2021 AMIA Board of Directors

Officers

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

Gretchen Purcell Jackson, MD, PhD, FACS, FACMI, FAMIA
IBM Watson Health &
Vanderbilt University Medical Center
Chair-elect

Neil Sarkar, PhD, FACMI
Brown University
Treasurer

Theresa A. Cullen, MD
Regenstrief Institure
Secretary

Directors

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

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

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

James J. Cimino, MD
University of Alabama at Birmingham

Susan C. Hull, MSN, RN-BC, NEA-BC, FAMIA
CareLoop, Inc.

Laura K. Heermann Langford, PhD, RN, FAMIA
Intermountain Health

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

Wanda Pratt, PhD, FACMI
University of Washington

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

Victoria L. Tiase, MSN, RN-BC
NewYork-Presbyterian Hospital

Adam Wright, PhD
Vanderbilt University Medical Center

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

Ex-Officio Board Members

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

Josette Jones, RN
Academic Forum Executive Committee Chair
Indiana University

Eileen Koski, MPhil, FAMIA
Informatics Partnership Council Chair

Jonathan R. Nebeker, MD, MS
Health Systems Chair
Veterans Health Administration (VHA)

Lynda Hoeksema, MSN, FNP-BC, RN-BC
Student Working Group Representative
The Ohio State University

Karen Greenwood
Executive Vice President and COO/Interim CEO

AMIA 2021 Virtual Informatics Summit Scientific Program Committee

Chair
Samuel Volchenboum, MD, PhD
University of Chicago

Vice Chairs

Majid Afshar, MD, MS
University of Wisconsin-Madison
Data Informatics Track

Robert Freimuth, PhD
Mayo Clinic
Translational Bioinformatics Track

Nicole Weiskopf, PhD
Oregon Health and Science University
Informatics Implementation Track

Laura Wiley, PhD
University of Colorado
Clinical Research Informatics Track

Members

Clinical Research Informatics Track

Hwayoung Cho, PhD, RN
University of Florida, College of Nursing

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

Eric Hall, PhD
Geisinger

Chris Harle, PhD
University of Florida

Jeremy Harper, MS
Indiana CTSI, Regenstrief Institute, Indiana University

Sabrina Hsueh, PhD
Viome, Inc.

Yun Jiang, PhD, MS
University of Michigan

Luke Rasmussen, MS
Northwestern University

Anthony Solomonides, PhD, MSc(Math), MSc(AI)
NorthShore University HealthSystem

Umit Topaloglu, PhD
Wake Forest Baptist Health

Wei-Qi Wei, MD, PhD
Vanderbilt University Medical Center

Tamara Winden, PhD
University of Kansas Medical Center

Data Science Track

Zhe He, PhD
Florida State University

Julian Hong, MD, MS
UCSF

Mei Liu, PhD
University of Kansas Medical Center

Anoop Mayampurath, PhD
University of Chicago

Sanjay Purushotham, PhD
University of Maryland Baltimore County

Justin Rousseau, MD, MMSc
Dell Medical School at the University of Texas at Austin

Ning Shang, PhD
Columbia University

Vignesh Subbian, PhD
University of Arizona

Lina Sulieman, PhD
Vanderbilt University

Richard Taylor, MD, MHS
Yale University School of Medicine

Kavishwar Wagholikar, MD, PhD
Harvard Medical School

Yiye Zhang, PhD
Weill Cornell Medical College

Maryam Zolnoori, PhD
Mayo Clinic

Informatics Implementation Track

Davera Gabriel, RN
Johns Hopkins ICTR

Jennifer Garvin, PhD, MBA
The Ohio State University/Department of Veterans Affairs

Yang Gong, MD, PHD
University of Texas Health Science Center at Houston

Ram Gouripeddi, MBBS, MS
University of Utah

Güneş Koru, PhD, FAMIA
University of Maryland, Baltimore County

Young Ji Lee, PhD, MSN, RN
University of Pittsburgh

Leslie McIntosh, PhD
Ripeta

Abu Mosa, PhD
University of Missouri School of Medicine and Center for Biomedical Informatics

Rimma Perotte, PhD
Hackensack University Medical Center

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

Russ Waitman, PhD
University of Kansas Medical Center

Lingling Zhang, ScD, MS, MPA
University of Massachusetts Boston

Translational Bioinformatics Track

Murthy Daverakonda, PhD
College of Health Solutions, Arizona State University

Blanca Himes, PhD
University of Pennsylvania

Harry Hochheiser, PhD
University of Pittsburgh

Indika Kahanda, PhD
Montana State University

Rebecca Levinson, PhD
University of Heidelberg

Faisal Mahmood, PhD
Harvard Medical School

Meeta Pradhan, PhD
Indiana Biosciences Research Institute

Arvind Rao, PhD
University of Michigan, Ann Arbor

Matthew Scotch, PhD, MPH
Arizona State University

Murat Sincan, MD
University of South Dakota

Theresa Walunas, PhD
Northwestern University

Ping Zhang, PhD
The Ohio State University

AMIA 2021 Informatics Summit Reviewers

Abdollahi, Behnaz

Abend, Aaron H

Abrams, Meredith

Achuthan, Srisairam

Ahmed, Kamran

Ahmed, Zeeshan

Akbilgic, Oguz

Akter, Sadia

Alag, Shray

Alaqeel, Abdulaziz

Alasmari, Ashwag

Albar, Aisha

Alekseyenko, Alexander V.

Alipanah, Neda

Alsheikh, Abdulrahman

Andreoletti, Gaia

Anton, Bonnie

Anzalone, Alfred Jerrod

Arbatti, Lakshmi

Arnaout, Rima

Arneson, Douglas

Arruda-Olson, Adelaide M.

Averitt, Amelia Jean

Baig, Furqan

Bailey, Charles

Balyan, Renu

Banda, Juan

Banerjee, Imon

Bangash, Hana

Bansal, Arvind

Barman, Arko

Barrett, Laura A

Bar-Shain, David

Batiste, Rebecca C

Bedoya, Armando

Bell, Douglas S

Belouali, Anas

Bennett, Kristin P.

Bennett, Tellen D

Beno, Mark F

Berano Teh, Jennifer

Bettencourt-Silva, Joao H

Bhanot, Karan

Bhattacharyya, Anirban

Bhattarai, Kriti

Bhavnani, Suresh K

Bhavsar, Nrupen Anjan

Bhuvaneshwar, Krithika

Bidwell, Jonathan

Binkheder, Samar

Bo, Na

Bodenreider, Olivier

Boguslav, Mayla R

Bokov, Alex F

Bona, Jonathan P

Bonney, Wilfred

Bope, Christian

Borglund, Erin

Borkowski, Steven A

Boyce, Richard D

Boyd, Andrew

Bradley, David J

Brickman, Arlen B

Brotman, David

Brown, Jeffrey

Bruce, Christa

Bucher, Brian T

Burton, Michael

Bush, Brian J

Buzalko, Russell J.

Campbell, Elizabeth A.

Cantor, Michael

Chakraborty, Prithwish

Chandak, Payal

Chandras, Rajan

Chaney, Kira

Chang, Marcello K

Chapman, Martin

Charpignon, Marie

Chartash, David

Chekalin, Evgenii

Chen, David

Chen, Jing

Chen, Nolan

Chen, Qingyu

Chen, Yue-ming

Chen, Zhaoyi

Chen, Ziqi

Cheng, Alex

Cheng, Christine M

Cheng, Jun

Choi, Jeeyae

Choi, Moon

Chu skychu, Yuan-Chia

Cimino, James J

Clarkson, Melissa

Co Jr, Manuel C.

Coker, Bolaji

Cole, Curtis L

Combs, Landon S

Contaxis, Nicole

Conway, Mike

Cook, Lily A

Cooke Bailey, Jessica

Cooper, Kathryn M

Craven, Catherine K

Crawford, Tami L

Croghan, Ivana T

Culbertson, Nick

Curcin, Vasa

Danahey, Keith

Daneshvari Berry, Shamsi

Das, Sudeshna

Datta, Surabhi

Davidson, Lena Marie

Davila, Jaime I.

Davila, Jessica A

Davis, Mary F

Davis, Sharon E

Day, Margaret A

De Alasei, Dana

De Freitas, Jessica

Dean, Julianna M

Debopadhaya, Shayom

Demirel, Doga

Deng, Yu

Deval, Shyam

Dey, Vishal

Dhawan, Andrew

Diaz-Garelli, Franck

Diederich, Catherine

Dikilitas, Ozan

Dinakarpandian, Deendayal

Ding, Xiruo

Ding, Ying

Doan, Son

Dobbins, Nicholas J

Doherty, Jennifer Anne

Dorr, David

Douthit, Brian James

Edgcomb, Juliet B

Eickhoff, Carsten

El-Azab, Sarah

Eldredge, Christina

Ellingson, Sally R

Estiri, Hossein

Evans, Clark

Facchini, David

Facelli, Julio C

Fan, Jungwei

Fan, Yadan

Fear, Kathleen

Feldman, Keith

Feng, Yujuan

Finkelstein, Joseph

Fitzmaurice, Laura

Florez-Arango, Jose F

Follett, Robert W.

Foraker, Randi E

Fort, Daniel

Foryciarz, Agata

Foster, Marva

Fouladvand, Sajjad

Fultz Hollis, Kate

Fung, Kin Wah

Furuhata, Hiroki

Gallego, Blanca

Gao, Grace

Gao, Jifan

Garcia-Milian, Rolando

Gartrell, Kyungsook

Garza, Maryam

Gaspar, Fraser

Gates, Evan

Gaudioso, Carmelo

Ge, Weiwei

Ge, Wendong

Gensheimer, Michael F

Giampanis, Stefanos

Gianfrancesco, Milena

Giannaris, Pericles S

Giuse, Dario

Gogia, Shashi Bhushan

Gold, Rachel

Goncalves, Luciana S

Gong, Yang

Goodwin, Travis Reed

Gordon, Geoffrey D

Gottlieb, Assaf

Green, Ian J

Greene, Sarah

Griffith, Brian

Gundelach, Justin H

Gunn, Martin

Guo, Aixia

Gupta, Aditi

Gutteridge, Charles

Hahn, Lewis

Haines, Jonathan

Hajaj, Chen

Hamid, Zeyana A

Hammond II, William Edward

Hannawi, Yousef

Hanrahan, Lawrence

Hardy, Lynda R

Harle, Christopher A

Harris, Debra Fasteson

Harris, Nomi L

Hasan, Md Mehedi

Hawthorne, Christopher

Haynes, Winston Andrew

He, Yongqun Oliver

He, Zhe

Heider, Paul M.

Hempelmann, Christian F.

Henry, Sam

Hmwe, Susan

Hoffman, Jeffrey

Hogan, William

Holl, Felix

Hong, Na

Hoyt, Robert E

Hsiao, Allen

hsieh, kang-lin

Hu, Ruifeng

Hu, Xinyu

Huang, Hu T

Huang, Yan

Huang, Ziming

Hui, Chi Ching Vivian

Hume, Sam

Hunter-Zinck, Haley

Hwang, Shelley

Hylock, Ray

Hynes, Kelly

Idrees, Ifrah

Israni, Sharat

Jaffe, Charles

jain, sarthak

Jeanselme, Vincent

Jeong, In cheol

Ji, Christina Xinyue

Jiang, Huizhen

Jiang, Yun

Jing, Xia

John, Jennifer Neda

Johnson, Darren

Johnson, Steven G

Jones, J.B.

Jones, Josette

Joshi, Shreekanth

Jung, Hyunggu

Jung, Jae-Yoon

Jung, Kenneth

Jungbauer Jr, Walter N

Kaelber, David

Kamdar, Maulik Rajendra

Kancherla, Jayaram

Kandula, Vijayabhaskar

Kang, Mengjia

Kannan, Vaishnavi

Kartoun, Uri

Karvir, Hrishikesh

Kasthurirathne, Suranga N

Kausar, Khadeja

Kayaalp, Mehmet

Kefayati, Sarah

Kehl, Kenneth L

Kember, Rachel L

Kerns, Ellen K

Khader, Shameer

Khan, Md Abdullah Al Hafiz

Khiabanian, Hossein

Khumrin, Piyapong

Kikinis, Ron

Kim, Yejin

Kim, Youngjun

Kimura, Eizen

King, Andrew J

Klann, Jeffrey G

Klasky, Hilda

Knosp, Boyd M

Kockara, Sinan

Konda, Pradap V

Kong, Sek Won

Kostka, Kristin
Kothari, Amit
Krichevsky, Spencer
Kuang, Zhaobin
Kuelbs, Cynthia
Kumar, Manish
Kumar, Sayantan
Kunisch, Joseph M
Kuo, Tsung-Ting
Kurc, Tahsin
Kury, Fabricio
Kuusisto, Finn
Kwon, KiBeom
Labilloy, Guillaume
Labkoff, Steven E
Ladhania, Rahul
Lai, Jiaying
Landman, Joshua M
LaPolla, Fred
Larson, Nicholas B
Laurio, Angela L
Le, Hoa Van
Ledbetter, David
Lee, Adam M
Lee, E. Sally
Lee, Kahyun
Lee, Robert
Lee, Young Ji
Lehmann, Harold P
Lemke, Klaus
Lenskaia, Tatiana
Li, Fang
Li, Fuhai
Li, Xiangrui
Li, Xin
Li, Ying
Li, Yun
Liang, Chen
Liang, Lifan
Lim, Hansaim
Lim Choi Keung, Sarah N
Lin, Esther
Lin, Rebecca Z
Liu, Hao
Liu, Hongfang
Liu, Nan
Liu, Ninghao
Liu, Sijia
Liu, Vincent
Liu-Ferrara, Ann
Lovely, Jenna
Lu, Dai-Yin
Lu, Yaoqin
Luo, Xiao
Luo, Zongwei
Lynch, Selah F
Lytle, Kay
Madan, Piyush
Madani, Sina
Magoc, Tanja
Mahajan, Satish M
Mahendran, Darshini
Major, Vincent J
Makeda, Kai
Malec, Scott A
Maltenfort, Mitchell
Manataki, Areti
Mandal, Meisha
Mandl, Kenneth
Manickam, Raj N
Maroilley, Tatiana
Masino, Aaron J

Mason, Jeremy C

Mason, Sarah

Masood, Muhammad

Matsil, Adam

Mayampurath, Anoop

Mays, Mary Helen

McClay, James C

McCoy, Matthew D

McCoy, Rozalina G

McCusker, James P

McDermott, Matthew

McGarvey, Peter

McGilchrist, Mark

McInnes, Bridget

McKillop, Mollie M

McPeek Hinz, Eugenia

Melamed, Rachel D

Meng, Frank

Meroueh, Chady

Metke Jimenez, Alejandro

Meystre, Stephane

Michelson, Andrew

Milgrom, Zheng

Miotto, Riccardo

Mironova, Maria

Mishra, Meenakshi

Mitchell, Sandra H

Moen, Erika L

Mogharabnia, Reyhaneh

Moldwin, Asher

Mork, James

Movahedi, Faezeh

Mowery, Danielle

Naeem, Jacqueline

Nagaie, Satoshi

Nair, Sujith Surendran

Nakhaei, Noor

Naumann, Tristan

Nelson, Therese A

Newman-Griffis, Denis R

Nho, Kwangsik T

Nicora, Giovanna

Nigam, Aastha

Ning, Xia

Nocera, Luciano

Norgeot, Beau

Noshad, Morteza

Novak, Laurie

Nuñez, John-Jose

Nye, Benjamin E

Ogallo, William

Oh, Wonsuk

Olex, Amy L

Ong, Toan

Oravecz, William T

Orlenko, Alena

Osborne, John D

Osinski, Kristen I

Ostasiewski, Brian

Ostovari, Mina

Ozmen, Ozgur

Pacheco, Jennifer

Padley, Michelle A

Palchuk, Matvey B

Palmer, Ellen Lorraine

Parbhoo, Sonali

Park, Albert

Park, Jung In

Park, Yoonyoung

Parra-Calderón, Carlos Luis

Patel, Shrawan

Payrovnaziri, Seyedeh

Pena, Danilo

Peterson, Kevin

Pfaff, Emily R

Pfohl, Stephen R

Philip, John

Philips, Santosh

Phillips, Mark H

Podchiyska, Tanya

Poley, Stephanie T

Poole, Sarah F

Popescu, Mihail

Popovic, Jennifer R

Porcino, Julia

Portales-Casamar, Elodie

Poterack, Karl

Pottinger, Tess D

Pradhan, Prajwal Mani

Prodduturi, Naresh

Provance, Jeremy

Pullum, Laura L

Raisaro, Jean Louis

Rajagopalan, Aravind

Raje, Satyajeet

Raju, Murugesan

Rasmy, Laila

Ravanmehr, Vida

Rehman, Shakaib U

Reimer, Andrew P

Richardson, Alexander

Richardson, Joshua E.

Ridgway, Jessica

Robasky, Kimberly

Roberts, Kirk

Robinson, Peter N

Rocheteau, Emma C

Rodriguez, Laritza Maria

Rodriguez, Victor A

Rollison, Dana E

Rosati, Robert J

Rosenbloom, S. Trent

Rossi, Lorenzo A.

Rotenberg, David

Rouhizadeh, Masoud

Rudrapatna, Vivek

Russell, Seth

ryan, james t

Ryan-Lora, Beatriz

Sadatis, Christal T

Saini, Divya

Sakaguchi, Farrant

Saleh, Sameh Nagui

Salsabili, Mahsa

Saltsman, Connie

Sanghavi, Devang

Santillan, Donna A

Santillan, Mark K

Sarker, Abeed

Sathees, Saraswathi

Sato, Jumpei

Sauta, Elisabetta

Sawyer, Tatiana

Scheufele, Elisabeth

Schlegel, Daniel R

Schlueter, David

Schmidt, Teresa D

Schrom, John

Seoane, Jose A

Shadmi, Efrat

Shankar, Rama

Sharma, Brihat

Sharma, Vishakha

Shaw, Timothy I

Shayegani, Shapoor

Sheets, Lincoln R

Shen, Li

Shendre, Aditi

Shi, Jingyi

Shin, Soo-Yong

Sholle, Evan

Shrestha, Aashara

Si, Yuqi

Sid, Eric

Siddicky, Safeer F

Simpkins, Carolyn

Simpson, Christopher L

Singh, Angad Preet

Sittig, Dean F

Smalheiser, Neil R

Smith, Corey

Smith, Jaime Y

Smith, Kelly M

Snowdon, Jane L

Soares, Andrey

Somani, Sulaiman

Song, Hsing-Yi

Song, Qianqian

Song, Xiaoyu

Song, Xing

Soni, Sarvesh

Sonkin, Dmitriy

Sonnenberg, Frank A

Sood, Akshay

Sotoodeh, Mani

Sottara, Davide

South, Brett

Speakman, Skyler

Staes, Catherine

Stewart, Ron

Stiepcich, Monica A

Strasberg, Howard R

Strasser, Zachary

Sulieman, Lina

Sulley, Saanie

Sun, Deyu

Sun, Zhaonan

Sunkara, Padageshwar

Sverchkov, Yuriy

Sward, Katherine A

Syed-Abdul, Shabbir

Syrowatka, Ania

Szymanski, Jeffrey

Tadesse, Girmaw Abebe

Tahmasebi, Amir

Tajgardoon, Mohammadamin

Talbert, Jeffery

Talmon, Geoffrey

Tannier, Xavier

Tao, Carson

Tarabichi, Yasir

Tate, Tia A

Tavakoli Hosseinabadi, Maryam

Taweel, Adel

Taylor, Bradley W

Thate, Jennifer

Tiase, Victoria

Tomita, Naofumi

Tommasi, Pierpaolo

Tonellato, Peter J

Torii, Manabu

Torres, Marisa

Trinkley, Katy E

Tutaj, Monika

Vashisht, Rohit

Vashishth, Shikhar

Vasilevsky, Nicole

Verspoor, Karin

Visweswaran, Shyam

Vunikili, Ramya

Wadia, Roxanne

Walden, Anita

Walling MD, PhD, Anne

Walters-Threat, Lois E

Wang, Lei

Wang, Lucy Lu

Wang, Peng

Wang, Shan

Wang, Wenjie

Wang, Yanshan

Wang, Yaqiang

Wang, Zengyan

Wang PhD, Li-San

Webb, Louise K

Weber, Griffin M

Weber, Jens H

Wei, Wenfei

Weiner, Mark

Weir, Charlene

Wen, Andrew

West, Vivian L

Westra, Bonnie L

Whitley, Eric W

Wi, Chung-Il

Wieland, Daryl L

Williams, Marc S

Williams, Nick

Windle, John

Wishnie, Lauren M

Wood-Wentz, Christina

Workman, Terri E

Wu, YiFan

Xie, Sherrie

Yan, Chao

Yan, Xiaowei Sherry

Yang, Kai-Chieh

Yang, Xinan Holly

Yang, Yuan-Chi

Yazdanparast, Aida

Ye, Cheng

Ye, Jiancheng

Yu, Sean Chonghwan

Yu, Yiqin

Yu, Yue

Yuan, Jianbo

Zachary, Iris

Zelle, David

Zhang, Aaron

Zhang, Lin

Zhang, Ping

Zhang, Tianlin

Zhang, Xiang

Zhang, Xiaoli

Zhang, Xinyuan

Zhang, Yuji

Zhao, Yiqing

Zhao, Yunpeng

Zheng, Chunlei

Zheng, Hua

Zheng, Yaguang

Zhou, Xin

Zhu, Ping

Zhu, Qian

Zhu, Vivienne

Zimmerman, Lindsay

Zimolzak, Andrew

Zong, Nansu

NOTICE

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

Accelerating an Application Programing Interface-based Ecosystem with Real-World Use Cases

Kevin Chaney, MGS[1], Kenneth D. Mandl, MD, MPH[2], Kristen Miller, DrPh, CPPS[3], Daniel J. Chavez, MBA[4], Anjum Khurshid, MD, PhD[5]
[1]Office of the National Coordinator for Health Information Technology, Washington, DC; [2]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA; [3]National Center for Human Factors in Healthcare, MedStar Health, Washington, DC; [4]San Diego Health Connect, San Diego, CA; [5] Dell Medical School, The University of Texas at Austin, Austin, TX

Abstract

The utilization of application programming interfaces (APIs) in healthcare has potential to enhance population health, patient care and research. Furthered by regulation issued by the Office of the National Coordinator for Health Information Technology (ONC) and the Centers for Medicare & Medicaid Services (CMS), patients and providers will have greater access to electronic health information. In anticipation of a new generation of health IT, ONC issued the Leading Edge Acceleration Projects (LEAP) in Health IT funding opportunity, to advance well-designed, interoperable, and scalable health IT for care and research. This panel will showcase four innovative initiatives, including a provider-payor use case employing a universal bulk-data API; a provider-facing clinical knowledge risk calculator app embedded in an electronic health record; an efficient, transparent and secure consent management prototype using a standards-based authorization framework; and a patient-engagement platform that empowers patients to gain and control access to their personal health data.

Introduction

The Office of the National Coordinator for Health Information Technology (ONC) is at the forefront of the administration's health information technology (IT) efforts and leads efforts to support the adoption of health information technology and promotion of nationwide health information exchange to improve healthcare. Over the last decade, ONC has made tremendous progress towards advancing not only the adoption and use of health IT across the healthcare ecosystem, but also increasing the rate of growth and innovation through its many programs to bridge policy with operational initiatives that can accomplish its mission.[1] Programs such as the Strategic Health Information Technology Advanced Research Projects (SHARP) cooperative agreements[2] aimed to close the gap between the promise of health IT and its realized benefits. Hailed as the major achievement from the SHARP program was the Substitutable Medical Applications, Reusable Technologies (SMART) Health IT,[3] a standards-based technology platform that enables innovators to create medical applications (apps) that seamlessly and securely run across the healthcare system.[4-5]

More recently, there has been a rapid progression in the types of technologies and innovations being utilized in all health domains, including personal use for patient generated health data, clinical tools and apps being used at the point of care, and surveillance and population health tools that may include social determinants of health data. As the electronic exchange of health data has matured, the amount and types of health data available has expanded as well. Data standards such as Health Level Seven International's (HL7®) Fast Healthcare Interoperability Resources (FHIR®) and application programming interfaces (APIs) are facilitating the sharing of health data in an interoperable way while using open standards that even non-healthcare industry technologists can leverage. Despite these advances, there is still much to learn about delivering and presenting emergent data seamlessly to patients and providers in both clinical and non-clinical settings in support of care and research. To better address these gaps and support alignment between the clinical and research ecosystems, ONC released National Health IT Priorities for Research: A Policy and Development Agenda (the Agenda),[6] which identifies nine priority areas and corresponding actions. ONC's is working to advance these priorities through a variety of actionable programs and policies aimed at: leveraging EHR data to support patient- and population-level research, analyses and services; improving patient engagement applications; enhancing consent management platforms; and improving tools to integrate clinical knowledge into routine clinical practice.

Leading Edge Acceleration Projects (LEAP) in Health IT

In 2018, ONC published the Leading Edge Acceleration Projects (LEAP)[7] in Health IT funding opportunity. The goal of this three-year funding opportunity is to further a new generation of health IT development and inform the innovative implementation and refinement of standards, methods, and techniques for overcoming major barriers and challenges in the field. This panel presents in detail the focus of LEAP in Health IT funding opportunity and the ongoing work of four funded projects. Key aspects of these projects focus on reducing provider and health system burden of utilizing health IT, while incorporating data access and use via an API for innovative purposes in support of care and research. LEAP in Health IT is an example of ONC's ability to fund projects that seek to overcome challenges that inhibit the development, use, or advancement of well-designed, interoperable health IT affecting care and research.

Panel Objectives and Presenters

This panel will provide an overview of how projects funded under the ONC's LEAP in Health IT funding opportunity is preparing the U.S. for an interoperable, modular, health IT ecosystem. Panelists will discuss results from projects awarded in 2018, and progress to date on the projects awarded in 2019, in addition to current and emerging challenges facing the field, results of canonical use case prototypes, and priorities for next steps in alignment with national health IT priorities for research. Each panelist will leave time for discussion around the development, implementation, and use of their respective solution, as well as early legal and policy implications. This panel aims to stimulate a highly interactive discussion and strengthen the community's knowledge of innovative uses of health IT, barriers and solutions for use, and areas ripe for future work.

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

Dr. Kenneth D. Mandl (panelist), directs the Computational Health Informatics Program at Boston Children's where he leads the transformative SMART Health IT initiative and is Principal Investigator of a 2018 ONC LEAP in Health IT project. He will provide an overview of a population health use case for the FLAT FHIR® Bulk API, which can drive change for populations, payers, providers, and patients at scale. Dr. Mandl is developing and testing a production scale, open source, reference population health app for use between payers and hospitals. The project relies on the emerging SMART/HL7® bulk data export standard, which has already been implemented in EHRs through the Argonaut process, and is an HL7 Standard for Trial Use (STU).

Dr. Kristen Miller (panelist), is the Scientific Director of the National Center for Human Factors in Healthcare at MedStar Health, an Associate Professor of Emergency Medicine at Georgetown University School of Medicine, and Associate Faculty at the Innovation Center for Biomedical Informatics at Georgetown Medical Center. She is the Principal Investigator of a 2018 ONC LEAP in Health IT project. She will provide an overview of MedStar Health's work transforming a stand-alone Million Hearts Risk Calculator into an interactive surveillance tool, as a SMART on FHIR app, and will discuss lessons learned in the technical design and integration, legal and policy implications, and opportunities for future enhancements.

Mr. Dan Chavez (panelist), is Executive Director of the San Diego Health Connect (SDHC) health information exchange (HIE) and provides leadership for a 2019 ONC LEAP in Health IT project. He will provide an overview of how SDHC has advanced efforts to test the FHIR Consent Implementation Guide and a package of open-source prototypes to demonstrate efficient, transparent, and secure consent management and data exchange. This work will be used to develop APIs that enable consent use cases that will advance patient-centered care, informed consent, and shared decision-making.

Dr. Anjum Khurshid (panelist), is the inaugural Director of Data Integration in Dell Medical School's Department of Population Health at the University of Texas at Austin and Principal Investigator of a FY 2019 ONC LEAP in Health IT project. He will provide an overview of a patient-engagement technology platform, *FHIRedApp*, being developed to support an ecosystem of mobile applications. The initiative uses a modified Community Engagement Studios approach to empower patients from underrepresented populations, to actively participate in the design of a standards-based, privacy preserving, and secure mobile platform to access and share their health data without special effort.

Panel Discussion Questions

- How are advancements from these projects available for other organizations to use and leverage?

- Are there disruptive technologies the field should be prepared for?
- What needs or gaps have these new technologies addressed and what remains?
- What technical and policy needs or gaps limit the usage of APIs for health care and research?
- Are there health IT infrastructure and/or standards barriers impeding the utilization of these solutions?
- How easily can these solutions be applied and scaled to other similar aspects in the field (e.g., other risk calculators, bulk data types, consent resources, or patient-engagement technologies)?
- How can ONC broaden the pool of use cases able to utilize these established technologies and standards (e.g., integrating with human services, mental/behavioral health services, or other non-clinical health services)?
- How does ONC ensure that new technologies improve access and use of health information for care and research?

Panel Leaning Objectives

1. Participants will understand ONC's priorities, with a strong focus on the LEAP in Health IT Program, and the newest focal areas of interest.
2. Participants will learn about the technical development and utilization of each LEAP in Health IT project to date.
3. Participants will learn the challenges and barriers experienced by each LEAP in Health IT project and the impact for broader uptake by the field.
4. Participants will learn about technical and policy gaps that are limiting the use of FHIR-based APIs.

Conclusion

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

Statement of Participation

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

References

1. The Office of the National Coordinator for Health Information Technology. About ONC [Internet]. Washington D.C.: ONC; 2019 [updated 2019 February 14; cited 2020 Feb 25]. Available from: https://www.healthit.gov/topic/about-onc
2. Friedman C. Health IT challenges and the future of healthcare [Internet]. Washington D.C.: Office of the National Coordinator for Health Information Technology; 2010 [updated 2010 Apr 2; cited 2020 Feb 25]. Available from: https://www.healthit.gov/buzz-blog/sharp/health-it-challenges-and-the-future-of-healthcare
3. SMART Health IT. What is SMART? [Internet]. Boston, MA: Harvard Medical School; 2018 [updated 2018 Mar 08; cited 2020 Feb 25]. Available from: https://smarthealthit.org/an-app-platform-for- healthcare/about/
4. Mandl K. Can Apple take healthcare beyond the fax machine? [Internet]. Boston, MA: Harvard Medical School; 2018 [updated 2018 Jan 30; cited 2020 Feb 25]. Available from: https://smarthealthit.org/2018/01/can-apple-take-healthcare-beyond-the-fax-machine/
5. Cartwright HJ. Lighting up healthcare data with FHIR: Announcing the Azure API for FHIR [Internet]. Seattle, WA: Microsoft Azure; 2019 [updated 2019 Feb 07; cited 2020 Feb 25]. Available from: https://azure.microsoft.com/en-us/blog/lighting-up-healthcare-data-with-fhir-announcing-the-azure-api-for-fhir/
6. The Office of the National Coordinator for Health IT (ONC). National Health IT Priorities for Research: A Policy and Development Agenda [Internet]. Washington, DC: ONC, 2020. [updated 2020 Jan. 15; cited 2020 Mar. 11]. Available from: https://www.healthit.gov/sites/default/files/page/2020-01/PolicyandDevelopmentAgenda.pdf
7. Office of the National Coordinator for Health Information Technology. Notice of funding opportunity: Leading edge acceleration projects (LEAP) in health information technology [Internet]. Washington D.C.: ONC; 2018 [cited 2020 Feb 25]. Available from: https://www.healthit.gov/sites/default/files/page/2018-06/LEAPHealthIT.pdf

What We Know About Data Warehousing in Support of Clinical and Translational Research: National Survey Results and the Work Ahead

Catherine K. Craven, PhD, MA, MLS, FAMIA[1], Thomas R. Campion, Jr., PhD[2], Dave A. Dorr[3], MD, Jeremy Harper, MBI[4] , Boyd M. Knosp, MS, FAMIA[5]

[1]Institute for Health Care Delivery Science, Icahn School of Medicine at Mount Sinai, New York, New York; [2] Dept. of Population Health Sciences, Weill Cornell Medicine, New York, New York; [3] Dept. of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, Oregon; [4]Regenstrief Institute, Indianapolis, Indiana; [5]Institute for Clinical and Translational Science, Roy J. and Lucille A. Carver College of Medicine, Univ. of Iowa, Iowa City, Iowa

Abstract

The panel comprises a moderator presentation of a 50-Clinical and Translational Science Award (CTSA) hub survey in 2019 about their enterprise data warehousing (EDW) for research and four clinical research informatics (CRI) experts presenting reactions to results from their operations and research perspectives. EDW operations directors and technical staff were surveyed to understand differences in knowledge and perceptions. Results covered here include DW and data characteristics, data quality practices, data access/delivery, and workforce. Proliferation of data and algorithmic approaches and centrality of EDWs as loci for clinical data for secondary uses present significant challenges for managing growth sustainably. **Learning objectives are to** 1) Understand the trajectory of capabilities and deployment of EDWs for research from 2010-2020; 2) Define the variability in EDW development at different sites and potential causes of the variability; 3) Explore value and sustainability of EDWs for research.

Introduction

Catherine Craven will open this reactor panel by presenting results from a national survey about clinical data warehousing for research that she led in 2019 of the CTSA hubs funded by the National Institutes of Health National Center for Advancing Translational Science (NCATS). The survey, by a seven-CTSA team, was vetted and aided in distribution by leadership from the CTSA Informatics Enterprise Committee (iEC), Clinical Data to Health (CD2H) initiative, Healthcare Data Analytics Association, with distribution help from AMIA's CRI Working Group. Panelists will react to survey results from their expert perspectives. For the final third, she will facilitate audience questioning and discussion about EDW for research progress, and work, challenges, and expectations for sustainable growth.

Motivation and rationale - The panel will be timely, needed, and attention grabbing for the audience because of the proliferation of data and algorithmic approaches - and the centrality of enterprise data warehouses (EDWs) as loci for clinical data for secondary uses – which present significant challenges for managing growth sustainably.

Audience - The audience will comprise all stakeholders interested in data warehousing for research, including Chief Information Officers; chief officers for research informatics, data, and analytics; CTSA principle investigators and informatics leads; research deans; EDW directors and staff; biomedical informatics faculty; and the CRI community.

Presentation of the Survey on CTSA EDWs for research

Background and Significance - Between 2008 and 2017, EDW adoption increased from 64% to 94% at CTSAs.(1, 2) The need for granular examination of EDWs was raised related to informatics sustainability(3) as our survey was vetted. Our objective was to replicate relevant questions from the 2010 CTSA EDW survey(1) and expand topics to cover current practices. We targeted EDW directors and technical staff; staff carry out much of the work and interact most closely with EDW clients, yet are usually not surveyed. Staff knowledge or lack thereof may impact their ability to execute queries and sustain a chain of EDW-related reproducibility best-practices, including for EHR data.

Methods – The REDCap-based survey comprised two links: one for Part I, for the clinical/enterprise DW operations director at each institution, which comprised 42 questions, and another for Part II for each of the non-director DW technical staff employed by the DW who maintain it and/or work with end-users, which comprised 36 questions. Distribution was June 10, 2019, through Labor Day, September 2019, via member listservs and calls of groups mentioned, with emails to CTSA PIs and informatics leads of most of the ~60 CTSAs. For analysis, fall 2019-winter 2020, categorical variables were summarized by n (%), while scaled data were summarized by the median and the

range. Distributions of categorical and scaled data were compared using the Fisher's exact test and Mann-Whitney test, respectively. Univariable analyses were performed. Here we will present those comparing CTSA directors vs. CTSA staff. Hypothesis testing was two-sided and conducted at the 5% level of significance. All statistical analyses were done using SAS v9.4 (SAS Institute, Cary, NC).

Results – ***Demographics*:** EDW ("DW") Directors from 50 CTSAs and 91 technical staff members responded.

***Data Warehouse Characteristics*:** All responding CTSAs have a DW up from 86.0% in 2010(1). However, 74.0% of CTSAs here but have a purpose-built DW geared toward research. Most DWs are built internally, a trend stable since 2010. Eight are part of a commercial software package. Most DWs are on-site on local servers. For 38, the DW comprises >million patient records, with another 9 DWs comprising 100,000-to-one-million. In 2010, median size for responding CTSAs was 1.6 million patients(1), indicating growth. More than half the CTSAs, 26, use an internally-developed, institution-specific data model. However, 17 CTSAs said they do structure their data according to a common data model (CDM), and 26 said they provide data to researchers according to a CDM. These CDMs developed since 2010 with the rise of their associated Clinical Data Research Networks (CDRNs).

***Data Characteristics and Data Quality (DQ)*:** Clinical data types absent in 2010 are now ubiquitous. Most DWs (80.0%) include 6-11 types, and >40.0% including genomics; CTSA staff and directors disagreed on data available. Data interoperability is improved: 96.0% of CTSAs are coding with LOINC and two-thirds coding problem lists with SNOMED or ICD codes. This is progress: in 2010 ~18% of respondents were planning for "standardization/ terminologies," considered among biggest challenges.(1) Data enhancement increased: Almost 2/3 of CTSAs add calculated fields; just over half geo-code addresses, but fewer standardize addresses. About 1/3 reconcile update deaths from the National Death Index, up from ~20% in 2010, and immunizations. Data provision for registries was not probed in 2010.(1) DWs now provide data or data-mart services from 44.0% for tumor registries to 10% for rare disease registries. DQ was among top 2010 challenges, although none was reported.(1) About 1/3 now check DQ in most source systems on entry, 62.0% check DQ during ETL into the DW and on data in the DW. Not quite 3/4 quality-check data extracted for users. Over half said DQ check results are stored and can be viewed/used later. Under half communicate checks to source system owners. Just under 40.0% assess changes in DQ over time. Under 1/3 provide DQ results to data recipients, to whom not probed. Staff and director responses often differed.

***Data access/delivery, workforce, fees*:** CTSAs provide more self-service access to de-identified data now than in 2010, so researchers can self-start: for aggregate patient data it's 70.0% v 54.0%; for de-identified patient-level data, 42.0% v. 31.0%;(1) Analyst-performed query-volume was not reported in 2010, but staffing capacity/expertise was the largest challenge; funding was also a challenge, and the avg. no. of end-user support staff was 6.42 full-time equivalents (FTEs).(1) The median analyst FTEs now employed to assist research users, including performing queries for them, is 5, with 28% of CTSAs reporting 6-10. A median of 4 maintain the DW; roles overlapped at some CTSAs. Staff has not increased at many CTSAs, yet researchers often need assistance, request volume likely has increased. 54.0% of CTSAs have a Chief Research Informatics Officer (CRIO).

Discussion/conclusion -- Results demonstrate a decade of progress. Areas for expected growth for maturity are potential challenges: adding and linking additional data types, data enhancements, and increasing the breadth and depth of DQ efforts.(4, 5) Findings show need for further education for DW staff, with implications for service-delivery quality and impact on research reproducibility. Rise of the unstandardized CRIO role, and now the sometimes overlapping Chief Data/Analytics Officer, demonstrates maturing understanding of the need for executive sponsorship to steer data governance and prioritize efforts. How CTSAs plan for, invest in, and budget to grow and sustain these efforts, a 2010 concern, and a sign of maturity, will remain a concern as will ensuring workforce capacity and informatics expertise to effectively drive them.

Table 1. Role, timetable, panelists, and presentation focus. All panelists here have agreed to participate.

Role / time	Panelist and Focus of Presentation

Moderator: Introduction (5 minutes) Survey presentation (15 minutes)	Catherine K. Craven is a Senior Clinical Research Informaticist in the Institute for Health Care Delivery Science, Dept. of Pop. Health Science and Policy, Icahn School of Medicine at Mount Sinai and appointed in the Clinical Informatics Group, IT Dept., MSHS. She conducts CRI research and operations work to inform, improve, and sustain EDWs for research and informatics approaches for research IT, and research for informatics engagement of vulnerable patients. She will present the survey and facilitate the audience discussion with panel members.
Panelist 1: Response to survey results (5 minutes)	David A. Dorr is CRIO and a Professor and Vice Chair of Medical Informatics and Medicine at OHSU. He researches how to improve systems for vulnerable populations and contemplates strategic research informatics and innovation needs and systems-based solutions for OHSU and the Oregon Clinical & Translational Research Institute. He will discuss survey results via evolution of the EDW for research at OHSU. The survey revealed significant variation in data integration, DQ checking, and staff /informatics leader understanding of capabilities, which highlight rapid, uneven growth of EDWs. He will discuss implications for strategic planning, communication, education, sustainability planning, and opportunities to address these through institutional leadership, collaboration, and new initiatives.
Panelist 2: Response to survey results (5 minutes)	Jeremy Harper is CRIO at Regenstrief Institute & Indiana Clinical and Translational Sciences Institute. For 10 years, he has served in leadership roles in hospitals and healthcare and led departments responsible for planning, implementation, and management of deployments and enterprise initiatives. He is developing a continuing education program for healthcare informatics. He will discuss survey results from an operations perspective, what results signify re staff data literacy, how his modules might fulfill needs, and additional areas needed.
Panelist 3: Response to survey results (5 minutes)	Boyd Knosp is Assoc. Dean for IT at the Univ. of Iowa's Carver College of Medicine and Assoc. Dir. of Informatics Operations at Univ. of Iowa Institute for Clinical & Translational Science where he leads a team to deliver state of the art informatics services. Boyd collaborates on projects to understand how DWs are used in medical education and clinical/translational science and to develop maturity models (eg, via CD2H) to help institutions understand and plan investments in IT and Informatics. He will discuss maturity model development for data quality, EDWs, and Informatics, and how survey results relate to/might inform that work.
Panelist 4: Response to survey results (5 minutes)	Thomas R. Campion, Jr., leads Weill Cornell Medicine's efforts to support clinical and translational investigators, especially through secondary use of EHR data. He is Assoc. Professor of Research in Pop. Health Science, Div. of Health Informatics, and Dir., Biomedical Informatics in the Clinical & Translational Science Center. He will provide commentary based on complementary findings from qualitative analysis of semi-structured interviews conducted through a CTSA iEC working group focused on data warehousing for research.

Questions to enhance audience participation during the questions and discussion period --

What DQ checking is your institution's EDW for research doing? Where are we going re CDRNs and CDM participation/sustainability? FHIR for research as a CDM? What challenges your institution re analytics maturity?

References

1. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. J Am Med Inform Assoc. 2012;19(e1):e119-24.
2. Obeid JS, Beskow LM, Rape M, Gouripeddi R, Black RA, Cimino JJ, et al. A survey of practices for the use of electronic health records to support research recruitment. J Clin Transl Sci. 2017;1(4):246-52.
3. Obeid JS, Tarczy-Hornoch P, Harris PA, Barnett WK, Anderson NR, Embi PJ, et al. Sustainability considerations for clinical and translational research informatics infrastructure. J Clin Transl Sci. 2018;2(5):267-75.
4. Zozus M HW, Green B, Kahn M, Richesson R, Rusinkovich S, et al. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0)2014 [cited 2020 August 17, 2020]. Available from: https://www.researchgate.net/profile/Meredith_Zozus/publication/283267713_Data_Quality_Assessment_Recommendations_for_Secondary_ise_of_EHR_Data/links/562f9d3908aeb1709b6000af.pdf.
5. Center DQSotNC. National Evaluation System for health Technology Ccoordinating Center (NESTcc) Data Quality Framework2020 July 23, 2020. Available from: https://nestcc.org/data-quality-and-methods/#frameworks.

Evidence-based Tools and Strategies for Evaluating the Safety of Health Information Technology Systems: The State of Practice, Challenges, and the Road Ahead

Moderator:
Aaron S. Dietz, PhD[1]

Panelists:
David Classen, MD, MS[2,3], Teja Kuruganti, PhD[4], Michael A. Rosen, PhD[5,6,7], Jeanie M. Scott, MS, CPHIMS[1]

[1]Department of Veterans Affairs, Office of Health Informatics, Informatics Patient Safety, Washington, District of Columbia; [2]University of Utah School of Medicine, Salt Lake City, Utah; [3]Pascal Metrics, Washington, District of Columbia; [4]Oak Ridge National Laboratory, Oak Ridge, Tennessee; [5]Johns Hopkins University School of Medicine, Baltimore, Maryland; [6]Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland; [7]Johns Hopkins School of Nursing, Baltimore, Maryland.

Background and Significance

Health information technology (HIT) has the potential to dramatically improve the safety and quality of care delivery. These systems, however, are complex[1] and have introduced new risks and challenges that may inhibit their true potential from materializing.[2] For instance, computerized provider order entry systems (CPOE) with decision support may not include adequate dosing alerts, engender alert fatigue, or include shortcomings related to system functionality.[3,4,5] These factors may explain why medication errors associated with adverse drug events still occur during medication ordering[6,7] despite electronic ordering mechanisms being linked to clinical treatment protocols and best practices.[8] There are also systemic issues related to a lack of interoperability between health information sources (e.g., electronic health records, bedside monitors, and infusion pumps) that create barriers to effective decision-making and care team coordination.[9]

A learning health system health system needs to (1) appreciate the complexity of HIT, (2) understand how the system can fail, and (3) identify ways to mitigate and manage risk. Recognition of these factors has propagated considerable investment to ensure the benefits of HIT are realized while eliminating or reducing the impact of unintended consequences that may jeopardize patient safety.[10,11,12,13] Yet the application of existing tools and best practices can be daunting for health care organizations due to resource constraints (e.g., time and staffing required to conduct testing), access to evaluators proficient in the appropriate use of the tool, and inherent limitations of different methodological approaches that may or may not cover the full spectrum of HIT risk points. Therefore, this panel brings together experts in the fields of health care, informatics, patient safety, and human factors to unpack the application of tools and methods for evaluating the safety of HIT systems. This panel will be dedicated to discussing the state of science and practice as it relates to ensuring the safety of HIT systems across life cycle phases through specific experiences. This panel will also include discussion on how these experiences can be applied to other applications to create robust learning health systems. Specific questions to enhance audience participation include:

- What are the most significant gaps/challenges to evaluating the safety of HIT systems across the health care community? How have these gaps/challenges impacted patient safety? How do we overcome these challenges?
- What are the challenges to scaling these methods? Are there unique considerations depending on the type of health care setting (e.g., a large integrated health system compared to a smaller rural hospital)? How do we overcome these challenges? What resources need to be in place before organizations can apply these tools?
- At what frequency should these tools and best practices be applied as part of a continuous monitoring or surveillance program? What are other instances that may prompt a need to conduct a full or partial evaluation of a HIT system?
- How can health care organizations learn from each other using these tools?
- How will the practice of HIT evaluation and monitoring change in the next year? Five years? Ten years?

Panelist Abstracts

Using Electronic Health Records for Realtime Patient Safety Detection and Prediction (David Classen, MD)

Twenty years after publication of the report *To Err is Human*, studies demonstrate persisting high levels of patient harm. Most patient safety measurement remains highly retrospective, relying on voluntary reporting and post discharge administrative coding. Progress has been limited by the lack of advances in measurement accuracy, detection sensitivity, and timely actionability. The broad adoption of electronic health records (EHRs) offers a significant opportunity to leverage digital information to improve safety measurement and management using real-time data. We developed a novel method to extract safety indicators from EHRs to identify harm and its precursors by implementing a patient safety active management system (PSAM) in hospitals within a national Patient Safety Organization (PSO). The PSAM generated validated adverse event outcomes and leveraged EHR data to develop a real-time safety predictive model. This study describes the PSAM's pilot at two large community hospitals in 2014–17. We found that the PSAM could detect harm in real time, at higher rates than current levels are detected, and that such harm could be predicted. In addition to outlining future opportunities and challenges with this EHR-enabled PSAM approach, we discuss implications and next steps for policy and practice.

Automated Hazard Detection Framework for Health Information Technology (Teja Kuruganti, PhD)

Adoption of electronic health records (EHRs) systems brought significant benefits to the system users, providers, and patients such as increased care quality and decreased healthcare costs. Most transactional systems in EHRs have evolved over decades and can have complex and sometimes redundant socio-technical interactions. The high degree of complexity can generate unintended consequences in safe use of healthcare information technology (HIT) systems and reveal themselves as hazards that can potentially interrupt care delivery. The safety concerns in HIT are categorized into three perspectives: (1) concerns unique or specific to malfunctioning hardware or software (safe operations), (2) concerns related to misuse of HIT (safe use of HIT), and (3) concerns related to potential outcomes (addressed by real-time or retrospective monitoring of the risks in the health care delivery processes).

ORNL in collaboration with Veteran's Affairs (VA) is developing an end-to-end framework for improving the reliability and performance of HIT systems. Our approach has three key objectives. Firstly, we developed a unified representation of the system state using standards-based methods. Secondly, we developed graphical model-driven approaches for evaluating HIT workflow to monitor state transitions and identify critical transitions that impact likelihood of faults. Thirdly, we developed data analytic techniques to explore corporate data warehouse (CDW) to detect hazards. The outcome of this effort is a prototypic tool for detecting hazards induced by HIT workflow in EHR systems. Key challenges that have to be addressed for wide-scale adoption such hazard detection frameworks are to identify performance metrics that drive detection requirements, generalization of methods to demonstrate applicability to diverse EHR systems, and scalability of detection techniques.

A Pilot Program for a Peer to Peer Learning and Assessment System for HIT Safety (Michael A. Rosen, PhD)

Peer-to-peer learning and assessment (P2P-LAS) systems are a critical practice for achieving highly reliable and safe operations in other high risk industries, such as commercial nuclear energy production and aviation. These are heavily regulated industries where external assessment by accrediting or commissioning organizations is a required component of operation. However, in these settings, it became apparent that these high stakes of assessments focused on adherence or compliance were not enough to encourage learning. P2P-LAS systems are different. They are completely confidential and focused exclusively on improvement and risk mitigation. This allows participants in these processes to be more open than they would under the eye of external regulators. Consequently, organizations can benefit from an external perspective without exposing themselves to financial, reputational, or regulatory risks. In healthcare, P2P-LAS systems have been developed for individual clinical skills and organizational governance and management of safety and quality. However, no P2P-LAS programs exist for improving Health Information Technology (HIT) Safety.

This presentation discusses results of a pilot program designed to develop and evaluate P2P-LAS tools and processes for HIT Safety. CPOE alerts were chosen as a focus area within the broader HIT Safety space due to operational priorities. Program tool content drew from industry standards (i.e., the SAFER Guides), literature review, and a Delphi panel process. The program process was developed from a literature review on P2P-LAS systems and the project team's prior experience implementing such programs for other safety and quality domains. The pilot program was implemented in two health systems. Qualitative results of the project will be shared. Effective P2P-LAS systems can form the basis of knowledge sharing between organizations and move safe practice forward.

Evaluating the Utility of a CPOE Assessment Tool for Widescale Adoption (Jeanie M. Scott, MS, CPHIMS)

Healthcare organizations need evidence-based tools and methods for identifying potential risk points prospectively. The purpose of this project was to evaluate the utility of the University of Utah's *EHR Flight Simulator* as a strategy for continuously assessing safety risks. The simulation involves a licensed provider entering simulated medication orders for simulated patients within an organization's test EHR system. The assessment process provides grades of CPOE performance along 10 medication ordering dimensions (e.g., drug-allergy interaction) and points to whether an order could have engendered alert fatigue or resulted in a fatality. We applied the *EHR Flight Simulator* at large medical centers and solicited feedback concerning logistical requirements and perceived utility of the assessment process. Although the *EHR Flight Simulator* is a valid approach for identifying potential risk points of CPOE systems, there is an opportunity to provide more explicit guidance and feedback to medical centers to foster continuous improvement and organizational learning. Specifically, we developed a structured feedback report focusing on CPOE configuration settings and drug mappings that medical centers could apply to address any shortcomings. In this presentation, we also address (1) logistical challenges confronted during this project (e.g., the assessment process requires front-line provider and informatics staff involvement), (2) future opportunities to address a wider range of orders placed within CPOE systems, and (3) automating aspects of the assessment process.

Acknowledgements

All participants have agreed to take part on the panel. The views expressed are those of the panelists and not necessarily those of their respective institution(s).

References

1. Institute of Medicine (US). Committee on Patient Safety and Health Information Technology. Health IT and patient safety: Building safer systems for better care. National Academies Press; 2012.
2. Harrison MI, Koppel R, Bar-Lev S. Unintended consequences of information technologies in health care—an interactive sociotechnical analysis. Journal of the American medical informatics Association.2007 Sep 1;14(5):542-9.
3. Tolley CL, Forde NE, Coffey KL, Sittig DF, Ash JS, Husband AK, Bates DW, Slight SP. Factors contributing to medication errors made when using computerized order entry in pediatrics: a systematic review. Journal of the American Medical Informatics Association. 2017 Oct 26;25(5):575-84.
4. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. Journal of the American Medical Informatics Association. 2004 Mar 1;11(2):104-12.
5. Phansalkar S, Van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, Middleton B, Bates DW. Drug—drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. Journal of the American Medical Informatics Association. 2012 Sep 25;20(3):489-93.
6. Nebeker JR, Hoffman JM, Weir CR, Bennett CL, Hurdle JF. High rates of adverse drug events in a highly computerized hospital. Archives of internal medicine. 2005 May 23;165(10):1111-6.
7. Amato MG, Salazar A, Hickman TT, Quist AJ, Volk LA, Wright A, McEvoy D, Galanter WL, Koppel R, Loudin B, Adelman J. Computerized prescriber order entry–related patient safety reports: analysis of 2522 medication errors. Journal of the American Medical Informatics Association. 2017 Mar 1;24(2):316-22.
8. Georgiou A, Prgomet M, Markewycz A, Adams E, Westbrook JI. The impact of computerized provider order entry systems on medical-imaging services: a systematic review. Journal of the American Medical Informatics Association. 2011 Mar 8;18(3):335-40.
9. Rosen MA, Tran G, Carolan H, Romig M, Dwyer C, Dietz AS, Kim GR, Ravitz A, Sapirstein A, Pronovost PJ. Data driven patient safety and clinical information technology. In Healthcare Information Management Systems 2016 (pp. 301-316). Springer, Cham.
10. Institute for Healthcare Improvement. Trigger Tool for Measuring Adverse Drug Events. Institute for Healthcare Improvement and Premier. 2004.
11. The Office of the National Coordinator for Health Information Technology. Saferty Assurance Factors for EHR Resilience (SAFER): Computerized Provider Order Entry with Decision Support. 2016.
12. Schiff G, Wright A, Bates DW, Salazar A, Amato MG, Slight SP, Sequist TD, Loudin B, Smith D, Adelman J, Lambert B. Computerized Prescriber Order Entry Medication Safety (CPOEMS): Uncovering and Learning From Issues and Errors. Sliver Spring, MD: US Food and Drug Administration. 2015.
13. Kilbridge PM, Welebob EM, Classen DC. Development of the Leapfrog methodology for evaluating hospital implemented inpatient computerized physician order entry systems.BMJ Quality & Safety.2006Apr1;15(2):81-4.

Panel: Intelligent Integrative Informatics Approaches for Big Data Aggregation, Sharing and Analytics in Stem Cell Research

Joseph Finkelstein, MD, PhD, FAMIA,[1] Fadia Shaya, MPH, PhD,[2] Kirill Borziak, PhD,[1] Ben D MacArthur, PhD,[3] Avi Ma'ayan, PhD[4]

[1]Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY; [2]University of Maryland School of Pharmacy, Baltimore, MD; [3]Schools of Mathematics and Medicine, University of Southampton, Southampton, UK; [4]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY

Abstract

Advancements in regenerative medicine have brought to the fore the need for increased standardization and sharing of stem cell product characterization to help drive these innovative interventions toward public availability. Although numerous stem cell databases exist and attempts have been made to standardize stem cell characterization, there is still a lack of a platform that incorporates heterogeneous stem cell information into a harmonized project-based framework. The aim of this panel is to introduce approaches for intelligent integration of heterogenous data sets generated in the course of preclinical and interventional stem cell research as well as discuss promising big data applications in this area targeting prediction of stem cell fate and explaining regenerative potency. Panelists from early adopter institutions will compare their experiences in aggregating and analyzing stem cell data. Key implementation issues that will be addressed by the panelist include common data elements, regulatory and ethical requirements, agile and flexible data hub development, and promising data analytics approaches.

General description of the panel

Heterogenous data streams collected during different phases of stem cell differentiation may represent a challenge for systematic integrative analysis. Harmonized integration of heterogenous data requires introduction of standardized metadata utilizing cross-linked biomedical ontologies. The goal of this panel is to foster substantive discussion of biomedical informatics community of optimal approaches to aggregate and share stem cell research data. To achieve this, the panel participants will engage the participants in an open-ended discussion using results from several ongoing NIH-funded projects as instructive use cases. The panel will cover a broad spectrum of relevant topics including (1) generalizable framework of common data elements for stem cell research; (2) regulatory and ethical issues of stem cell data sharing; (3) implementation of NIH-sponsored data hub for stem cell research; (4) machine learning applications in stem cell research; (5) an integrated library of cellular signatures and its application in systems biology. The panel will employ an interactive format with at least third of the time devoted to soliciting open-ended input from the audience.

The objectives of this panel are:

- Review multi-module framework of common data elements for stem cell research
- Learn regulatory and ethical issues of stem cell data sharing
- Compare and contrast early adopters' approach to develop and implement data hubs for stem cell research
- Understand the role of metadata and biomedical ontologies in intelligent integration of stem cell research data
- Learn major building principles of Regenerative Medicine Data Repository (ReMeDy)
- Understand design of the Library of Integrated Network-Based Cellular Signatures (LINCS)
- Formulate optimal approaches for machine learning and system biology in stem cell research

Implementing Multi-module Framework of Common Data Elements for Stem Cell Research Data Sharing. We will provide a systematic overview of the main features of available stem cell databases in order to identify specifications useful for implementation in stem cell data hubs [1]. The data elements reported in these databases represented a broad spectrum of parameters from basic socio-demographic variables to various cells characteristics, cell surface markers expression, and clinical trial results. The data hubs features consisted of the following: common data elements (CDE), data visualization and analysis tools, and biomedical ontologies for data integration. Dr. Finkelstein will present a multi-module CDE framework, which aims to capture all facets of information related to regenerative medicine trials. Our multi-module CDE framework supports a broad range of stem cell studies by introducing modules delineating CDEs from preclinical studies to randomized controlled trials.

Regulatory and Ethical Aspects of Stem Cell Research Data Sharing. Broad sharing of genomic- and health-related data requires proper governance and security [2] and will be discussed by Dr. Shaya. In the context of stem cell research, data and sample sharing represent a scientific and ethical challenge to ensure appropriate protection of individual interests as well as maintaining public trust. Effective data protection requirements are necessary along with the future data harmonization efforts for building successful stem cell research data sharing [3]. Deployment of the common framework for responsible sharing of genomic and health-related data established by the Global Alliance for Genomics and Health (GA4GH) in stem cell databases can facilitate the use of data in compliance with national and international laws and general ethical principles and standards [4].

Regenerative Medicine Data Repository (ReMeDy) for Harmonized Aggregation and Sharing of Stem Cell Research Data. To accelerate stem cell research NIH as a part of the 21st Century Cures Act. NIH has established, with the coordination of the FDA, the Regenerative Medicine Innovation Catalyst (RMIC) data hub for sharing stem cell research. The platform, called Regenerative Medicine Data Repository (ReMeDy), is an implementation of the Signature Commons, an NIH-funded project designed to store and search diverse metadata in an agile and flexible manner [5]. The platform architecture will be presented by Dr. Borziak.

Artificial Intelligence Applications to Predict Stem Cell Fate and Explain Regenerative Potency. Dr. MacArthur [6] will present a collated a library of single cell gene expression patterns from various different cell types in the hematopoietic hierarchy taken from young, adult and aged mice. The library was used to train an artificial neural network (ANN) to accurately predict both cellular identity and developmental age directly from gene expression profiles. The resulting classifier was used it to investigate division patterns of human stem cells.

Data Integration for Machine Learning for Drug and Target Discovery. Dr. Ma'ayan will present extensions to the Harmonizome project [7]. The Harmonizome, available at https://maayanlab.cloud/harmonizome, is a collection of processed datasets gathered to serve and mine knowledge about genes and proteins from major biomedical resources. To create the Harmonizome the Ma'ayan Lab extract, abstract and organizes data into functional associations between genes/proteins and their attributes. Such attributes could be physical interactions with other biomolecules, expression in cell lines and tissues, genetic associations with knockout mouse or human phenotypes, or changes in expression after drug treatment. The Harmonizome is a comprehensive resource of knowledge about genes and proteins, and as such, it enables researchers to discover novel relationships between biological entities, as well as form novel data-driven hypotheses.

Panelists

Joseph Finkelstein, MD, PhD, FAMIA is Chief Research Informatics Officer, Senior Associate Dean for Information Technology, and Professor of Population Health Science and Policy at the Icahn School of Medicine at Mount Sinai. Dr. Finkelstein is also the Director of the Center for Biomedical and Population Health informatics. Dr. Finkelstein has extensive experience in biomedical informatics with a particular emphasis on innovative health information technologies supporting collection and analysis of heterogeneous data streams, intelligent data aggregation, predictive analytics, and the conduct of clinical trials. Dr. Finkelstein is the Principal Investigator of a nation-wide regenerative medicine data hub aimed at aggregation, harmonization and analysis of clinical trial results assessing stem cell-based therapies funded by NIH as a part of the 21st Century Cures Act.

Fadia Shaya, MPH, PhD is Professor at the University of Maryland Schools of Pharmacy and Medicine, and Director of Informatics at Institute for Clinical and Translational Research, University of Maryland Baltimore. Dr. Shaya also serves as Director, Center on Drugs and Public Policy and Executive Director, Behavioral Health

Research Program at the University of Maryland Baltimore. Dr. Shaya leads digital transformation in health services research, and has built research and training capacity to support all stages of translational research, from pre-clinical trials to post-marketing surveillance. She is a member of the Food and Drug Administration (FDA) funded Maryland Center for Regulatory Science and Innovation (CERSI) and is a lead on the Patient Preference Information Group, spanning all FDA funded CERSIs, setting the structure at the FDA for building infrastructure and developing methods to incorporate patient input into drug and device development and safety and effectiveness evaluation.

Kirill Borziak, PhD is a bioinformatics expert with extensive experience in molecular biology and development of computational pipelines for big data visualization and analysis. His current research focuses on approaches for optimal reuse and harmonization of shared big data with particular focus on sequencing and expression data for hypothesis generation and knowledge discovery. Dr. Borziak oversees technical implementation of **Re**generative **Me**dicine **D**ata Repositor**y (ReMeDy)**, a NIH-funded data hub aimed at harmonized aggregation, sharing and analysis of data generated by pre-clinical stem cell research projects and interventional stem cell clinical trials.

Ben MacArthur, PhD is a Professor in the Faculty of Medicine and the School of Mathematics at the University of Southampton, and a Fellow of the Alan Turing Institute, the UK national institute for data science and artificial intelligence. His work combines experimental methods and mathematical models to investigate molecular regulation of stem cell fate. His projects resulted in ability to collect increasingly detailed information about molecular expression patterns in individual stem and progenitor cells using high-throughput single cell profiling technologies. The resulting data are complex and require advanced data analytics to fully ascertain impact of multiple interrelated factors. Dr. MacArthur's work combines modern machine learning methods with mechanistic mathematical models to do this. Mathematical models include both deterministic and stochastic mechanisms with a particular focus on cell-cell variability and its role in collective decision-making.

Avi Ma'ayan, PhD is the Director of the Mount Sinai Center for Bioinformatics and Mount Sinai Endowed Professor of Bioinformatics in the Department of Pharmacological Sciences. Dr. Ma'ayan is also Principal Investigator of the NIH-funded BD2K-LINCS Data Coordination and Integration Center and Mount Sinai Knowledge Management Center for Illuminating the Druggable Genome. The Ma'ayan Laboratory applies computational and mathematical methods to study the complexity of regulatory networks in mammalian cells. His research team applies statistical mining techniques to study how intracellular regulatory systems function as networks to control cellular processes such as differentiation, dedifferentiation, apoptosis and proliferation. The Ma'ayan Laboratory develops software systems to help experimental biologists form novel hypotheses from high-throughput data, while aiming to better understand the structure and function of regulatory networks in mammalian cellular and multi-cellular systems.

References

1. Finkelstein J, Parvanova I, Zhang F. Informatics Approaches for Harmonized Intelligent Integration of Stem Cell Research. Stem Cells Cloning. 2020;13:1-20.
2. Knoppers BM, Isasi R, Benvenisty N, et al. Publishing SNP genotypes of human embryonic stem cell lines: policy statement of the international stem cell forum ethics working party. Stem Cell Rev Rep. 2011;7(3):482–484.
3. Morrison M, Bell J, George C, et al. The European general data protection regulation: challenges and considerations for iPSC researchers and biobanks. Regen Med. 2017;12(6):693–703.
4. Bredenoord AL, Mostert M, Isasi R, Knoppers BM. Data sharing in stem cell translational science: policy statement by the international stem cell forum ethics working party. Regen Med. 2015;10(7):857–861.
5. Stathias V, Koleti A, Vidović D,et al. Sustainable data and metadata management at the BD2K-LINCS Data Coordination and Integration Center. Sci Data. 2018;5:180117.
6. Stumpf PS, MacArthur BD. Machine Learning of Stem Cell Identities from Single-Cell Expression Data via Regulatory Network Archetypes. Front Genet. 2019 Jan 22;10:2.
7. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016 Jul 3;2016:baw100.

Affirmation

This is to affirm that all proposed members have been personally contacted by Joseph Finkelstein and have agreed to participate in this panel.

Are We There Yet? Creating, Finding, and Using Better Information for Pandemics

Kate Fultz Hollis, MS, MBI[1], Nicholas Tatonetti, PhD[2], Scott McGrath, PhD[3], Karen A. Monsen, PhD, RN[4], John Lauerman[5]
[1]Oregon Health & Science University, Portland, OR; [2]Columbia University, New York, NY; [3]Providence Health and Services, Missoula MT; [4]University of Minnesota School of Nursing, Minneapolis, MN; [5]Bloomberg News, Reporter and Editor, Boston, MA

Abstract

How do we obtain and analyze better and accurate information to study and report public health emergencies such as the COVID-19 pandemic? This panel brings together translational informaticians and researchers using COVID-19 data, a public health professional working with standards for data and informing the community about COVID-19, and a prominent journalist from Bloomberg News who specializes in medical science reporting. We present both good examples of pandemic data science appearing in journals and the news as well as look where we might not be best at explaining a pandemic to a community. We describe what informaticians need: to present an integrated view of COVID-19 information resources for public information, scientific research, and clinical interventions.

Introduction

As the *New England Journal of Medicine* reported, (i)n December 2019, a cluster of patients with pneumonia of unknown cause was linked to a seafood wholesale market in Wuhan, China. A previously unknown betacoronavirus was discovered through the use of unbiased sequencing in samples from patients with pneumonia.[1] As we are biomedical informaticians, we are trained to find accurate and relevant information for diseases and use the information from the laboratory, the clinic, and the community. In 2020, we saw hundreds of database resources with information about COVID-19, coronavirus, CoV 2, and many variations of names for the coronavirus disease. Some of these resources showing data and explaining outbreaks also turned up in news sources all over the world, for example: Johns Hopkins University School of Medicine Coronavirus Resource Center[2]; COVID-19 Resources from Institute for Health Metrics and Evaluation[3]; CDC Weekly Review (https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html); and State and local departments of public health.

The purpose of this panel is to bring different perspectives on the biomedical and social determinants information we use to study and combat disease, and the special case that the COVID-19 pandemic has brought to different communities studying and fighting the disease.[4] We have data and use it from translational bioinformatics including genomics and other specialties. We are presented data from variety of sources including the biomedical sources we commonly use (PubMed, BioRxiv, medRXiv), and then there are all the news organizations like *New York Times*, STAT, *Bloomberg News*, and local news sources (and many more outside the US). There is also data we obtain from the clinic and community; and so how do we find and use all this information effectively when we treat our patients? In many ways, the COVID-19 pandemic has been a unifying situation: all of us have a need to know about the virus and all of us have an interest in what might be the best way to combat the virus and protect us from the pandemic. Are we on the right path for relevant and correct information?

This panel brings together translational informaticians and researchers using COVID-19 data, a public health professional working with standards for data and the community about COVID-19, and a prominent journalist from Bloomberg News who specializes in medical science reporting. We present both good examples of pandemic data science appearing in journals and the news as well as look where we might not be best at explaining a pandemic to a community.

Panel Outline

Many databases and articles have appeared in biomedical and informatics journals but this might be a rare occasion where the databases and science often appear in major news sources, social media and community newspapers. For

example, this table includes an extremely small subset of thousands of information sources on COVID-19 available from institutions and from news media:

Table 1. Sample Online Databases

Database Website	Database Description	Update Frequency
COVID-19 Trials Tracker, Oxford University	Tracking COVID-19 Trials and their Results	Monthly
Coronavirus Disease 2019 (COVID-19) | CDC	Information and Directions to CDC databases	Weekly
COVID Tracking Project, Atlantic Monthly Magazine	A volunteer organization from The Atlantic: collects and publishes the data to understand the COVID-19 outbreak in the United States.	Daily

With so many information resources to choose from, the panel will proceed as follows:

1. Introductions and 5 to 10 minutes remarks from each panelist about the topic of finding, using and analyzing data on COVID-19.
2. In depth questions directed to each panelist on the COVID-19 pandemic information, in particular:
    a. What information sources do you use and what have been hard to access?
    b. What sources have been the best for your work?
    c. Do you think data has been effectively shared among research or news groups in the pandemic?
    d. How can science be portrayed correctly to the public (how can informatics be portrayed)? Where have we encountered misinformation and how is science portrayed incorrectly? Are there ways we can correct misinformation?
3. The panel will end with audience participation and a call for particular questions. The moderator will start with specific questions for each panelist to start the discussion.

Panelists

The panelists will be as follows and all participants have agreed to take part on the panel:

Moderator: Kate Fultz Hollis, MS, MBI Oregon Health & Science University, Portland OR
With many years at Harvard Medical School and UCLA, Ms. Fultz Hollis has developed a strong interest in how to find accurate medical research data to answer questions. She became particularly interested in research to find relevant medical data, particularly where data is found in research studies and in electronic medical records, and then used to discover new treatments for patients. She has been extremely active for the past 7 years in editing and evaluating biomedical informatics research for MEDINFO (conference of the International Medical Informatics Association or IMIA) and she is currently one of the senior editors for IMIA's Yearbook.

Nicholas Tatonetti, PhD, Associate Professor of Biomedical Informatics, Director of Clinical Informatics, Institute for Genomic Medicine, Director of Clinical Informatics, Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY
Dr. Tatonetti directs the Tatonetti Lab at Columbia University. The lab is making drugs safer through the analysis of data. Adverse drug reactions are experienced by millions of patients each year and cost the healthcare industry billions of dollars. In the Tatonetti Lab, they use advanced data science methods, including artificial intelligence and machine learning, to investigate these medicines. Using emerging resources, such as electronic health records and genomics databases, the lab is working to identify for whom these drugs will be safe and effective and for whom they will not. In 2020, Dr. Tatonetti teamed up with Dr. Scott McGrath to produce the C19 Weekly, a weekly information show on coronavirus science in the journals and the news.

Scott McGrath, PhD, Providence Health and Services, Missoula MT
Dr. McGrath is a clinical informatics education specialist at Providence Health & Services. He teaches several remote informatics courses at Ohio University, University of New England, University of Nebraska at Omaha, and the University of Montana. In addition to serving as the chair of the AMIA Student Working group, he is also the producer of The C19 Weekly, hosted by Dr. Nicholas Tatonetti of Columbia University. They aim to provide some assistance to informaticians, data scientists, and anyone who is interested in this global pandemic, and wants to learn more. C19 Weekly has four primary goals: 1. Help to improve scientific literacy and promote good research practices; 2. Advance COVID-19 research awareness and help explain the latest papers; 3. Showcase informatics and data science contributions on the front lines of this fight; and 4. Help sound the call to action and spotlight worthy causes and collaborations to join, no matter what your skill set.

Karen A. Monsen, PhD, RN, FAMIA, FAAN, Professor and Chair, Population Health and Systems Cooperative, Director of the Center for Nursing Informatics, Director of the Omaha System Partnership, School of Nursing, University of Minnesota, Minneapolis, MN
Dr. Monsen is an internationally recognized researcher, educator, and public health nurse who leads multiple interprofessional initiatives around COVID-19 data, guidelines, resilience, and community response. She directs the Omaha System Partnership practice-based research network and works with collaborators globally to advance the use of standardized data to discover new knowledge, inform care quality, and improve population health.

John Lauerman, Bloomberg News, Reporter and Editor, Boston, MA
John Lauerman is a reporter and editor with Bloomberg News who joined in 2002, writing primarily about health and education. He's covered disease outbreaks around the world including SARS, HIV, bird flu, the 2009 swine flu and now the Covid-19 pandemic. He's also written about a wider range of health issues and events, including Hurricane Katrina, human embryonic stem cell research, the proliferation of new, immune cancer drugs, and the rise of cheap, accessible genomic analysis. Before coming to Bloomberg, he worked in communications at Harvard Medical School, was a health reporter for the Springfield Union-News in Massachusetts, and freelanced for 12 years, writing articles and books about healthy aging, diabetes and other subjects.

Importance of this Panel Discussion

We describe what informaticians need: to present an integrated view of COVID-19 information resources for public information, scientific research, and clinical interventions.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020;382(8):727-33.
2. Gardner Lea. Johns Hopkins University of Medicine Coronavirus Resource Center: Johns Hopkins Bloomberg School of Public Health; 2020 [08/14/2020]. Available from: https://coronavirus.jhu.edu.
3. Murray C. COVID-19 Resources: Institute for Health Metrics and Evaluation; 2020 [08/14/2020]. Available from: https://covid19.healthdata.org/united-states-of-america.
4. Monsen, K. A. Rapid Development and Deployment of an International Omaha System Evidence-Based Guideline to Support the COVID-19 Response, 2020. Computers, Informatics, Nursing: CIN, *38*(5), 224–226. https://doi.org/10.1097/CIN.0000000000000648.

Working Towards Shareable and Interoperable Patient and Clinician Facing Clinical Decision Support — Experiences from the Field

Roland Gamache, PhD, MBA, FAMIA[1] (Organizer); Laura Haak Marcial, PhD[2]; Kristen Miller, DrPH, CPPS[3]; Joshua E. Richardson PhD, MS, MLIS[2]
[1]Agency for Healthcare Research and Quality, Rockville, MD; [2]RTI International, Research Triangle Park, NC; [3]National Center for Human Factors in Healthcare MedStar Health, Washington, DC

Abstract
A confluence of efforts is changing clinical decision support (CDS) from siloed knowledge management efforts within single institutions to open source efforts that leverage rapidly evolving standards that enable healthcare organizations to share and implement CDS. This panel provides perspectives from the Agency for Healthcare Research and Quality and its goals for promoting a more robust CDS ecosystem through two funded projects that leverage standards-based, interoperable, and shareable knowledge artifacts in real-world settings for pain management. The panelists will each provide their unique perspectives based on their projects' activities for developing and implementing those knowledge artifacts and share how their experiences can inform future efforts to meet the needs for standards-based and interoperable CDS that support patients and providers. The speakers will also address the challenges for developing and implementing CDS that meet the needs of multiple stakeholders within the key area of chronic pain management.

Significance
Few areas in healthcare today are as pressing as managing chronic pain, defined as pain that occurs on at least half the days for six months or more.[1] It affects as many as one in three American adults,[1] and one in ten American adults suffer from chronic pain that significantly disrupts their "work, social, and/or self-care activities."[2] Furthermore, the related use of opioids is linked to co-morbidities and mortality that some have described as "twin crises" in healthcare.[3] In all, the annual costs of chronic pain range from $560 billion to $630 billion[4]—exceeding the costs from heart disease, diabetes, or cancer.

Chronic pain cases—such as those with low back pain, fibromyalgia, or tension headache—particularly affect utilization in primary care, as 52% of patients including those with chronic non-cancer pain (CNCP) receive their care in the primary care setting. The COVID-19 pandemic further complicates chronic pain and opioid management in these settings due to quarantines and lockdowns.[5] Patients in these settings benefit from providers who educate them on ways to seek appropriate care and participate in shared decision-making (SDM) that addresses patient needs, priorities, and values. Evidence includes the Centers for Disease Control and Prevention's (CDC's) Guideline for Prescribing Opioids for Chronic Pain, which calls out SDM as a key strategy for providers to educate patients about the risks and benefits of treatment options,[6] and when effectively implemented, increases patient knowledge of treatment options and aligns treatment options to patient values. Enabling partnerships between patients and providers through SDM—"therapeutic alliances"[7]—is key to delivering patient-centered care for CNCP and opioids.

Emerging open-source solutions like CDS Connect via the Agency for Healthcare and Research Quality (AHRQ), CDS Health Level 7 (HL7) CDS Hooks, clinical quality language (CQL), and SMART on FHIR are promising ways to disseminate evidence for chronic pain and opioids in CDS, but more work is needed within and across the domains to align architectures, knowledge artifacts, and implementation strategies to make shareable CDS effective for providers and patients.

This panel will describe the panelists' efforts via two AHRQ-funded projects—led by MedStar Health and RTI International—to develop and implement standards-based patient- and clinician-facing CDS that improves SDM for pain management in primary care. The panelists will use their unique perspectives from the field to discuss the

challenges and opportunities they are encountering when leveraging standards-based and interoperable CDS. Their insights will address the strategies they are employing to balance informatics standards with local EHR solutions while also meeting the user-centered needs of patients and providers.

Description of the Panel

Time	Speaker	Topic
15’	Marcial (moderator)	Moderates the panel and leads the discussion around CDS for chronic pain management
15’	Gamache	Address the broad goals and expectations of AHRQ with these two contracts
15’	Miller	Discusses MedStar Health’s project challenges, achievements, and lessons learned
15’	Richardson	Discusses RTI’s project challenges, achievements, and lessons learned
25’	All	Discussion, Q&A with audience

Learning Objectives

1. Understand the major challenges, and opportunities, to the use of interoperable and publicly-shareable CDS in healthcare organizations.
2. Learn about ability—and limits—to standards-based approaches for developing and implementing CDS;
3. Learn about technological and usability solutions that projects employ to achieve CDS that is both patient- and provider-facing for enabling SDM.

Individual Speaker Contributions

Laura Haak Marcial, PhD – Dr. Marcial will serve as the panel moderator and introduce the panel participants. In addition, Dr. Marcial will discuss the challenges and opportunities associated with patient- and clinician-facing CDS based on experiences from the AHRQ-funded, Patient-Centered CDS Learning Network, as well as the challenges and opportunities for using CDS to support SDM for chronic pain management. Later in the program, Dr. Marcial will summarize the takeaways from the panelists and lead the audience Q&A.

Roland Gamache, PhD, MBA, FAMIA – Dr. Gamache will provide background and context for the two funded efforts featured in this panel. Started in 2016, AHRQ’s CDS initiative has two broad aims: to advance evidence into practice through CDS and to make CDS more shareable, interoperable, and publicly-available. The initiative features open-source, standards-based tools (e.g., for CDS authoring) as well as a web-based repository of publicly-available CDS artifacts. Funded projects provide real-world demonstrations of applying CDS standards in the field, publish lessons learned, and disseminate artifacts for re-use and adaptation by other healthcare organizations.

Kristen Miller, DrPH, CPPS – Dr. Miller will discuss the MedStar Health CDS for Chronic Pain Management project (Tapering and Patient Reported (TAPR) Chronic Pain Management Tool" TAPR-CPM Tool). The proposed research aims to advance knowledge for patients and providers through CDS tools that enhance the quality of clinical discussion and share decision-making for optimizing pain management therapy – specifically opioid tapering. CDS for patients will help track and manage pain and daily function to support reduced opioid use while facilitating continued patient engagement. CDS for providers will help detect patients at high risk of harm from opioids while also optimizing presentation of patient data and evidence-based guidelines to support opioid tapering. She will also discuss ethical, legal, and strategic challenges of the project.

Joshua Richardson, PhD, MS, MLIS – Dr. Richardson will discuss the RTI-led CDS for Chronic Pain Management (CDS4CPM). The aims of the project are to develop, implement, and evaluate the lessons learned for disseminating publicly-accessible, standards-based knowledge artifacts for chronic pain management within primary care clinics at two academic medical centers. He will discuss the technical challenges to applying standards-based solutions to implementing CDS for SDM as per AHRQ’s SHARE Framework. Dr. Richardson will also address the challenges of balancing CDS standards with the constraints at local institutional governance and workflows.

Expected Discussion and Discussion Questions

We expect the audience will want to engage the panelists in discussions on approaches for effective implementation of patient-facing CDS, approaches to standards-based CDS, and approaches to patient-facing CDS systems. The discussion questions include:

1. How are informatics standards meeting the requirements of patient-facing CDS?
2. In what ways can projects at the level of healthcare organizations balance the availability of standards-based solutions that promote CDS interoperability while achieving electronic health record (EHR)-specific solutions that meet localized needs, e.g. workflow, usability, etc.?
3. In what ways can CDS standards adapt for people, processes, and technology need to promote adoption within the culture of medical care?

Urgent Topics for Intended Audiences

This panel addresses pressing issues around implementation at multiple levels: translating evidence into logic such as CQL and executable code; addressing challenges with balancing rapidly evolving CDS standards and local EHR environments; and lessons learned with integrating CDS into clinical workflows that address both patients' and providers' needs.

- **CMIOs and CNIOs:** those responsible for implementing CDS in healthcare organizations;
- **EHR Implementers:** implementation staff responsible for implementation and effective use of EHR tools, and system optimization**,** including patient-facing tools;
- **CDS Systems Developers:** development staff building clinical decision support tools and services for providers and patients;
- **Patients and Patient Advocates:** those who receive and/or represent the patients who receive CDS to support chronic pain care.

Attestation

The panel moderator has assurances from all participants that they will be available to participate at AMIA 2020.

References

1. Interagency Pain Research Coordinating Committee. National pain strategy: A comprehensive population health-level strategy for pain [Internet]. Washington, D.C.: Department of Health and Human Services; 2016 Mar [cited 2020 Mar 16]. Available from: https://www.iprcc.nih.gov/sites/default/files/HHSNational_Pain_Strategy_508C.pdf
2. Dahlhamer J, Lucas J, Zelaya, C, Nahin R, Mackey S, DeBar L, et al. Prevalence of Chronic Pain and High-Impact Chronic Pain Among Adults — United States, 2016. MMWR Morbidity and Mortality Weekly Report [Internet]. 2018 Sep 14 [cited 2019 Jan 28];67(36):1001–6. Available from: http://www.cdc.gov/mmwr/volumes/67/wr/mm6736a2.htm?s_cid=mm6736a2_w
3. Mackey S, Kao M-C. Managing twin crises in chronic pain and prescription opioids. BMJ [Internet]. 2019 Mar 6 [cited 2019 May 9];l917. Available from: http://www.bmj.com/lookup/doi/10.1136/bmj.l917
4. Skelly AC, Chou R, Dettori JR, Turner JA, Friedly JL, Rundell SD, et al. Noninvasive Nonpharmacological Treatment for Chronic Pain: A Systematic Review [Internet]. Agency for Healthcare Research and Quality (AHRQ); 2018 Jun [cited 2020 Mar 16]. Available from: https://effectivehealthcare.ahrq.gov/topics/nonpharma-treatment-pain/research-2018
5. Puntillo F, Giglio M, Brienza N, Viswanath O, Urits I, Kaye AD, et al. Impact of COVID-19 pandemic on chronic pain management: Looking for the best way to deliver care. Best Practice & Research Clinical Anaesthesiology [Internet]. 2020 Jul [cited 2020 Aug 25];S1521689620300562. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1521689620300562
6. CDC OPIOID GUIDELINE: Implementation Guide for Electronic Health Records [Internet]. Electronic Health Record Association; 2018 Nov. Available from: https://www.ehra.org/sites/ehra.org/files/EHRA-CDC-Opioid-Guideline-Implementation-Guide-for-EHRs.pdf
7. Report on Pain Management Best Practices: | HHS.gov [Internet]. 2019 [cited 2020 Mar 16]. Available from: https://www.hhs.gov/ash/advisory-committees/pain/reports/index.html

FHIR in the Research Continuum: The emerging Vulcan FHIR Accelerator

Charles Jaffe, MD, PhD[1]; Micky Tripathi, PhD[2]; Viet Nguyen, MD[3]; Janet Campbell, BA[4]; Clem McDonald, MD[5]; Christopher G Chute, MD, DrPH[6]

[1]Health Level 7, Del Mar, CA; [2]Arcadia, Burlington, MA; [3]Stratametrics, Salt Lake City, UT; [4]Epic, Verona, WI; [5]National Library of Medicine, NIH, Bethesda, MD; [6]Johns Hopkins, Baltimore, MD

Abstract

HL7 FHIR (Fast Healthcare Interoperability Resources) is now 10 years old. To date, it has been implemented in over 5000 sites worldwide. FHIR supports the broad continuum including patient care, population health, evolving payment models and clinical research. In March of last year, both the Center for Medicare and Medicaid Services (CMS) and the Office of the National Coordinated for Health IT (ONC) released two complementary Final Rules to improve the Interoperability of health data. The HL7 FHIR Accelerator Program has been the principle enabler of FHIR implementation. Vulcan, the newest FHIR Accelerator, is dedicated to enabling regulated and non-regulated research. This panel will explore the emergence of FHIR implementation across the broad continuum of biomedical research, regulated clinical research and population health.

Introduction

This panel will provide the attendees with the background, the rapidly evolving processes, the technical elements, and the innovative approaches to solving the complex problems of interoperable data exchange. In the last decade, FHIR[1] has been embraced by developers of technology solutions, by government regulatory bodies, by academic institutions, and by Public Health agencies worldwide. The adoption of FHIR-based solutions has been accelerated by coalescence around a single API structure. The process has been embraced by both public- and private-sector initiatives and by reliance upon a highly consistent *maturity model* and a reliable strategic roadmap.

Within the scope of this panel, we will highlight the innovative approaches to these strategic goals and articulate the framework for their solution. Unprecedented collaboration by private-sector companies and by broad based coalitions have largely refined the business model for application development. Moreover, innovative government-based initiatives have fostered the sharing of genomic data for both applied and basic research.

Perhaps the most far-reaching acceleration of FHIR adoption for research has been the embrace of the Vulcan FHIR Accelerator Program by a coalition of Biopharma, FDA, NIH, NLM, and CDC, with support from Clinical and Translational Science Awards (CTSA) program recipients.

Defining the challenges

For standards development organizations and the supporting programs that provide implementation solutions, there are major hurdles for enhancing interoperability while respecting the enormous investment in legacy systems. During the debate, in both academia and industry, some solutions have accelerated enhancements to the integration of basic and clinical science into the challenges of patient care. At the same time, public health agencies, both local and national, have benefited from innovative solutions for collecting data and reporting critical healthcare emergencies and population-based recommendations for preventative care. This is apparent in the fields of cancer care and metabolic diseases. Even more prominent is the integration of genetic and genomic data into diagnosis and treatment. In addition, government agencies are requiring standards that support the complexity of our reimbursement systems. Moreover, as our fiscal model for healthcare and research financing shifts from an outmoded and increasing costly model, the FHIR platform has been embraced as a vehicle for change.

SMART on FHIR[2]

Currently, an ONC-funded project at Boston Children's Hospital emerged to leverage the commitment to the API model of data exchange and reuse. In close partnership with HL7, SMART on FHIR emerged as a critical solution to achieving inter-system interoperability. The support for OpenAuth 2 and Secure ID, two ISO standards that enable a consistent trust framework for internet-based authentication and financial transactions and provide a much-needed layer to the emerging FHIR stack. Most critically, the partnership enabled a fabric for both cross-system and cross-

EHR platform data exchange. SMART on FHIR grew on two fronts. Not only did it become a critical enabler of future interoperability programs, such as the Argonaut Project, but it also fostered the development of an app store for production level FHIR-based applications and solutions.

The Technical Growth of HL7 FHIR

The success of the Argonaut Project[3], the landmark FHIR Accelerator, provided a clearly defined catalog of FHIR-based profiles and implementation guides that support the objectives defined in the US-centric *Meaningful Use* requirements. Because of the technical rigor and consistency of the process, the Argonaut framework has been adopted by government outside of the US, with particular attention on the FHIR-compliant APIs. To the significant impact of the Argonaut Project profiles on intersystem interoperability, the private sector collaboration has been further enhanced by the adoption of these implementation guides by the Sequoia Project Carequality initiative, built upon a national health information network and trust frame work, as well as the CommonWell Health Alliance program for patient matching. This commitment to the FHIR-enabled platform has fostered far reaching enhancement of patient matching and the concomitant reduction in matching errors and the attendant costs.

During the last two years, the broad landscape of compensation for care has evolved from a fee-for-services model to one in which compensation is driven by clinical outcomes. This value-based care scenario has been embraced by the Centers for Medicare and Medicaid Services (CMS) as well as by the private-sector payers. This had lead to the creation of the Da Vinci Project[4], in which this community of payers collaborate in a pre-competitive environment to leverage FHIR for delivering clinical data for a broad range of use cases that foster value-based care. In addition to the payer community, Da Vinci includes EHR vendors, academic health systems, and application developers. The vision of the Da Vinci consortium embraces a broad range of innovative approaches to streamlined payment systems including real-time prior authorization.

Lastly, ONC has funded HL7 to develop a critical addition to the FHIR platform for collecting bulk data for integration and analytics. Now referred to as *Bulk Data on FHIR*, the uses for this data extend far beyond payment systems, including the CMS introduction of Blue Button 2.0, which enables Medicare recipients to download their clinical data. Other Federal agencies, including the Centers for Disease Control & Prevention (CDC) envision the use of this specification for bio-surveillance, as well as morbidity and mortality reporting.

In December 2019, HL7 announced the publication of Release 4 of FHIR. Now an ANSI standard, R4 is backward compatible, more stable, and capable of incorporating additional data and information sources. At the same time, HL7 continues of the development of R5[9], expected to be published in the fall of 2020. R5 will provide more normative resources, more seamless integration of HL7 v2 and CDA specifications, as well as multi-language support and federated servers. The recognition of R4 in the Final Rules from both the Office of the National Coordinator for Healthcare IT and CMS (Centers for Medicare and Medicaid Services) in March of this year heralded a commitment for FHIR implementation worldwide.

Growing the Community of Implementers

Google Brain and Verily leverage FHIR for a host of development projects, not limited to analytics. In conjunction with several leading EHR vendors, the clinical decision support landscape if being re-imagined with the emergence of *CDS Hooks*.[5] This technology enables a broad range of clinical decision support data sources to be integrated into the point of care without leaving the EHR environment. Most critically, Apple announced that it had integrated the FHIR platform into iOS 11.3, in collaboration with now more than 500 health systems and providers[7], to potentially transform the means by which patient access their data from across multiple sources and systems.

Lastly, the **FHIR Accelerator Program** has been created to streamline the on-ramping of new FHIR implementation communities. When introduced early this year the program was comprised of already established implementation initiatives, including the Argonaut Project, the Da Vinci Project and the CARIN Alliance. Throughout the year, additional communities including, Gravity, supporting Social Determinants of Healthcare, the Consortium for Agile Genomics, as well as Codex, the diverse oncology consortium, have all implemented FHIR to advance care delivery. Most recently, the Vulcan Project[8], which has aligned Biopharma, FDA, NIH, NLM, and the CDC has encouraged

FHIR adoption for clinical recruitment, patient reported outcomes (PRO), post-marketing bio-surveillance, and real-world clinical trials (RWCT).

Conclusions

At the conclusion of this interactive panel, participants will be able to 1) articulate the value model of the HL7 FHIR platform for patient care, applied research, population health, or patient engagement; 2) formulate technical, business, and workflow strategies that enable the integration of open APIs (Application Programing Interfaces) and HL7 FHIR resources to enhance interoperability initiatives and data integration; 3) exploit the ease of FHIR implementation to enhance technical strategies, reduce development time, decrease project implementation costs. Finally, they will be able to evaluate the early impact of the emerging Vulcan Project on a broad research landscape.

1. FHIR: Fast Healthcare Interoperability Resources (http://www.hl7.org/implement/standards/FHIR-Develop/?ref=learnmore)
2. SMART on FHIR. http://smarthealthit.org/smart-on-fhir/
3. Argonaut Project-HL7. http://argonautwiki.hl7.org/index.php?title=Main_Page
4. Da Vinci Project. http://www.hl7.org/about/davinci/index.cfm?ref=common
5. CDS Hooks. https://github.com/argonautproject/cds-hooks/wiki/Introduction-to-CDS-Hooks-and-the-patient-view-hook
6. FHIR integration into Apple iPhone. http://www.healthcareitnews.com/news/apple-launch-health-records-app-hl7s-fhir-specifications-12-hospitals
7. Health systems deploying the FHIR-based Apple iPhone platform. https://support.apple.com/en-us/HT208647
8. Vulcan FHIR Accelerator Program. https://confluence.hl7.org/display/VA/Vulcan+Accelerator+Home
9. FHIR R5 Roadmap. https://onfhir.hl7.org/2019/01/20/fhir-r5-roadmap/

Pandemic Informatics: Tuning Expectations of Real World Data – Lessons Learned from the National COVID Cohort Collaborative (N3C)

Kristin Kostka, MPH[1,2], Michele Morris, BA[3], Matvey Palchuk, MD, MS[4], Emily Pfaff, MS[5], Robert Miller, MS[2,6]

[1]Observational Health Data Sciences and Informatics, New York, NY; [2]Real World Solutions, IQVIA, Cambridge, MA; [3]University of Pittsburgh, Pittsburgh, PA; [4]TriNetX, Cambridge, MA; [5]NCTraCS Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC;[6] Tufts Clinical and Translational Science Institute, Boston, MA

Abstract

The first case of a novel coronavirus, subsequently named SARS-CoV-2, was detected in Wuhan, Hubei Province, China in 2019. By the end of August 2020, the coronavirus has since spread across the world, causing over 25 million cases of COVID-19 (the disease caused by SARS-CoV-2) and over 844,000 deaths. The use of real world data is important piece of understanding the epidemiology of COVID-19, the natural history/severity of disease and potential therapies. The National COVID Cohort Collaborative (N3C), sponsored by the National Center for Advancing Translational Sciences (NCATS), is a multi-site collaborative learning health network designed to overcome barriers to rapidly build a scalable infrastructure incorporating multi-organizational clinical data for COVID-19 analytics. This panel is composed of informaticians supporting the harmonization of COVID-19 data for downstream analytics. Here we will discuss the need to balance pragmatism versus perfectionism in informatics projects during a pandemic.

Mobilizing a Global Response in a Pandemic: The Power of Common Data Models – Kristin Kostka, MPH

The first case of a novel coronavirus, subsequently named SARS-CoV-2, was detected in Wuhan, Hubei Province, China in 2019. The coronavirus has since spread across the world, causing over 25 million cases of COVID-19 (the disease caused by SARS-CoV-2) and nearly 844,000 deaths, according to the Johns Hopkins University Center for Systems Science and Engineering.[1] In late March 2020, the Observational Health Data Sciences and Informatics (OHDSI) community, a global network that uses large-scale analytics on real-world data (RWD), annual European research symposia was cancelled due to the COVID-19 pandemic. Instead, organizers repurposed the allotted time to conduct a virtual "study-a-thon." In the span of 88 hours, more than 350 epidemiologists, statisticians, clinicians, and software engineers from around the world worked remotely to conduct rapid retrospective observational research studies to answer various research questions related to the pandemic.[2] Following this intensive 4-day exercise, the OHDSI community continues to run COVID-19 research studies on more than 18 databases from 5 countries with COVID-19 case data.[3] In May 2020, the OHDSI community joined forces with i2b2/ACT, TriNetX and PCORNet to participate in the National COVID Cohort Collaborative (N3C). N3C's goals are to demonstrate that a "multi-site collaborative learning health network can overcome barriers to rapidly build a scalable infrastructure incorporating multi-organizational clinical data for COVID-19 analytics.[4] Through weekly workgroups, subject matter experts from these common data model communities are faced with the toughest challenge: translating research questions into what is feasible in local data. This expertise is paramount in steering the development of designing COVID-19 cohort definitions and subsequent data extraction scripts to drive communal research.

Bridging COVID data in EHRs to CDMs: Perspectives from the i2b2/ACT Network – Michele Morris, BA

The ACT Network (SHRINE) is a real-time platform allowing researchers to explore and validate feasibility for clinical studies across the NCATS Clinical and Translational Science Award (CTSA) consortium, from their desktops. The ACT Network (SHRINE) helps researchers design and complete clinical studies, and is secure, HIPAA-compliant and IRB-approved. The ACT Network leverages SHRINE, or the Shared Health Research Information Network, to support multi-site research projects by enabling study feasibility/cohort discovery at partnered institutions.[5] Through the N3C program, ACT has developed COVID-19 specific ontologies to streamline and standardize local electronic health record (EHR) data into the ACT schema. This work allows local sites to have variability in lab coding systems and local condition coding variations while still having a way to be queried by the broader network for multi-center studies. ACT provides a conduit to show how EHR data can be wrangled and leveraged to answer pressing questions. During routine N3C data quality checks, ACT found a need for integration of UMLS coding into the N3C shared ontologies to increase capture of qualitative lab results captured by the program.

ACT continues to evaluate site-level challenges in harmonizing data feeds and works across CDMs to create informatics decisions that scale between models.

Streamlining Research Ready Data: Perspectives from TriNetX – Matvey Palchuk, MD, MS

TriNetX is a global health research network that optimizes clinical research and enables discoveries through the creation of real-world evidence.[6] By partnering with TriNetX, local sites receive TriNetX hardware and have hands-on support from TriNetX staff in the transmission of COVID-19 data for the N3C program. In this capacity, TriNetX is taking the disparate feeds of EHR modules and streamlining these data into a single format transmitted for consumption by the N3C data ingestion pipelines. Where available, TriNetX is already harmonizing information by supporting the breadth and depth of HL7 objects and use of common terminologies like RxNorm. This provides a significant value to the N3C program by cleaning up the enumeration of encounter-types in local EHR feeds, de-duplicating patient records and flagging places where records are connected to other clinical attributes. Even still, there are site-level challenges in capturing and pulling the information that N3C researchers request. TriNetX and the N3C program regularly work together to understand what's feasible at each site.

Enforcing Order in a Network: Perspectives from PCORNet – Emily Pfaff, MS

PCORnet is a "network of networks" that brings together patients, clinicians, researchers, health plans, and healthcare systems to share information and participate in research.[7] All the core data elements needed to support COVID-19 research and surveillance have a home in the PCORnet CDM. However, data partners may need to prioritize loading them. Current expectations within PCORnet are that partners refresh their CDM every quarter and run a comprehensive data quality assessment (January, April, July, October). To meet the demands of the pandemic, PCORnet also created a rapidly refreshed stand-alone version of the CDM that includes coronavirus patients plus other patients with respiratory illnesses since January 2020.[8] The goal for PCORnet is to characterize the cohort of COVID-19 patients and provide detailed information on demographics and pre-existing conditions. As the N3C program evolves, PCORnet sites are receiving feedback from the data ingestion pipeline to understand opportunities for increasing standardization of local data elements. In this mechanism, PCORnet sites are even being notified when variables are clinically divergent from other sites in the N3C program to ensure that this variation is actually an artifact of local care delivery and not a systems issue. This can be challenging to debug and creates an iterative loop of prioritizing which fixes are necessary to generate high quality informatics projects.

Balancing Perspectives, Moderator – Robert Miller, MS

As a site lead supporting multiple multi-center research projects, Mr. Miller brings extensive first-hand expertise on the challenges with designing extract-transform-load (ETL) processes for institutional data marts. Mr. Miller provides a neutral voice to balance panelist perspectives and ensure the discussion is rooted in a pragmatic evaluation of what can be done with the resources available. For data to be useful in research, they have to be standardized across systems. In particular, a critical emphasis will be placed on "low hanging fruit" (what's easiest for a site under duress of a pandemic to reliably make available for research) versus data elements which are inherently difficult to standardize because of care variation, artifacts of overburdened healthcare systems (e.g. the loss of consistency in information when care settings are improvised to meet demand), and other challenges that are unavoidable in a pandemic.

Questions

- How does one develop a universal phenotype when case definitions are variable by geography and biased by local testing strategies?
- For COVID-19 cases, what data elements are the most challenging to standardize across a network? What, if any, data elements are the easiest?
- How does a site who wants to bring their data into a research network like this go about the process to implement a common data model? What kind of effort is involved?
- How do ETL processes evolve as source vocabularies / ontologies evolve?
- Healthcare providers have an urgent need to understand the safety and efficacy of the various therapies being used to treat COVID-19. What's feasible today? What's not?

Acknowledgements

The work presented in this panel reflects the collaboration of teams of individuals from across CD2H, Observational Health Data and Informatics (OHDSI), NCATS Accrual for Clinical Trials (ACT) Network, Patient-Centered Clinical Research Network (PCORnet) and TriNetX organizations. The panelists would like to acknowledge the contributions

of Christopher Chute, Davera Gabriel, Harold Lehmann, Tricia Francis, Stephanie Hong, Xiaohan Tanner Zhang and Richard Zhu representing Johns Hopkins University; Clair Blacketer of Janssen; Lora Lingrey of TriNetX; Shyam Visweswaran representing the University of Pittsburgh; Marshall Clark, Kellie Walters, Adam Lee, Evan Colmenares, Robert Bradford representing the University of North Carolina; Lisa O'Keefe representing Northwestern University; Karthik Natarajan representing Columbia University; Andrew Williams representing Tufts University; Charles Yaghmour and Smita Hastak of Samvit Solutions; Raju Hemadri of Digital Infusion; Tell Bennett representing University of Colorado; Joel Saltz and Richard Moffitt representing Stony Brook University; Anita Walden, Andrew Neumann, Connor Cook and Melissa Haendel representing Oregon Health & Science University; and Ken Gersing of NCATS. This work has been funded through the National Center for Advancing Translational Sciences, National Institutes of Health, under award number U24 TR002306. Ms. Kostka, the panel organizer, has an affirmation from each of the participants that each agree to participate on this panel.

References

1. Center for Systems Science and Engineering (CSSE). COVID-19 Global Dashboard [Internet]. 2020 [cited 2020Aug30]. Available from: https://coronavirus.jhu.edu/map.html.
2. Sachson C. 88 Hours: OHDSI's Signature Moment [Internet]. 2020 [cited 2020Aug30]. Available from: https://www.ohdsi.org/88-hours/
3. Lane JCE, Weaver J, Kostka K, Duarte-Salles T, Abrahao MTF, Alghoul H, et al. Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. The Lancet Rheumatology. 2020Aug21
4. Haendel M, Chute C. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. JAMIA Open, 2020; [in press].
5. i2b2 / ACT Network (SHRINE) [Internet]. CTSI. 2020 [cited 2020Aug30]. Available from: https://www.ctsicn.org/i2b2-shrine-act
6. TriNetX. 2020 [cited 2020Aug30]. Available from: trinetx.com
7. PCORnet [Internet]. The National Patient-Centered Clinical Research Network. 2020 [cited 2020Aug30]. Available from: https://pcornet.org/
8. Carton TW, Marsolo K, Block JP. PCORnet COVID-19 Common Data Model Design and Results. PCORNet; 2020.

Precision Medicine facilitated through Registry Science: The Challenges and Solutions Associated with Constructing National-breadth Registries by Aggregating a Multi-faceted Patient-focused Datasets

Steven E. Labkoff, MD, FACP, FACMI, FAMIA[1], Leon Rozenblit, JD, PhD[2],
Claudio Faria, Pharm.D., MPH, Kathleen Hewitt, DNP, RN, CPHQ[4]
[1]The Multiple Myeloma Research Foundation, Norwalk, CT, [2]Prometheus Research, an IQVIA company, New Haven, CT, [3]Alexion Pharmaceuticals, Boston, MA,
[4]The ASH Research Collaborative, Washington, DC

Abstract:
Large data sets with disparate kinds of data are needed to facilitate precision medicine. Data such as patient journeys, outcomes, genomics, immunologic, and proteomic data - joined with real-world evidence data from EHRs and claims are needed to facilitate research. Many organizations have taken up the challenge to help generate these data sets via the construction of longitudinal registries focused on a single disease state. This panel will highlight experiences, challenges, and solutions in bringing together data sets for research. The discussion will focus on informatics, data science, outcomes research, legal, regulatory and social challenges.

Discussion:
One of the largest challenges to precision medicine is that of having sufficient numbers of patients in representative data sets that focus on the disease in question from multiple perspectives. One approach to this challenge is to create disease registries meant to focus on the specific disease state - then aggregate data from various perspectives such as EHR, genomic, proteomic, transcriptomic, patient-reported outcomes, immunologic data, and medical claims. Despite the enormous investment in data generation that resulted from EHR adoption (as an example), major challenges and hurdles continue to exist in creating high-utility data sets. When complete, they do afford the ability for scientists to gain a multi-faceted perspective on all aspects of patient care, their disease journey, and outcomes. By having a longitudinal registry, aspects of temporality can be taken into account for these studies, such as the construction and analysis of patient journeys.

The construction of these registries, however, brings into focus many of the enormous technical, legal, regulatory, and privacy challenges needed to curate such data sets. While the benefit may seem obvious, laws and regulations created in previous eras for local protection of business interests can interfere with macro programs of this nature today. In addition, the variable implementation of current HIT and other data standards also makes alignment and growth of these tools challenging. When you add a pandemic to the mix, all manner of reliable business relationships can become unglued, causing even more obstacles.

Each speaker will discuss his/her experiences in building registries with national or international footprints focused on precision medicine and public health. We draw upon viewpoints from informatics companies, life science firms, professional societies, and patient support foundations. Each speaker is leading large, national efforts to build registries and will discuss how they faced challenges in getting

these registries constructed, sustained, engaged with by participants, and utilized by researchers. The execution of programs focused on generating such data sets is critical to finding therapies and cures for diseases, especially rare diseases. Understanding these hurdles may make it easier for others to construct these resources for the larger research, clinician and patient communities.

Steven Labkoff, MD, is the Chief Data Officer of the Multiple Myeloma Research Foundation. In his role, he is the program sponsor of the CureCloud Direct-to-Patient Registry (CC-DTP). This registry is working to create a longitudinal, linked data set comprising (at launch) of four different data types – including patient donated data (surveys), EHR clinical data (abstracted into a 175-field data dictionary), a 70-gene myeloma panel, and nurses notes. Eventually, the registry will also include patient-reported outcomes, immune profiling, and medical claims – all linked to a unifying patient identifier. The CC-DTP has run up against a myriad of challenges including legal, privacy, regulatory, and data governance issues. However, the most complex and daunting challenge for this registry is the fact that it is direct-to-patient – with a goal of returning data to patients. Because of this constraint, all lab tests run on behalf of the registry (the 70-gene MM panel) must be run on a CLIA-validated (informatics) pipeline. Some of the issues that arose in building out the CC-DTP included dealing with telemedicine laws as they related to the CLIA pipeline in a direct-to-patient registry.

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

Claudio Faria, Pharm.D., MPH is currently the Executive Director of Global Health Economics and Outcomes Research at Alexion Pharmaceuticals. He has 18 years of experience in the area of Real-World Data & Evidence - in particular expertise in observational methodology, the development of patient registries, and the use of registry/registry-like data for synthetic arm trials. Claudio spent the last 13 years in various HEOR Leadership roles within the Pharmaceutical/Biotech sector - where he has developed and analyzed registry data across a multitude of therapeutic areas mainly for regulatory and reimbursement decisions. He has also had the opportunity to teach Research Methods and Pharmacoepidiemology at the UMass Medical School, Northeastern University and the Massachusetts College of Pharmacy. He is well-published, with over 200 peer-reviewed publications. He is also adjunct faculty at Rutgers University and Thomas Jefferson University.

Kathleen Hewitt, DNP, RN is the Director of the ASH Research Collaborative (ASH RC), a non-profit organization established by the American Society of Hematology (ASH) to improve the lives of people affected by blood diseases by fostering collaborative partnerships to accelerate progress in hematology. The foundation of the ASH RC is its Data Hub, a technology platform that facilitates the exchange of information by aggregating research-grade data on hematologic diseases. Prior to joining the ASH RC Dr. Hewitt served as Associate Vice President at the American College of Cardiology where she oversaw the strategic direction of ACC's NCDR, Accreditation Services, and National Quality Campaign programs. During her leadership, the ACC's quality portfolio, including the NCDR, grew into an unprecedented 3,000 hospital and healthcare system network.

References:

Gliklich, R. E., Leavy, M. B., & Dreyer, N. A. (2020). Patient registries. In Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 4th edition. Agency for Healthcare Research and Quality (US).

Gliklich, Leavy, and Dreyer, "Registries for Evaluating Patient Outcomes"; Gliklich, Leavy, and Dreyer, "Tools and Technologies for Registry Interoperability, 2nd Addendum of Registries for Evaluating Patient Outcomes, A User's Guide, 3rd Ed."

Richesson R.L., Rozenblit L., Vehik K., Tcheng J.E. (2019) Patient Registries for Clinical Research. In: Richesson R., Andrews J. (eds) Clinical Research Informatics. Health Informatics. Springer, Cham. https://doi.org/10.1007/978-3-319-98779-8_13

Unlocking clinical concepts embedded in unstructured text to advance COVID-19 Analytics for National COVID Cohort Collaborative (N3C)

Hongfang Liu, PhD[1], Rafael Fuentes[2], Justin Guinney, PhD[3], Sijia Liu, PhD[1], Peter Szolovits, PhD[4], and Hua Xu, PhD[5]

[1] Mayo Clinic, Rochester, MN; [2] National Institute of Health, Bethesda, MD; [3] Sage Bionetworks, Seattle, WA; [4] Massachusetts Institute of Technology, Boston, MA; [5] The University of Texas Health Science Center at Houston, Houston, TX

Abstract

The National COVID Cohort Collaborative (N3C) aims to assemble a multi-site learning health infrastructure with electronic health record (EHR) data for a nationwide cohort made available for COVID-19 analytics. One known challenge of EHR-based observational studies is that detailed patient information required by a study often resides in clinical narratives. Various natural language processing (NLP) technologies have been investigated to accelerate the use of clinical narratives for rapid clinical research. This panel describes a collaborative effort among Clinical Data to Health (CD2H), Open Health Natural Language Processing (OHNLP), and Observational Health Data Sciences and Informatics (OHSDI) NLP towards a national clinical NLP ecosystem.

Background

The wide adoption of electronic health record (EHR) systems has accumulated large amounts of patients' clinical data, which becomes an enabling resource for clinical and translational research such as evidence generation for clinical characterization and treatment outcomes. One known challenge of secondary use of EHRs for clinical research is that detailed patient information required by a study (e.g., symptoms, social determinants, mental status related to COVID-19) often resides in clinical narratives. Therefore, natural language processing (NLP) technologies have been extensively investigated in the medical domain to accelerate the use of clinical narratives for rapid clinical research. Despite the successes of open source NLP tools as well as their demonstrated applications , the adoption rate of those open source NLP tools in other major CTSA sites is still slower than expected. In reality, delivering practical NLP solutions for clinical research requires the existence of an institutional text analytics infrastructure that can be challenging to develop due to the incentive and regulatory complexity.

The NLP effort in National COVID Cohort Collaborative (N3C) is a collaboration among Clinical Data to Health (CD2H), Open Health Natural Language Processing (OHNLP), and Observational Health Data Sciences and Informatics (OHSDI) NLP with the goal to build a national ecosystem towards clinical NLP for COVID-19 analytics . The N3C NLP working group aims to understand barriers of establishing text analytics infrastructure and develop solutions for different stakeholders (e.g., NLP developers or consumers) to share, test, customize, and deploy diverse NLP tools by defining appropriate technologies, standards, and governance strategies. We hypothesize that through community engagement and team science, we can better enhance NLP capabilities across CTSA sites and accelerate the use of clinical narratives for clinical research.

Objectives

Our objectives in the N3C NLP project include:

Objective 1. To work with CD2H Sandbox and build a prototype of an open NLP ecosystem for sharing, testing, integrating, and deploying diverse types of NLP algorithms. Specific activities include:

1. conducting capability and needs analysis through user surveys and interviews leveraging CD2H/N3C and iEC text analytics working group,
2. defining and building appropriate architectures/technologies to accommodate diverse NLP tools and different use scenarios,
3. standardizing NLP output representations (e.g., by leveraging OMOP CDM) to improve interoperability and integration among different tools,

4. developing governance structure and strategies for collaborative activities in the open platform, and
5. organizing training sessions and prepare education materials.

Objective 2. To develop benchmark datasets and tools for COVID-19 and use them to drive the development and evaluation of the open NLP ecosystem. Specific activities include:

1. assembling a proof-of-concept collection of COVID-19 related deidentified clinical narratives from multiple sites through community participatory,
2. establishing community engagement and crowdsource environments,
3. collecting, prioritizing, and annotating data elements to be extracted from the clinical text,
4. developing COVID-19 NLP algorithms leveraging the open NLP ecosystem, and
5. encouraging community adoption of the open NLP ecosystem (e.g., organizing shared tasks using the annotated corpus).

Objective 3. To streamline community engagement and participation regarding NLP needs, which in turn will make NLP more accessible to a wider base of researchers and developers. Specific activities include :

1. creating an onboarding and registration system, allowing users and groups to seamlessly join a listserv for community engagement,
2. creating a polling system, which in turn will allow the community to ask questions and receive answers in a templated format, and
3. creating recurring meetings to allow community members to discuss a variety of topics.

Panelists' Presentations

Dr. Hongfang Liu is a professor in biomedical informatics, and currently leading the biomedical informatics division at Mayo Clinic. Her primary research focus is to facilitate the use of clinical NLP for the secondary use of EHR data for clinical and translational science research and health care delivery improvement. Additionally, she has extensive collaborative research experience in cancer genetics and molecular pharmacology. She has given lectures and demos in various settings on clinical NLP including local graduate programs and CTSA. More info: http://www.mayo.edu/research/faculty/liu-hongfang-ph-d/bio-00055092

Dr. Liu will moderate and provide the overview of the N3C NLP project.

Dr. Justin Guinney is the Vice President of the Computational Oncology group at Sage Bionetworks. His scientific career has been focused on the design and application of computational methods for translational cancer medicine. His current research specializes in integrative data analysis for prognostic and predictive modeling of cancer outcomes and response to therapy, with an emphasis in colorectal cancer. Dr. Guinney is leading the N3C Collaborative Analytics workstream which is responsible to develop the N3C Enclave. More info: http://bime.uw.edu/faculty/justin-guinney/

In this presentation, Dr. Guinney will introduce his experience developing systems and instracture for benchmarking informatics tools, and how this is being used to crowd-source the development of an NLP ecosystem for the CTSA community.

Dr. Peter Szolovits is Professor of Computer Science and Engineering in the MIT and leading the Clinical Decision-Making Group within the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). His research centers on the application of AI methods to problems of medical decision making, natural language processing to extract meaningful data from clinical narratives, and the design of information systems for health care institutions and patients. He has worked on problems of diagnosis, therapy planning, execution and monitoring for various medical conditions, computational aspects of genetic counseling, controlled sharing of health information, privacy and confidentiality issues in medical record systems, and integration of clinical and genomic data for translational medicine. More at: http://people.csail.mit.edu/psz/

In this presentation Dr. Szolovits will discuss privacy and confidential issues in medical record systems and his effort on the application of AI methods for the de-identification of clinical documents.

Dr. Hua Xu is a professor at the School of Biomedical Informatics in The University of Texas Health Science Center at Houston (UTHealth). He directs the Center for Computational Biomedicine at UTHealth. Dr. Xu's primary

research interest is to develop NLP methods and systems and apply them to clinical and translational research. He has worked on different clinical NLP topics, including syntactic parsing, word sense disambiguation, and active learning and has built multiple clinical NLP systems including the medication information extraction tool MedEx and a recent comprehensive clinical NLP system CLAMP. Methods and tools developed in Dr. Xu's lab have been widely used in large clinical consortia. Currently, he is the Chair of NLP working group at OHDSI. More at: https://sbmi.uth.edu/faculty-and-staff/hua-xu.htm.

Dr. Xu will present OHDSI NLP working group's effort on standardizing textual data in OMOP CDM using NLP, by introducing stanards/tools and their applications specific to COVID-19.

Dr. Sijia Liu is an Informatics Specialist II at the Department of Health Sciences Research, Mayo Clinic at Rochester, Minnesota. Working closely with Dr. Hongfang Liu on many NIH funded grants, his primary interest falls under the application of clinical NLP methods and the implementation of clinical NLP systems to support clinical decision making. He received his PhD in Computer Science and Engineering at University at Buffalo, The State Unviersity at New York in February 2019. He works with Dr. Hongfang Liu on managing the text analytics service line at Mayo Clinic via triaging and intaking tasks, providing edutional sessions and balancing workloads among the text analytics team. More at http://sijialiu24.com/.

Dr. Sijia Liu will discuss the design and implementation of N3C NLP infrastructure, including a cloud-based NLP engine builder and end-to-end NLP pipeline for N3C, on behalf of the N3C NLP development team.

Mr. Rafael Fuentes is a computer scientist, with a research focus on Advanced Analytics, Aritificial Intelligence, and Emerging Technologies. Using these specialties and experiences, Rafael has worked with a variety of Government and private organizations to design and implement a variety of AI and NLP applications. Besides development, Rafael has led large AI, Data Analytics, and NLP development projects across the government. Currently, Rafael is providing services to NCATS, with the primary goal to empower groups and individuals to design and deveop analytics, visualization, and NLP applications using COVID-19 datasets. Rafael is driving and encouraging collaboration among these groups with the goal to further enhance the applications and insights these groups and organizations produce.In this presentation, Rafael will discuss overall effort to encourage community participation, and empower engagement from a variety of groups organiations, and individuals.

List of discussion questions to enhance audience participation

- What are the key changes in the field that have enabled the current advances in clinical NLP?
- What are the primary gaps that need to be resolved?
- Who are the key players in this field?
- What tasks are involved in implementing an NLP system at each institution?
- What are the barriers and missing opportunities of clinical NLP research?
- Who are the key stakeholders to engage?
- How can we as a community to advance clinical NLP to support clinical and translational research?

Plan for interaction between panelists and the audience

The panelists will interact with audience through questions and commentaries discussing and reflecting on challenges that current medical environment and informatics infrastructure faces to implement patient stratification.

The panelist will take audience questions individually at the end of their individual presentations and will take questions as a group after all the panelist presentations. More than a third of the session will be reserved for audience participation. All the presenters have agreed to take part on the panel.

Lessons Learned from Healthcare Organizations Contributing Clinical Data to the National COVID Cohort Collaborative (N3C)

Stephane M. Meystre, MD, PhD[a], Ramkiran Gouripeddi, MBBS, MS[b], Jeremy Harper, MS[c], Jeffery Talbert, PhD[d]

[a] Medical University of South Carolina, Charleston, SC
[b] University of Utah, Salt Lake City, UT
[c] Regenstrief Institute, Indianapolis, IN
[d] University of Kentucky, Lexington, KY

Abstract:
The COVID-19 pandemic was officially declared by the World Health Organization in March 2020, after initial cases declared in China and a worldwide expansion. The first case in the U.S. was confirmed in January and a rapid expansion to all 50 U.S. states followed. We have seen a scattered approach for data collection and public data sets not being adequately available to explain the disease progression and key insights remaining unavailable to the public. To enable data sharing and collaborative research focused on COVID-19 across healthcare organizations in the U.S., the National COVID Cohort Collaborative (N3C) was created in the Spring of 2020 with support from NCATS and a focus on CTSA program hubs. It fostered a rapidly growing collaborative network of healthcare organizations and research communities. At the end of August 2020, more than 20 healthcare organizations were already sharing clinical data with N3C regularly, To share clinical data with N3C, participating healthcare organizations have to go through several steps and ensure availability of clinical data in a selection of data models (OMOP CDM, PCORnet, ACT, or TriNetX). This panel features speakers from four academic healthcare organizations currently sharing clinical data with N3C. They will tell about their institution and practical experiences, ideas and advices for healthcare organizations already sharing or planning to share clinical data with N3C.

Introduction:
The Coronavirus Disease 2019 (COVID-19) pandemic was officially declared by the World Health Organization (WHO) on March 11, 2020, after initial cases declared in China and a worldwide expansion. The first case in the U.S. was confirmed January 21 and a rapid expansion to all 50 U.S. states followed, with about 5.9 million confirmed cases and more than 180,000 deaths as of August 27, 2020.[1] We have seen a scattered approach for data collection and public data sets not being adequately available to explain the disease progression and key insights remaining unavailable to the public.

To enable data sharing and collaborative research focused on COVID-19 across healthcare organizations in the U.S., the National COVID Cohort Collaborative (N3C) was created in the Spring of 2020 as a partnership between the National Center for Data to Health (CD2H), healthcare organizations and subject matter experts with support from NCATS.[2] N3C focuses on Clinical and Translational Science Award (CTSA) program hubs and fostered a rapidly growing collaborative network of healthcare organizations and research communities with more than 600 individuals and 100 organizations. At the end of August 2020, more than 20 healthcare organizations were already sharing clinical data with N3C regularly, and more than 50 were at various stages of the administrative and legal process required before sharing clinical data.

Current N3C efforts are organized in several work groups corresponding to the overall organization (Data Partnership and Governance), data acquisition and sharing (Phenotype and Data Acquisition group), aggregation and cleaning (Data Ingestion and Harmonization group) and finally making the data available to the broader research community through the N3C 'Enclave' (Collaborative Analytics group).

To share clinical data with N3C, participating healthcare organizations have to go through several steps listed below. The N3C Phenotype and Data Acquisition group has already developed and shared detailed documentation and tools to ease this process, but local clinical data representation and organization specificities, legal requirements and sometimes limited available resources all potentially contribute to making this process non-trivial.

- Contact with N3C and enrollment in information sharing resources (GitHub, Slack, Google documents)
- Data transfer agreement (DTA) execution
- Single IRB reliance (John's Hopkins University) or local IRB approval
- Availability of local collaborators with the required database and Python or R knowledge
- Local clinical data represented using a compatible data model (Observational Health Data Sciences and Informatics (OHDSI) OMOP Common Data Model (CDM), Accrual to Clinical Trials (ACT), PCORnet, or TriNetX).
- Clinical data included in the N3C COVID-19 Phenotype is available
- All clinical data required to run the SQL scripts or Exporter tools is available
- Data extraction, cleaning and preparation for sharing with N3C
- SFTP credentials acquisition
- Extracted local data sharing, preferably 1-2 times every week

Panel overview:
This panel will focus on activities lead by the Phenotype and Data Acquisition N3C group, activities resulting in the definition of COVID-19 phenotypes, data models and standard terminologies to use, and tools to ease data querying, preparation and sharing with N3C. Discussions will aim at broad sharing of practical experiences, ideas and advices for healthcare organizations sharing clinical data with N3C, along with the viability of this approach for other disease states for their own organizations.

Learning objectives: During and after this session, participants should be better able to:
- Contrast local characteristics and challenges of clinical data sharing with N3C.
- Evaluate practical options for efficient local data preparation and sharing.
- Establish a collaboration with N3C and share their local clinical data.

Intended audience: This panel is adressed to professionals with activities and interests in clinical data sharing for large-scale collaborative research. It will mostly interest professionals planning to contribute to the COVID-19 pandemic response and share local patients clinical data with N3C.

Expected discussion and strategies to engage the audience: The panel moderator and presenters will start with presentations of key challenges and experiences with sharing clinical data with the N3C and ask the audience questions related to their presentation and how it relates with the audience's experience. Panel moderator and presenters will invite the audience to share and discuss their own experiences and how they relate to the presentations.

Panel organizer and participants:

Stephane Meystre will moderate this panel and introduce **MUSC**'s experience with clinical data collection including information extracted from clinical notes using Natural Language Processing (NLP), data cleaning and preparation, and sharing with N3C. To help assess the local extent of the COVID-19 pandemic and support patient care as well as research, a new database was created at MUSC in March 2020, along with an NLP-based COVID-19 information extraction tool enriching this database. This database and extracted COVID-19 related information is used for operations and clinical care guidance,[3] for testing results prediction[4] enabling data-driven decision support and for testing optimization. It was also used as a key resource for sharing MUSC clinical data with N3C, along with a local ACT database. Options based on PCORnet and the OMOP CDM were also assessed.
Dr. Meystre, MD, PhD, FACMI, FIAHSI, is Professor and SmartState Chair in Translational Biomedical Informatics at MUSC (Charleston, SC) with research activities focused on easing access to clinical data for research and clinical care purposes, using techniques such as NLP for information extraction and automated de-identification.

Ramkiran Gouripeddi will share the **University of Utah** (UU) experience in participating in the N3C. The UU was a vanguard site in the N3C data sharing effort. It provided its clinical data as per the N3C COVID-19 phenotype and was extracted from the PCORnet common data model using the Python scripts provided by the N3C Data Ingestion team. Data extraction was periodically updated based on inputs from the Ingestion team and regular data assessment checks were performed to improve the quality of the submitted data. In addition to data sharing, the informatics team at UU is coordinating a University wide effort for participating in the N3C research opportunity.

This includes regular presentations on N3C, one-to-one consults, and planning, performing and collaborating on research opportunities. In addition to supporting the researchers of the University, the informatics team is also developing state-of–the-art informatics artificial intelligence methods to assess the sequence of events in COVID-19 disease progression, the effect of risk factors such as type 2 diabetes, and the role of environmental and social determinants.
Dr. Gouripeddi, MBBS, MS, is an Assistant Professor in the Department of Biomedical Informatics, and the Assistant Director of the Informatics Core, Center for Clinical and Translational Science, University of Utah. His research interests are in clinical and translational research data integration, assimilation, data infrastructure and in artificial intelligence methods.

Jeremy Harper Chief Research Information Officer will present the Indiana and **Regenstrief Institute**'s experience with implementing a COVID19 specific dataset based off a health information exchange. How rapid changes to expectations for healthcare systems across the state both contributed to better data coverage while impacting expectations for data contributions. That data process informs the infrastructure data contributions in onboarding to the N3C. Also discussed will be the process required to implement N3C locally and long-term feasibility of similar approaches.

Jeffery Talbert will discuss the **University of Kentucky**'s (UK) experience sharing data with N3C using the ACT framework. The UK CTSA biomedical informatics core maintains a research data warehouse that supports data extracts for the ACT network, TriNetX, and local i2b2 SHRINE projects. Once the regulatory requirements were completed (DTA and IRB), the UK team ran the scripts supplied from the N3C Phenotype and Data Acquisition group to generate the N3C extract. The data feed was simplified given the scripts but did require some local data modifications and cleaning. Once the data was submitted, the data ingestion process also revealed some local miscoding for some lab values. The final step was developing an automated process to extract and submit data to the N3C twice a week. Dr. Talbert, PhD, is Professor and Division Chief of Biomedical Informatics and Director of the Institute for Biomedical Informatics. His work is focused on translational and public health informatics.

Statement of the panel organizer: All participants listed in this proposal have agreed to take part in this panel.

References

1. Johns Hopkins University Center for Systems Science and Engineering (CSSE). COVID-19 Dashboard. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
2. Haendel M, Chute C, Consortium. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. *J Am Med Inf Assoc. 2020.* In Press.
3. Ford D, Harvey J, McElligott J, et al. Leveraging Health System Telehealth and Informatics Infrastructure to Create a Continuum of Services for COVID-19 Screening, Testing, and Treatment. *J Am Med Inf Assoc*. 2020.
4. Obeid JS, Davis M, Turner M, Meystre SM, Heider P, Lenert L. An AI approach to COVID-19 infection risk assessment in virtual visits: a case report. *J Am Med Inf Assoc*. 2020.

Tales from the Front Line – What Happens Between Here and There with HL7 Data Exchanges

Sandra Mitchell, RPh, MSIS, FASHP[1], Fran Martin, RPh[2], Maureen Layden, MD, MPH[3], Jason Vogt, BS[4], Eric LaChance, MBA[5]

[1]Veterans Health Information Exchange (VHIE), J P Systems, Inc., Clifton, VA; [2]Retired Health Care Executive and Health Care Consultant, Raleigh, NC; [3]VA National Medication Reconciliation Initiative VHA, Washington, DC; [4]Meditech, Minnetonka, MN; [5]PathogenDx, Scottsdale, AZ

Keywords: Data quality, Data-driven research and discovery, Data sharing / interoperability, Data standards

General Description

This panel has launched a landmark collaboration to accelerate the quality of clinical data exchanges. Their work is grounded in a vision of a health care ecosystem where health care actors (e.g., clinicians, vendors, clinical data quality analysts) strive to improve the health care decision making process, advance patient care, and improve operational efficiency.

The VHIE Clinical Data Quality Team program serves as a model to educate the health care community about each health care actor's opportunity to take immediate action towards mitigating risks, improving workflow, and ensuring that the clinical intention of electronic exchanges is conveyed.

Each panelist represents a key health care actor illuminating the gap between expectation and experience. With the patient at the center of their health care continuum, a wide variety of principal and supporting health care actors weave in and out of the patient's wellness lifecycle.

Methodology

Four years ago, the VHIE Clinical Data Quality Team implemented a strategy to focus on the nation's largest HL7 data exchanges with an emphasis on data elements with the highest clinical impact. Analytics and metrics provide the cornerstone to build the data quality story for each specific organization. The C-CDAs exchanged are the primary insight stream to develop quantitative assessments and trends, supplemented with clinician reporting captured by the Community Coordinators.

To gain better qualitative insight into clinician issues, VHIE Community Coordinators conduct semi-structured interviews across VA's health care system. Additionally, we apply a community-based participatory approach to engage with data stakeholders (e.g., software vendors and/or health care organizations) to gain insight into their perceptions, real-life experiences, and suggestions for implementing change.

Evaluation Results

Vendor/organization analysis findings reveal that very few HL7 messages are clean, correct, consistent, and complete. For example, in one Health Information Exchange's Allergy domain, 46.2% of data is missing a code and/or code system. In one Medications domain, 55.9% of data is missing a code and/or code system, immunizations are misplaced in the CCD-A, and fields are mis-aligned in format. Both of these examples show huge issues in data quality. Community Coordinator interviews with front-line clinicians confirm that the absence of high-quality information negatively affects clinical decisions, increases clinician and patient frustrations, fosters distrust of data, significantly decreases work efficiency, and leads to bottlenecks in workflow.

The community-based participatory approach with data stakeholders uncovered numerous issues, including variations in implementation of regulations/standards due to weak language, incorrect client customizations, inconsistent software versions, and improper migration of legacy platforms. This created a health care ecosystem with unreliable clinical data flow that requires data transformations along the way. The HL7 message data exchange technology may succeed, but the clinical data quality content often fails due to risky data transitions between systems.

The VHIE Clinical Data Quality Team uses a collaborative presentation that includes Community Coordinator input in order to engage the vendor/organization, with the goal of developing consensus plans to resolve the issues of clinical data quality. In the analysis example above, reviewing the domain code/code system mapping tables for specific stakeholders is a starting point. As a deliverable, a workbook of de-identified data issue examples (at the C-

CDA document level) provide an easy tool for the source health care team to identify roles that can address specific data issues.

Through these approaches, we have come to understand the importance of educating a broad spectrum of health care actors on clinical data quality. An abundance of data is useless and can even become a liability without the ability to find and act on information as needed. Even the most sophisticated Clinical Decision Support (CDS) algorithms and the most advanced Artificial Intelligence (AI) protocols cannot function without complete, high-quality, and robust clinical data. "What is true for any scientific inquiry is true for improving health care: the better the data, the more meaningful the results" (Schmidt, 2012, para 17).[1] Better access to accurate and timely clinical data content leads to better clinical and business decision-making. This is the true key to successful interoperability.

Additionally, active participation in standards evolution is critical in addressing how health care actors are impacted in practice, not just theoretically. We need more national conversations about interoperability and deeper explorations of solutions so that we can work together to improve clinical content within the health care interoperability space.

Brief Description of Each Panelist's Presentation

Fran, the Caregiver, will discuss her experience navigating the health care system for her 95-year-old father, a WWII Veteran. Most caregivers are ill-prepared for their role, but Fran has a background in the health care industry. However, even with her background knowledge, she still finds the landscape difficult to navigate. Fran hopes to help the audience understand the experience of a caregiver when poor clinical data content impacts navigation of the health care system, causing delays and frustration. Fran will help the audience understand the role of caregivers in providing clinicians with a holistic view of the patient, the technological diversity in caregiver skillsets, and the importance of supporting family members' health care decisions.

Maureen, the Clinician, will discuss her experience working within the health care system and the complications in caregiving she has experienced due to COVID-19. She will help the audience understand the balancing act of the clinician's role throughout the health care ecosystem, which is affected by data access, patient needs, business requirements, and barriers to the Electronic Health Record (EHR) supporting clinician workflow processes.

Jason, the Vendor, will discuss his experience as a provider of health care software, including pain points and steps he is taking to resolve issues. He will help the audience understand how standards and client implementations directly impact the way clinicians see and utilize data, including external data from other electronic medical records.

Eric, the Patient, will discuss his experience as a patient with multiple caregivers and the ways in which poor interoperability have affected his life. He will help the audience understand the importance of supporting patients through the health care decision making process by providing easy access to educational resources and health care data from all clinicians.

Conclusion

For data to be accessible and actionable, it must be detailed, robust, and complete while retaining its clinical intention across numerous transmissions, both within internal systems and across organizations. Petabytes of health care data accumulating rapidly every day across many disparate systems do not equate to more insight. It is critical to understand that high quantity does not necessarily indicate high quality, and therefore technical success is not necessarily clinical success.

In order to begin the difficult task of improving the clinical content of data exchanges, it is necessary to engage a full cast of health care actors across the health care ecosystem. The solution intricacy requires awareness of the importance of embracing a holistic understanding of the health care data environment across all the modalities, business models, potential treatment protocols, and mitigation strategy tools.

Health care actors in the standards space can begin by facilitating a more unified approach to implementation of requirements through more specific data guidelines and reducing ambiguity in the present ones. A critical example is Artificial Intelligence modeling, where current development now requires many data transformations, each one adding risk as each data transition parallels each patient health care transition's risk profile. Tight, clear, enforceable standards would provide a safer data environment, rather than each project re-inventing the transformations.

Involvement of a broad spectrum of health care actors to play a role in collectively shaping a very real and tangible impact on people's lives is necessary. "The future of health care will be centered around the broad and more effective use of data from any source" (Cardon, 2014, para 19)[4].

Topic Justification

COVID-19 has put a spotlight on health care data interoperability within the health care ecosystem and exposed the newly coined "infodemic" that includes dissemination of inaccurate information as well as bad data.[3]

Isolation of health care data to a single enterprise no longer exists. As patients see providers across organizations, it is necessary to exchange complete and accurate electronic medical records. To do this, national health care automated exchanges are required for health care decisions involving many health care actors, who need to be educated to treat the problem of poor clinical data quality and focus on the origin of the semantic, syntactic or configuration-related problem which automatically resolves the problem. Treating the symptom of poor data quality with complex analytic transformations at both ends of the data exchange is spending resources in the wrong place in the process.

The health care ecosystem has many stories. ONC feels that Health Information Exchanges (HIEs) have created a "one-stop shop"[2] for providers on reporting. However, the reality is that aggregation of missing codes, normalization of values in accordance with loose standards, varying vendor configurations, and inconsistent business rules can all inhibit the goal of true interoperability.

Anticipated Audience

The intended audience of this panel includes professionals across the health care ecosystem: operational and clinical, CMIO, CNIO, IT leadership, risk management, regulatory, standards organizations, clinicians, patients, vendors, dentists, pharmacists, academic researchers, data scientists, and other professionals involved with decision making based on the collection, dissemination, and analysis of health care data.

Discussion Questions

- Do you know the health care actors in your health care neighborhood?
- When is FHIR coming, and what will it bring to the health care ecosystem in terms of interoperability?
- What keeps data exchange resources up at night?
- How can health care institutions support the coming interoperability data tsunami from the Internet of Medical Things (IoMT)?
- What is the gap between clinician and patient data expectation and the reality of their experience?
- How can clinicians spend less time using technology during patient engagements?
- Why do networks, regulatory organizations, and standards organizations allow so much flexibility?
- What are the vendor responsibilities to the client to educate, provide options, and continually assess clinical data quality?
- What are the responsibilities of medical specialties to define minimum data requirements and data needs?
- How can clinical research programs be optimized to utilize current standards?

Statement of Agreement

All panelists have provided written agreement to participate in this AMIA 2021 Informatics Summit panel discussion.

References

1. Schmidt, Brooke. 2012 August 6. The critical importance of good data to improving quality; [accessed 2020 August 22]. https://www.psqh.com/analysis/the-critical-importance-of-good-data-to-improving-quality/.
2. Miliard, Mike. 2020 August 11. At ONC tech forum, Rucker touts value of HIEs in COVID-19 response; [accessed 2020 August 19]. https://www.healthcareitnews.com/news/onc-tech-forum-rucker-touts-value-hies-covid-19-response.
3. World Health Organization. 2020 February 15. Munich Security Conference; [access 2020 March 11]. https://www.who.int/dg/speeches/detail/munich-security-conference
4. Cardon, Drew. 2014 August 28. Healthcare databases: purpose, strengths, weakness; [access 2020 August 22]. https://www.healthcatalyst.com/insights/healthcare-database-purposes-strengths-weaknesses

Harmonizing What to Analyze in the National COVID Cohort Collaborative (N3C): Lessons Learned

Richard Moffitt, PhD[1], Andrew Girvin, PhD[2], Harold P Lehmann, MD PhD[3], Kristin Kostka, MPH[4],[5] Joel Saltz, MD PhD[1]
[1]Stony Brook University, Stony Brook, NY; [2]Palantir Technologies, Washington, D.C.; [3]Johns Hopkins, Baltimore, MD; [4]IQVIA, Cambridge, MA; [5]Observational Health Data Sciences & Informatics, New York, NY

Abstract

The National COVID Cohort Collaborative (N3C) was established to provide patient-level data to further COVID-19 research. While the data are made available for research in the form in which they were submitted, providing constructs of the data should make that research go faster, be more consistent across projects, and more transparent. This panel, comprising informaticians, domain expertise, data scientists, and experts in common data models, will describe the process by which we selected which variables needed harmonization, how we accomplished and vetted that harmonization, and how we made public that process and their results for over hundreds of variables and code sets. Our experience has implications for others involved in large-scale projects of pooled electronic health record data.

Introduction — Joel Saltz

The National COVID Cohort Collaborative (N3C; covid.cd2h.org) "aims to aggregate and harmonize EHR data across clinical organizations in the United States (US), especially the Clinical and Translational Science Awards (CTSA) Program hubs that encompass more than 60 organizations and their partners." [1] Initialized in March 2020, it is funded by the National Center for Advancing Translational Science (NCATS) built on the principles of "partnership, inclusivity, transparency, reciprocity, accountability, and security" and aims to effect the goals of open science.[1] While the initial efforts focused on governance and ingestion of analyzable data, the needs to characterize the cohort and to support roughly a dozen initial observational research studies led to the need to generate a library of systematically curated data elements. These data elements come in different flavors: Harmonized values (e.g., temperature in a common scale), harmonized code sets (e.g., which lab tests comprise "serum creatinine"), harmonized value sets (e.g., what are the standard choices for SARS-CoV-2 qualitative tests), and computable phenotypes (e.g., how do we express "diabetes" as a comorbidity). A crucial aspect of this project involved development and deployment of software pipelines and review processes to iteratively assess and improve data quality — quality assessment and control is a crucial aspect of creating reliable data elements from EHR information ingested from many sites.

In this panel, we describe the organization, roles, workflow, processes, and tools required to create these constructs and to socialize them throughout the N3C project, which encompasses hundreds of individuals and dozens of research projects. While our panel sections divide the process up into individual presentations, the actual work was a recurring collaboration of the panel members and others (see the Acknowledgements).

In the overview, we will describe the organizational context, the various workstreams, working groups, and Task Teams involved, and the shared principles and goals that governed this work.

The Platform — Andrew Girvin

The N3C Enclave — implemented entirely under the control of NCATS and hosted in AWS GovCloud— comprises the data being analyzed, the tools used for the analysis, and the reports that result. The data comprises EHR data supplied (as of Dec 22, 2020) by 36 CTSA sites (hundreds of thousands of COVID-positive patients and millions of COVID-negative) in 3 forms: Limited Data Set (LDS), Safe Harbor, and Synthetic. Custom tools were configured to manage the construct process. The data themselves are in the OMOP Common Data Model (CDM),[2, 3]} having been transformed, if necessary, from native CDMs (PCORnet,[4] ACT,[5, 6] TriNetX[7, 8]), so each "field name" was replaced by an OMOP concept, with its own id and context. In this section we will describe the needs these tools were configured to satisfy and how such tools are configured in the N3C Data Enclave environment.

The Dictionary — Kristin Kostka

The core activity was in selecting sets of concepts to capture the appropriate data elements for downstream analysis. Through ongoing multi-center COVID-19 network research, many of the issues faced by the N3C in this work were already addressed by the OHDSI community. [9] For instance, the Atlas tool provided an initial infrastructure that motivated the Enclave tools.[10] But beyond tool building, the OHDSI experience provided expertise for vetting the process of assessing the constructs as well as providing insight into how concepts chosen from the OMOP vocabulary would be expected to behave and what their semantics generally were. Finally, the OHDSI experience with ongoing curation proved crucial in creating a robust foundation to N3C's approach: ensuring data quality while minimizing workload on the local sources.

The Work — Richard Moffitt

Defining and validating concept sets required the construction of a suite of visualization tools that provided important feedback during this iterative process. In this section, we will describe the methods and visualizations used to assess the quality of the data and how those assessments were used. Of particular note was the investigation of units and value ranges for each variable, and the subsequent set of conversion formulas that would be used to harmonize them. Beyond just measurement data, candidate concept codes for hundreds of medications, phenotypes, and categorical variables were also assessed. While the "Kahn Framework" gave the essential structure to the analysis,[11] issues crucial to a pooled-project like ours were raised. In particular, variability across sites was important to characterize (and in some cases correct), while maintaining anonymity of the sites. The notion that new data would always be coming in, and so early assumptions might be obviated also influenced the construct process. Finally, the work here provided "signals" regarding data quality issues that often had to be managed by sites or the mapping process more upstream.

The Transparency — Harold Lehmann

Beyond the work internal to defining the constructs was the need for transparency during and after creation of the constructs. "Issues" are reported upstream to the data-quality staff by analysts throughout the Enclave, while documentation of Issues and Limitations accompany data and concept sets. Review by clinicians was crucial as part of the process, as well as review by other workstreams and Task Teams within the N3C project. Furthermore, availability for review by others outside the project remains important to maintain trust and believability in N3C results, whether formally, in terms of compliance with STROBE requirements,[12, 13] or informally, in terms of face validity. Finally, visibility to the analysts downstream from the construct-creation process was vital. This panel will discuss how attention to these downstream issues affected the construct process.

Acknowledgements

The work presented in this panel reflects the collaboration of teams of individuals from across the N3C project, Observational Health Data and Informatics (OHDSI), Palantir, and the many organizations and universities whose members provided ongoing input, support, and participation. Thanks to Kenneth Gersing, Melissa Haendel, Christopher Chute, representing N3C leadership, Tell Bennett from University

of Colorado and the many co-authors of the Cohort Paper, Sandeep Mallipatu, Janos Hajagos and Jacob Woolridge from Stony Brook, Kate Bradwell, Benjamin Amor, and Nabeel Qureshi from Palantir, Shawn Murphy from Harvard, Anita Walden and Andrew Neuman from Oregon Health Sciences University, Clair Blacketer from Janssen Research & Development; Karthik Natarajan, Anna Ostropolets and Matthew Spotniz from Columbia University; Andrew Williams from Tufts University; Emily Pfaff, Marshall Clark, Kellie Walters, Adam Lee, Evan Colmenares, Robert Bradford from the University of North Carolina. This work has been funded through the National Center for Advancing Translational Sciences, National Institutes of Health, under award number U24 TR002306. Dr. Saltz, the panel organizer, has an affirmation from each of the participants that each agree to participate on this panel.

Questions to enhance audience participation

1) What "constructs" are important to pooled-EHR research and why?

2) What are the high-level steps required for managing such constructs?

3) What are the principles that govern the creation and management of such constructs?

4) How does the process described differ from what is done in single-center studies? From multi-institutional, single- project studies? From national CDM-based projects?

5) What attributes of a data-management environment help or hinder these activities?

References

1. Haendel M, Chute C, Gersing K. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. J Am Med Inform Assoc. 2020;2020 Aug 17;ocaa196.
2. OMOP Common Data Model. OHDSI; 2019; Available from: https://www.ohdsi.org/data-standardization/the-common-data-model/.
3. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015 May;22(3):553-64. https://www.ncbi.nlm.nih.gov/pubmed/25670757.
4. Common Data Model (CDM) Specification, Version 5.1: pcornet2019 Sept, 2019. https://pcornet.org/wp-content/uploads/2019/09/PCORnet-Common-Data-Model-v51-2019_09_12.pdf.
5. Visweswaran S, Becich MJ, D'Itri VS, Sendro ER, MacFadden D, Anderson NR, et al. Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. JAMIA Open. 2018 Oct;1(2):147-52. https://www.ncbi.nlm.nih.gov/pubmed/30474072.
6. i2b2 / ACT Network (SHRINE) [Internet]. CTSI; 2020 [cited 2020 Aug 26]; Available from: https://www.ctsicn.org/i2b2-shrine-act.
7. Topaloglu U, Palchuk MB. Using a Federated Network of Real-World Data to Optimize Clinical Trials Operations. JCO Clin Cancer Inform. 2018 Dec;2:1-10. https://www.ncbi.nlm.nih.gov/pubmed/30652541.
8. : TriNetX; 2020 [cited 2020 Aug 26]; Available from: trinetx.com.
9. C. S. 88 Hours: OHDSI's Signature Moment. 2020 [cited 2020 Aug 26]; Available from: https://www.ohdsi.org/88-hours/.
10. Atlas. OHDSI; 2020 [cited 2020 Aug 27]; Available from: http://atlas-covid19.ohdsi.org/#/home.
11. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Washington, DC). 2016;4:9-11.
12. Kahn MG, Brown JS, Davidson BN. Transparent Reporting of Data Quality in Distributed Data Networks. eGEMs. 2015;3:Article 7.
13. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clinical Pharmacology and Therapeutics. [Article]. 2017;102(6):924-33. https://www2.scopus.com/inward/record.uri?eid=2-s2.0-85029886626&doi=10.1002%2fcpt.857&partnerID=40&md5=944f230075e551b47fcda52a6ee75d96.

Performance of COVID-19 Research in the CTSA ACT Network

Shawn Murphy MD, PhD[1,2], Jeffrey Klann, PhD[1], Michele Morris, BA[3], Dipti Ranganathan, MS[4], Griffin M Weber, MD, PhD[2]

[1]Mass General Brigham, Boston, MA 02114; [2]Harvard Medical School, Boston, MA 02115, [3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA; [4]University of Chicago, Chicago, IL

Abstract - *The Accrual for Clinical Trials (ACT) NIH Research Network was established for CTSA-attached organizations to perform Electronic Health Record research on COVID-19 infections. The goal for performing research in the ACT network is to allow participation from all joining institutions to create hypotheses and perform analyses across the network. We began during the 2020 pandemic to assemble and work with the pieces that could enable not only general health record queries, but focused questions that were arising from the pandemic. This could be achieved by a joint effort that allowed specifically constructed COVID-19 ontologies to be applied to the site data, analytic programs to be built and run at local sites, data quality to be assessed and validated at the sites, and governance to allow results to be pooled and published. The methods, tools and data structures used are open source and can be feely learned, exchanged, and reproduced.*

Description of Panel

In April of 2020, a pilot was launched to perform COVID-19 research across the ACT network. The pilot was focused upon the usefulness or harmfulness of medications that may affect who gets infected with COVID-19 and whether the severity of the illness is affected by the medication. Performing a robust analysis of this type required several steps.

First, research questions were assembled from the CTSA-attached organizations by ACT leadership reaching out to experts at their associated hospitals. It became apparent that critical questions regarding effective therapeutics and harmful medications were being asked. It was also apparent that many different medication classes and mechanisms of action were hypothetically important, although empirical evidence was lacking. Therefore, a framework to measure disproportionate numbers of patients infected who had taken specific classes of medications, and those who became more or less severely ill who were given medications in the hospital was deemed to be desirable for research.

Second, a COVID-19 specific ontology was created that could be used to ask research questions and organize research data. Outcomes and test results were specially defined so that they could be implemented at the sites where queries would be performed. The data that was defined by the ontology could be queried through the network and ensured that those data were being collected at the sites. This served as the principle method to standardize data collections and add new data elements to the collections.

Next, once the data was specified through the ontology, the data could be profiled to ensure it met requirements for completeness and quality. This could be established preliminarily following a series of network queries. Generally, this meant comparing data profiles at similar sites to ensure uniformity and making sure the evolution of data over time was occurring uniformly. Often the focus was to look for missing data, especially with regards to how recently data had been loaded for analysis.

Now the analytic programs are assembled both in SQL and in the R programming language. Both can be run directly against data that was extracted from the site i2b2 or OMOP standardized datasets as defined by the ACT ontology. The general format of these programs is to define a cohort by local data completeness criteria (a loyalty cohort) and then proceed to perform comparisons between the patients on defined medication classes.

Finally, visual displays perform "reality checks" against the analysis to search for characteristic patterns that show biases in the populations and discern discrepancies with the analysis. Often results were compared between hospitals to distinguish characteristics of overfitting and quantitative mismatches.

Following the analyses at the participating sites, the results are pooled. Depending on the nature of the results, statistics are generated at the sites and pooled, or intermediate data sets are generated at the sites and pooled. An ongoing publication on a document-sharing platform is then created.

Panelists:

Dipti Ranganathan - Ms. Ranganathan will describe the considerations that informed the design of the governance model for the COVID-19 ACT pilot as the pilot represented a transition from a cohort identification and accrual network to include the use of the network for the creation of hypotheses and the analysis of data for research. The urgency to take a network with 150 patient records across 48 participating institutions and build an adjunct, well-curated data set specifically targeted to COVID-10 infections required agility from the governance structure, quick collaboration from the various workgroups and tested the network's communication structure. Ms. Ranganathan will also describe the ACT Governance and these considerations and decisions that led to the successful pilot.

Michele Morris - Ms. Morris will discuss the process for building the ontology collaboratively for the COVID-19 ACT pilot. She will discuss the concepts selected for the ontology, how it differs from the organization of the other ACT ontologies and the challenges in developing an ontology with emergent concepts. She will also reflect on the impact of other efforts on the development.

Jeff Klann - Dr. Klann will discuss the data quality initiative in the COVID-19 ACT pilot. This uses a unique approach to count the number of patients in every possible single-element cohort across the network, leveraging the ACT COVID ontology to capture cohorts at varying levels of granularity (e.g., both 'COVID-related diagnoses' and 'Acute Respiratory Distress Syndrome'). This accumulation of counts across sites allows analytics to find site outliers, coding differences, missing data elements, and unexpected trends across refreshes (e.g., a decrease in total number of patients with a COVID-positive test). A web-based dashboard allows both researchers and site administrators to view these data quality problems interactively. Anomalies are flagged in an ontology browser and then visualized using bar and line graphs.

Griffin Weber - Dr. Weber will present the technical architecture and implementation challenges of the pilot and important considerations of using ACT for COVID-19 research. For the pilot we leveraged the ACT test network, which consists of 9 of the 47 ACT sites. When using ACT for COVID-19 research, investigators must take into account several things, including: (1) Patients can have multiple COVID-19 test results, which include false positives or negatives. (2) Patients, in general, have been avoiding hospitals and delaying visits. This can make them appear healthier than compared to before the pandemic. (3) Data censoring is significant since the disease course of COVID-19 can be weeks or months, and there hasn't been enough time yet for many patients to either fully recover or die. (4) Direct measures of disease severity, such as critical care or death may only be accurate in notes and not in coded data. (5) Because

patients receive care from multiple providers, a single EHR might not have all of a patients' comorbidities or medications.

Shawn Murphy - Dr. Murphy is a Professor of Neurology at the Harvard Medical School and Associate Director of the Laboratory of Computer Science at Massachusetts General Hospital and Chief Research Information Officer at Mass General Brigham. His research focuses on development of methodology and tooling to facilitate the use of machine learning in the clinical setting. Dr. Murphy will serve as moderator on the panel.

Need for the Panel

Many skills were leveraged from across the United States at a time of great need for performing rapidly iterating research using the Electronic Health Record. Because the local teams at each hospital became partners in the research, it was possible to account for many changes occurring in the EHR during this time. Standardized test results and derived codes could be created and used quickly in analytics that would have been difficult to verify in a central database. The demonstration of both the method and the value of this approach is important for future planning and will be discussed with the audience.

Discussion questions:

1. What interesting data quality approaches and discoveries have been made in the ACT COVID-19 network?
2. How can other networks utilize the tools and approaches that have been developed here?
3. What challenges do investigators face when using EHR data for COVID-19 research?
4. How can the ACT COVID ontology be leveraged to harmonize EHR data across networks?
5. Given that an original scope of the network was expanded for the pilot network, did you consider writing a separate agreement and new governance documentation? Did the pilot groups have to sign new agreements?
6. How will you transition from the pilot to the entire network?

All participants have agreed to take part on the panel.

References:

Visweswaran S, Becich MJ, D'Itri VS, Sendro ER, MacFadden D, Anderson NR, Allen KA, et al.. Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. JAMIA Open. 2018 Oct;1(2):147-152. PMID: 30474072; PubMed Central PMCID: PMC6241502.

Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):624-30. doi: 10.1197/jamia.M3191. Epub 2009 Jun 30. PubMed PMID: 19567788; PubMed Central PMCID: PMC2744712.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30. doi: 10.1136/jamia.2009.000893. PubMed PMID: 20190053; PubMed Central PMCID: PMC3000779.

Data Acquisition and Harmonization of COVID-19 Case Data Across Common Data Models: Early Field Reports from the National COVID Cohort Collaborative (N3C)

Emily Pfaff, M.S.[1], Stephanie Hong, B.S.[2], Dazhi Jiao, M.S.[3], Xiaohan Tanner Zhang M.D., M.S.[4], Christopher G. Chute M.D. Dr.PH[4]
[1]NC TraCS Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [2]Institute for Clinical and Translational Research, School of Medicine, Johns Hopkins University; [3]Department of Medicine, Johns Hopkins University; [4]Schools of Medicine, Public Health, and Nursing, Johns Hopkins University, Baltimore, MD, USA

Abstract

The National COVID Cohort Collaborative (N3C), sponsored by the National Center for Advancing Translational Sciences (NCATS) is a partnership among Clinical Translational Science Awardees (CTSAs) and other academic medical centers; the National Center for Data to Health (CD2H); and members and subject matter experts from Observational Health Data Sciences and Informatics (OHDSI), PCORnet, the Accrual to Clinical Trials (ACT) network, and TriNetX. N3C's goals are to demonstrate that a "multi-site collaborative learning health network can overcome barriers to rapidly build a scalable infrastructure incorporating multi-organizational clinical data for COVID-19 analytics. This panel is composed of informaticians supporting the data acquisition, ingestion and harmonization processes of N3C, with a focus on phenotype development and implementation, building the data ingestion pipeline, improving COVID test results from early data sets and developing and implementation of a data quality framework for the project.

Introduction – Christopher Chute M.D., Dr.PH

Many research projects whose aims include aggregation of data across multiple research databases and common data models (CDMs) face resource bottlenecks associated with the requirement to support data transformation. The urgency and need for continually updated, timely access to population data presented by the COVID-19 pandemic compounds this issue. The National COVID Cohort Collaborative (N3C), sponsored by the National Center for Advancing Translational Sciences (NCATS) is a partnership among Clinical Translational Science Awardees (CTSAs) and other academic medical centers; the National Center for Data to Health (CD2H); and members and subject matter experts from Observational Health Data Sciences and Informatics (OHDSI), PCORnet, the Accrual to Clinical Trials (ACT) network, and TriNetX. N3C's goals are to demonstrate that a "multi-site collaborative learning health network can overcome barriers to rapidly build a scalable infrastructure incorporating multi-organizational clinical data for COVID-19 analytics.[1]"

From March through July of 2020, N3C organized topic-focused and inclusive working groups centered on governance and legal agreements; COVID-19 cohort identification and data extraction; data quality processes; data harmonization pipeline development; and analytics tools and methods.[2] This panel consists of representatives from the N3C Phenotype & Data Acquisition work stream and the Data Ingestion &Harmonization work stream. Together, these two work streams represent the informatics "back end" of N3C, and focus their efforts on defining, acquiring, and harmonizing the electronic health record data from tens of sites that are eventually made available for analysis in N3C's Secure Enclave. Panelists will discuss the methods, processes, and innovations used to plan and execute computable phenotyping, data extraction, data harmonization, and data transformation for N3C on a very short timeline.

Phenotype and Data Acquisition – Emily Pfaff, M.S.

The purpose of the Phenotype & Data Acquisition work stream is threefold: (1) to determine the data inclusion/exclusion criteria for import to N3C (computable phenotype); (2) to create and maintain a set of scripts to execute the computable phenotype in each of four CDMs—ACT, OMOP, PCORnet, and TriNetX—and extract relevant data for that cohort; and (3) to provide direct support to sites throughout the data acquisition process.

Defining the cohort. Because our knowledge of COVID-19 is evolving, it is challenging to define a stable computable phenotype to identify COVID-19 patients from their EHR data. We chose to bring together existing inclusion criteria and code sets from a number of organizations into a "best-of-breed" phenotype. The N3C phenotype[3] is designed to be inclusive of any diagnosis codes, procedure codes, lab tests, or combination thereof that may be indicative of

COVID-19, while still limiting the number of extracted records to meaningful and manageable levels. The panel will discuss the process of developing the initial phenotype, engaging the community for input, and updating and maintaining the phenotype over time.

Devising extraction methods. We did not want participating sites to have to write their own code to extract data per our specifications, as this would be burdensome for sites and could also introduce inconsistency and errors among data payloads. For this reason, we wrote a series of SQL, Python, and R scripts to help sites easily and consistently extract their CDM data for the N3C cohort. The panel will discuss the design of these scripts, overcoming issues encountered, and other potential uses of this architecture.

Supporting sites. A significant part of the Phenotype & Data Acquisition work stream's workload is supporting participating sites who have questions about the phenotype or extraction scripts, or who are having trouble getting up and running. As a small team charged with supporting tens of sites, we devised a support structure that enables us to support as many sites as possible in an efficient way. The panel will discuss this support structure, which includes individual help sessions, twice-weekly office hours, and large community calls.

Data Ingestion and Harmonization - Stephanie Hong, B.S.

The N3C Data Ingestion and Harmonization (DI&H) pipeline is a hybrid solution utilizing customized scripts working in tandem with a commercial workflow management product. The DI&H pipeline functions under the principle of minimizing transformations to render the source data faithful to their source instances into the selected target, OMOP 5.3.1, using the N3C domain identifiers created at the beginning of the ingestion cycle. Preceding the mapped transformations in the pipeline, checks occur for key missing data and common data model conformance. The mapping process aims to achieve consistent transformations to OMOP 5.3.1 en masse across the many sites contributing a continuously updated pipeline of COVID-19 data. The hybrid N3C ingestion architecture uses both static and dynamic maps for all terms presented in the native CDM payloads: N3C OMOP domain identifiers generated from all source data. The OHDSI Data Quality Dashboard (DQD) runs on the ingested, transformed payloads once in the OMOP 5.3.1 format. Contributing sites receive the results of the DQD and OMOP transformed payloads via site-specific, secure folders. The N3C aggregate data store receives each site OMOP payload. The data refreshes every time a site contributes a new payload: completely replacing previous data. A safe harbor data store is created from limited data set aggregates through zip code and date / time obfuscation, persisting date / time shifts applied to each person record for every site payload data refresh cycle.

Improving COVID Laboratory Data - Dazhi Jaio, M.S.

COVID-19 case identification is of paramount importance to the project. Consistent early case detection at any individual site is challenging. Multiple sites contributing data compounds the problem. Normalization of test data is challenging due variability of testing in early phases of the outbreaks as well as editorial lag of standard content from coding systems. The N3C data ingestion team applied an open-source rule-based tool: COVID-19 TestNorm[4] and other methods to identify COVID 19 cases and align testing data in the N3C aggregate dataset. The development team used the COVID-19 TestNorm tool to create scripts in Python to discover and transform fields in CDM native databases coded to 'NULL' in place of LOINC COVID-19 and related test in contributed lab test tables. The scripts generated were then integrated into the N3C data ingestion pipeline as a component of the overarching data normalization workflow. As the pandemic progressed and new diagnostics methods to detect COVID-19 developed, corollary LOINC reference information required continual updates in the pipeline test discovery and mapping code.

Addressing Data Quality - Xiaohan Tanner Zhang M.D., M.S.

Data quality is an important aspect of data aggregation projects supporting research secondary use cases. Compounding the issue in N3C is its use of multiple CDMs, each of which contribute data transformed from primary sources. DI&H convened a series of meetings with a broad group of CDM and data aggregation subject matter experts from the translational informatics community charged with formulating strategies for data quality within the N3C ingestion pipeline. The DI&H quality configuration reflects this feedback while remaining true to the principle of minimum but necessary data transformations. The DI&H team in partnership with N3C Phenotype and Data Acquisition and Collaborative Analytics teams has developed a data quality framework that aims are to provide a systematic analysis and remediation methodology, incrementally addressing quality issues in individual site contributed and aggregate data stores. The DI&H group utilizes the same analytics platform that supports N3C

researchers to do concept aggregation, visualization and develop data harmonization tasks deployed both upstream in the DI&H pipeline as well as a feature of the production analytics platform supplied to researchers.

Acknowledgements

The work presented in this panel reflects the collaboration of teams of individuals from across CD2H, Observational Health Data and Informatics (OHDSI), NCATS Accrual for Clinical Trials (ACT) Network, Patient-Centered Clinical Research Network (PCORnet) and TriNetX organizations. The panelists would like to acknowledge the contributions of Davera Gabriel, Tricia Francis, and Richard Zhu representing Johns Hopkins University; Clair Blacketer of Janssen; Kristen Koskta of IQVIA; Matvey Palchuk of TriNetX; Michele Morris and Shyam Visweswaran representing the University of Pittsburgh; Marshall Clark, Kellie Walters, Adam Lee, Evan Colmenares, Robert Bradford representing the University of North Carolina; Karthik Natarajan representing Columbia University; Robert Miller representing Tufts University; Charles Yaghmour and Smita Hastak of Samvit Solutions; Raju Hemadri of Digital Infusion; Sandeep Naredla and Srini Rao of Adpetia and Ken Gersing of NCATS. This work has been funded through the National Center for Advancing Translational Sciences, National Institutes of Health, under award number U24 TR002306. Dr. Chute, the panel organizer, has an affirmation from each of the participants that each agree to participate on this panel.

Questions to enhance audience participation

1) What data challenges did sites face when participating in the network? What remedies best addressed these challenges?
2) How did the phenotype definition influence data extraction and quality issues in the acquisition and ingestion processes?
3) What were the most prevalent data quality issues for N3C? How were these addressed?
4) What was the approach to normalize semantic differences between contributing common data models?

References

1. National COVID Cohort Collaborative (N3C) [Internet]. National Center for Advancing Translational Sciences. 2020 [cited 14 July 2020]. Available from: https://ncats.nih.gov/n3c
2. Haendel M, Chute C. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. JAMIA Open, 2020; [in press].
3. National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition [Internet]. GitHub. 2020 [cited 14 July2020]. Available from: https://github.com/National-COVID-Cohort Collaborative/Phenotype_Data_Acquisition/wiki/Latest-Phenotype
4. Dong X, Li J, Soysal E, Bian J, DuVall S, Hanchrow E et al. COVID-19 TestNorm - A tool to normalize COVID-19 testing names to LOINC codes. Journal of the American Medical Informatics Association. 2020;.

Extracting clinical data from EHRs to support research: Early lessons from the Cancer Moonshot's IMPACT Consortium

Jennifer R. Popovic, DVM, MA[1]; Roxanne E. Jensen, PhD[2]; Parvez Rahman, MHI[3]; Firas H. Wehbe, MD, PhD[4]; Michael Hassett, MD, MPH[5]

[1]RTI International, Waltham, MA; [2]National Cancer Institute, Bethesda, MD; [3]Mayo Clinic, Rochester, MN; [4]Northwestern University, Chicago, IL; [5]Dana-Farber Cancer Institute, Boston, MA

Abstract

In 2018, the National Cancer Institute established an initiative to fund a consortium that aims to improve the monitoring and management of patients' cancer-related symptoms. The Improving the Management of symPtoms during And following Cancer Treatment (IMPACT) consortium is supported by funding provided through the Cancer Moonshot℠, and is comprised of three research centers (RCs) and a coordinating center (CC) tasked with collecting and sharing symptom data from across the cancer care continuum. One data stream includes elements from each RC's electronic health record (EHR) system, including patient demographics and associated clinical data such as diagnoses, procedures and medications. Despite each RC utilizing the same EHR vendor platform, a standardized approach to extracting data across sites has proven deceptively challenging. This panel will highlight RCs' varied approaches, challenges, and solutions. Learning objectives include understanding current EHR data extraction and standardization approaches and the potential that emerging technologies may hold.

Background and significance

IMPACT has funded three RCs, that are supported by a CC and NCI scientists, to conduct pragmatic trials in oncology settings. The studies are utilizing patient-reported outcomes and clinical data from the EHR to assess and respond to patient symptoms. Data elements from RCs' EHR systems include patient demographics, cancer and comorbid condition diagnoses, procedures, and medication orders and administrations. To support the ability to perform pooled Consortium-level analyses, data from all RCs will be extracted from their source systems and aggregated by the CC into a single, harmonized data asset that will be structured in accordance with a common data model (CDM) designed by the CC.

Despite the fact that all of the IMPACT RCs use the same EHR vendor, a consistent approach to extracting EHR data in a manner that maximizes the ability to standardize, harmonize and make them interoperable and interpretable for research remains elusive. Proper identification of the specific source data tables and elements requires analysts with specialized knowledge of proprietary EHR platform data table structures (in this case, Epic) that serve as the source of the patient clinical data. Knowledge-sharing of those proprietary data structures across provider systems who use the same EHR platform is often of limited utility due to facility-specific customizations of their EHR platforms; there is at-best partial transferability of data content standards between provider systems using the same EHR vendor platform[1].

There is legislative momentum to make bulk extraction of EHR data more efficient and interoperable via open standards and technologies, to support multiple purposes including research[1]. Under the Cures Act, for example, providers will need to have Fast Healthcare Interoperability Resources (FHIR)-based bulk extraction capabilities in place by 2023[2]. This would theoretically allow data queries such as, "return all data for a cohort of patients", that retrieve data in accordance with FHIR-based data content and exchange standards. Until then, FHIR-based data exchange standards are designed to support export of data one patient at a time[1]. And even after 2023, it remains to be seen how bulk data exchange implementation will play out across various EHR vendor platforms and the provider systems that rely on them.

Description of the Panel

This panel provides an early-look at the potential, promise and challenges of curating and extracting clinical data concepts and elements from multiple sites' EHRs into a centralized, standardized, harmonized and computable research data asset.

This panel aims to, (a) describe the approaches, challenges and solutions encountered by each IMPACT RC in identifying and extracting clinical data elements from their respective source systems for this multi-site cancer research consortium, (b) describe the IMPACT CC's approach to standardizing and harmonizing these disparately-sourced data to support consortium-wide analyses (e.g., adaptation of existing CDMs and data standards), and (c) engage in discussion about the promise of emerging standards and technologies, such as the FHIR bulk data standard or EHR platform-specific apps (e.g., REDCap EHR integration module for Epic), to facilitate use of EHR data for research in the future.

Individual Panelists

Jennifer R. Popovic, DVM, MA: Dr. Popovic will introduce the panel participants to frame the importance and challenges of using EHR data to support consortium-level research, describe the IMPACT CC's approach to standardizing data from disparate systems, and facilitate the discussion after panel presentations. Dr. Popovic is a Senior Director at RTI International, with expertise in the design and development of research infrastructure to support the secondary-use of multiple healthcare data sources, such as claims and EHR data.

Roxanne E. Jensen, PhD: Dr. Jensen will frame the priorities for using patient clinical data from the observational EHR to support the IMPACT Consortium research agenda by describing NCI's goals for IMPACT. Dr. Jensen is a Psychometrician and Program Director in the Outcomes Research Branch within the Healthcare Delivery Research Program at NCI.

Parvez Rahman, MHI: Mr. Rahman will describe the early efforts of the informatics team at Mayo Clinic for the Enhanced, EHR-facilitated Cancer Symptom Control (E2C2) trial to identify and extract EHR patient and clinical data from source systems. Mr. Rahman is a principal health services analyst in the Mayo Clinic Robert D. and Patricia E. Ker Center for the Science of Health Care Delivery. He has a background in applied informatics and brings his expertise in creating integrated data and visual analytics platforms to the E2C2 project.

Firas H. Wehbe, MD, PhD: Dr. Wehbe will present Northwestern University's IMPACT consortium efforts to identify and extract clinical data elements from their EHR and enterprise data warehouse. Dr. Wehbe holds several leadership roles at Northwestern University: Chief Research Informatics Officer at Northwestern Medicine; Director of the Applied Research Informatics Group within the Northwestern University Clinical And Translational Sciences Institute; and Associate Director of Quantitative Data Sciences Core, Robert H. Lurie Comprehensive Cancer Center.

Michael Hassett, MD, MPH: Dr. Hassett will provide an overview of the Dana-Farber Cancer Institute-led Symptom Management IMplementation of Patient Reported Outcomes in Oncology (SIMPRO) Research Center's unique role as an multi-site coordinating center embedded within the IMPACT consortium and the bearing that role has on SIMPRO's approach to extracting EHR data to support the broader consortium-level research agenda. Dr. Hassett is a medical oncologist and was the physician lead for the Epic deployment at Dana-Farber.

The panel will be organized as follows:

Time	Speaker	Topic
7 min	Popovic	Introduction of the panel topic and panelists
12 min	Jensen	Review of the Cancer Moonshot^{SM}, IMPACT Consortium research aims and how clinical data elements support IMPACT's broader vision
12 min	Rahman	Overview of the E2C2 project and early lessons about identifying and extracting EHR patient and clinical data from source systems to support a multi-site research consortium agenda

Time	Speaker	Topic
12 min	Wehbe	Overview of the NU IMPACT project and early lessons about identifying and extracting EHR patient and clinical data from source systems, including their enterprise data warehouse, to support a multi-site research consortium agenda
12 min	Hassett	Overview of the SIMPRO project, their unique role as a "network within a network", and how that role shapes their approach to EHR data extraction
35 min	Popovic	Brief overview of the CDM developed for the IMPACT Consortium, followed by facilitation of discussion and Q&A amongst panelists and audience

Learning Objectives

1. Understand the practical issues surrounding extracting data from EHRs for secondary-use research purposes, including staff roles involved in performing the work, common challenges encountered and solutions to meet those challenges
2. Understand the varied approaches each IMPACT RC is using to extract data from their EHR systems, as well as the approaches the IMPACT CC is taking to leverage existing clinical research infrastructure (e.g., common data models and terminology standards) to create a standardized, interoperable and computable data asset from disparate clinical source systems
3. Learn about the potential of emerging technologies, such as the FHIR bulk data standard, to support EHR data extraction efforts for research in the future, from the perspective of IMPACT RC panelists

Anticipated audience

The anticipated audience includes informatics professionals, clinicians, researchers, and patients who have an interest in secondary-use of EHR data and data standards and interoperability issues surrounding use of EHR data to support multi-site research, particularly for cancer research.

Discussion questions

1. Are RCs making use of any specific technologies (e.g., apps, APIs) to extract data from their respective source systems?
2. Are there any EHR data retrieval and use issues that are unique to supporting cancer research?
3. How will future bulk EHR data exchange standards change current processes of extracting/exchanging EHR data to support cancer research?

Attestation

The panel organizer has assurances that all named participants have agreed to take part on the panel.

Conclusion

A standardized approach to extracting EHR data elements to make them interoperable and interpretable for research is challenging. NCI IMPACT panelists will discuss their current approaches, challenges, and solutions. The panelists will also discuss the potential of emerging standards and technologies to facilitate use of EHR data for research in the future.

References

1. Meeting to Advance Push Button Population Health: SMART/HL7 Bulk Data Export/FLAT FHIR. Proceedings; 2019 Nov 6. 2019; Boston, MA: Health Level Seven International (HL7) and Boston Children's Hospital Computational Health Informatics Program; [cited 2020 Aug 21]. Available from: http://smarthealthit.org/wp-content/uploads/SMART-2019_FHIR-Bulk-Data-Meeting_final.pdf

2. Department of Health and Human Services, Office of the Secretary. Final Rule. 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program. Federal Register 85, Issue 85. May 1, 2020: 25642-25961 [Internet]. Department of Health and Human Services: Washington, DC; 2020. Available from: https://www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification

Implementability of Deep Learning based Predictive Models: Bridging Data Science Research and Real-World Practice

Laila Rasmy, MSc[1], Angela Ross, DNP[1], Kathleen McGrow, DNP[2], Robert E. Murphy, MD[1], Degui Zhi, PhD[1]

[1]School of Biomedical Informatics University of Texas Health Science Center, Houston, TX; [2]Microsoft Corporation, New York, NY

Abstract

With the abundant availability of secondary electronic health record (EHR) data, researchers start to focus on developing predictive algorithms using such big clinical data. Deep learning (DL)-based models are offering promising prediction accuracy and proved to outperform traditional statistical or machine learning algorithms. Yet, we rarely see those models implemented or validated in practice. Data scientists need to consider the implementability of those models during the development phase to facilitate clinicians' acceptance for further clinical validation. The main objective of this panel is to identify factors associated with the implementability of DL-based predictive models. We will define each factor in the context of predictive modeling of clinical events and describe how it impacts the feasibility of the model implementation. We will also discuss the magnitude of the impact of each factor and how to objectively measure or evaluate it.

Introduction

We use the term implementability to refer to the feasibility of a system or a model to be implemented, so it is an evaluation process before deciding on the actual implementation. Shiffman et al[1] provided a clear definition of implementability for clinical guidelines. They defined implementability as the set of characteristics that predict ease of use or determine the key obstacles for guideline implementation. On the other hand, there is no clear definition in literature for implementability in the context of artificial intelligence (AI) and clinical events predictive modeling. Therefore, we proposed a framework to define and evaluate the implementability of AI predictive models for clinical events. Our framework utilized the key criteria used for machine learning model evaluation as emphasized by the food and drug administration (FDA) good machine learning practice(GMLP)[2] to identify the possible obstacles that may impact the implementation of the predictive models as well as subject matter experts (SMEs) feedback. Our proposed framework includes six evaluation criteria. The first criterion is the prediction accuracy, which should be calculated based on practical and meaningful metrics. The second criterion is the reliability of the predictions which is highly impacted by the availability of a clear interpretation or explanation of the given prediction. The third criterion is the generalizability of the predictions, which can be defined by providing the calibrated accurate prediction in different settings even under distribution shift. The fourth criterion is the optimized data mechanics. Data mechanics is a term used in the industry that refers to the data flow from the source (EHR system) to the destination (predictive model) including all required data wrangling, normalization, and any additional preprocessing steps required to reach the final format consumed by the model. The fifth criterion is computational efficiency including factors like model size and running time that has a direct impact on cost as well as other factors like portability. The last and most important criterion is data privacy and security. Although security measures are a part of actual implementation and it might not be realistic to consider in early development phases, developed models need to comply with HIPAA regulations and avoid the use or need of any PHI data, which are commonly unnecessary for disease predictions.

We reviewed predictive modeling articles indexed in PubMed in the last 5 years that used structured EHR data to train deep learning based models. We found that all articles reported the model prediction accuracy as the major evaluation criteria, 50% of the articles discussed the proposed model interpretability or explainability and Less than 25% evaluated the model generalizability, efficiency, ease of use, computational cost, or running time. Only a couple of articles described data mechanics, while no article discussed the security of the proposed models but the majority were not using any PHI data as input features.

Learning Objectives

Upon completion of this panel, attendees will be able to:

- Understand the value of Artificial intelligence (AI) and DL-based models to predict clinical events.

- Anticipate the role that AI research will play to help improve health outcomes.
- Understand the gap between AI research and clinical application and how we can bridge between them.
- Define AI implementability.
- Enumerate factors associated with the implementability of DL-based predictive models and rank them based on importance.
- Consider ways to measure those factors to evaluate the feasibility of DL-based predictive models for further clinical validation

Panel Description

This interactive panel will encourage discussion on the feasibility of further validating DL-based predictive models for different clinical events in clinical settings. The panel will start with an introduction on the value of AI in healthcare with a focus on disease and other clinical events predictive modeling. We will provide a brief description of the common process of developing DL-based predictive models for clinical events in a research environment and the role of clinicians involved in model development early phase. Then, we will describe the gap between AI research and real-world application and our proposed solution to bridge this gap.
We have a diverse panel from all aspects including gender, ethnicity, location, and professional background. There will be one moderator and four speakers who are representing different stakeholders' perspectives, as follows:

- Degui Zhi, Ph.D., MS (Associate Professor, UTHealth SBMI). Dr.Zhi has a unique expertise with a solid background in computer science, bioinformatics, and biomedical informatics. Dr. Zhi focuses on developing computational and statistical methods and practical strategies for biomedical big data analytics. His team has published a number of papers on predictive modeling of electronic health record (EHR) data using deep learning in biomedical informatics journals.
  Dr. Zhi will cover the data science research perspective. He will talk about deep learning concepts and the state-of-the-art of their applications to the prediction of clinical events. He will also discuss the role of clinicians' involvement in the early phase of model development.

- Kathleen McGrow, DNP (CNIO and Industry Executive, Microsoft). Dr. McGrow advises and supports organizations on how the innovative use of technology can support their digital transformation imperatives of consumer engagement, provider enablement, analytics for population health, and cognitive computing to support a learning health system. She believes that technology can improve the lives of providers and patients. Dr. McGrow's unique ability lies in her expertise in both clinical and IT domains. Trained as a nurse, educated in technology, and working for organizations at the forefront of developing today's most innovative HIT solutions, she has a great understanding of all stakeholders.
  Dr. McGrow will cover industry and vendor perspective. She will explain how AI can impact the healthcare quadruple aim, enhance clinical, operational and financial performance, maximize capacity and patient experience, and transform to new care models.

- Angela Ross, DNP, MPH, PMP, PHCNS-BC, LTC (Assistant Professor, UTHealth SBMI). Dr. Ross has held positions as CMIO, acting chief of system service and design, and project manager for the U.S. Army Medical Information Technology Center Defense Health Agency (DHA) for more than 25 years. She is an informatics consultant and leader in the implementation, integration, and operation of emerging and fielded clinical information technologies supporting health care and administrative functions for over 30 medical treatment facilities. Dr.Ross interests include workflow analysis, performance improvement; project management; system implementation; program and project evaluation; and policy development.
  Dr. Ross will cover the applied research and evaluation perspective. She will discuss the AI implementability factors and how to evaluate them.

- Robert E. Murphy, MD (Associate Dean for Applied Informatics and Associate Professor, UTHealth SBMI). Dr.Murphy was named the CMIO at Memorial Hermann Healthcare System (2005-2015) where he provided system-wide leadership on clinical information system projects, including computerized physician order entry, clinical decision support, and quality informatics. During his tenure, Memorial Hermann received the 2012 Eisenberg Award, the 2009 National Healthcare Quality Award from the National Quality Forum, and all nine acute-care Memorial Hermann hospitals reached Stage 6 in the HIMSS Analytics EMR Adoption Model by 2013. Dr. Murphy was named by Modern Healthcare magazine in 2010, 2011 and 2012 as one of the nation's Top 25 Clinical Informaticists. His project for "CDS Good

Catches", which prevented over 7,000 medical errors over a 1-year period, was awarded the Breakthrough of the Year in Quality at Memorial Hermann in 2010. He has published and lectured widely on physician adoption and change management, development of evidence-based content for electronic health records, and using information technology to improve patient safety and quality.
Dr. Murphy will cover healthcare organization management as well as clinicians' perspectives. He will discuss the importance of evaluating the implementability of predictive models before introducing to clinicians and healthcare organization for further validation and adoption.

- Laila Rasmy, MSc, MBA, BPharm (PhD. Candidate, UTHealth SBMI) will moderate the panel. Laila's research focus is on developing implementable DL-based models for disease predictions.

Discussion Questions

In order to enhance audience participation, we will encourage audience input on the following:

- Do you think that AI can improve healthcare? In what aspects?
- How many AI solutions have you been exposed to in a healthcare setting? (will link to audience background)
- What metrics you commonly use to represent the accuracy of the prediction?
- What reliability means for you?
- Which one you prefer, a black-box prediction model that provides 95% prediction accuracy or a model that provides 88% accuracy with clear explanations of the predictions?
- Out of the 6 implementability factors proposed, can you please provide the top 3 from your opinion and rank them? (will link to audience background)

Timeliness

With the emerging need for AI and high hope in promising results, we believe that is the best time to discuss the implementability of AI models and how to facilitate the model transition from the research lab to clinical settings even for at least further clinical validation.

Intended Audience

We anticipate that this panel will be of interest to a wide range of individuals attending the AMIA Summit, including:

- Healthcare organizations executives
- Biomedical Informatics Researchers
- Data Scientists
- Clinicians
- Policymakers
- Vendors
- Students and trainees

Statement from Panel Organizer:

Laila Rasmy, the panel organizer affirms that all panel participants have agreed to participate and have contributed to the preparation of this document (as of Aug 26th, 2020)

Acknowledgments:

LR is supported by UTHealth Innovation for Cancer Prevention Research Training Program Pre-doctoral Fellowship (Cancer Prevention and Research Institute of Texas (CPRIT) grant # RP160015).

References

1. Shiffman RN, Dixon J, Brandt C, Essaihi A, Hsiao A, Michel G, et al. The GuideLine Implementability Appraisal (GLIA): development of an instrument to identify obstacles to guideline implementation. BMC Med Inform Decis Mak . 2005 Dec 27
2. Artificial Intelligence and Machine Learning in Software as a Medical Device | FDA [Internet]. [cited 2020 Aug 26]. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Virtually Ready: Technological Tools and Adaptations in a Children's Hospital during the COVID-19 Pandemic

Tiranun Rungvivatjarus[1,2], MD, Begem Lee[1,2], MD, Amy Chong[1,2], MD, Mario Bialostozky[1,2], MD, Jeannie Huang[1,2], MD, MPH, Cynthia L. Kuelbs[1,2], MD.
[1]Rady Children's Hospital, San Diego, California; and [2]University of California – San Diego, San Diego, California.

Abstract

In this current COVID-19 pandemic, healthcare systems across the country have undergone drastic changes in patient care and operational workflow. The distribution of up-to-date and reliable information to healthcare workers, patients, and the community is also of paramount importance.[1,2] Pediatric institutions are undergoing similar changes. In this panel, we outlined the various technological tools/adaptations and organizational changes that occurred at our academic institution in the midst of the COVID-19 pandemic. With panelists from various backgrounds and administrative roles, we will share our experiences in leading organizational changes from the informatics perspective and engage the audience in the different ways to leverage telemedicine, conserve personal protective equipment (PPE), provide clinical decision support, disseminate real-time organizational data, and optimize communication and access to care for families.In a pandemic, healthcare systems need to prepare for sudden and frequent changes in patient care and disruption of usual workflows.

Objectives

1) Identify EHR-based tools/adaptations to leverage telemedicine and promote physical distancing at an academic institution.
2) Interactively discuss solutions to common barriers institutions face while implementing institution-wide technological and operational changes.

Panel Structure:

1) Introduction: Moderator Tiranun Rungvivatjarus MD (5 min)
2) Panelist 1: Amy Chong MD (10min + 10min Q&A)
3) Panelist 2: Mario Bialostozky MD (10min + 10min Q&A)
4) Panelist 3: Cynthia L. Kuelbs MD (10min + 10min Q&A)
5) Panelist 4: Jeannie Huang MD (10min + 10min Q&A)
6) Wrap up (5 min)

Panelist 1: Amy Chong MD, Physician Informaticist and Tiranun Rungvivatjarus MD, Physician Informaticist
Division: Hospital Medicine
Topic/Issue discussed: clinical decision support

<u>COVID-19 Pediatric Clinical Pathway and Order Set</u>

Clinical pathways and order sets are essential to establish a collaborative, systematic approach to care delivery, decrease unnecessary variation in care, and ensure safe practices. A multidisciplinary team of content experts (infectious disease, critical care, hospital medicine, emergency medicine, nursing, and pharmacy) created a COVID-19 pediatric clinical pathway. This pathway provides instructions for safe transfer of patients, outlines admission criteria, defines high risk pediatric populations, recommends laboratory and imaging testing, and provides guidelines for escalation of care.

<u>COVID-19 Clinical Alert and the evolution of COVID-19 testing algorithm</u>

We implemented a nurse-administered screening alert in the EHR to identify patients at higher risk for COVID-19 at intake. Initially the screening was based on travel history and close contact exposure which eventually evolved to include symptoms suggestive of COVID-19. A positive screening triggered a pop-up clinical alert to direct the nurse to initiate isolation precautions, place an isolation precaution order, and place a flag in the chart for COVID-19 risk factors. We will also discuss the various COVID-19 testing algorithms and clinical decision support tools that help guide physicians during the time of limited testing availability. As testing availability increases, our institution also implemented testing of parents/caretakers in the inpatient setting.

Panelist 2: Mario Bialostozky MD, Physician Informaticist, Division:Emergency Medicine
Topic/Issue discussed: Ambulatory telemedicine and patient portal activation

Expanded Telemedicine and Remote Activation for Patient Portal

To minimize patient and staff exposures, our institution prioritized the rapid conversion of ambulatory in-person visits to telemedicine. We formed a multidisciplinary team (clinicians, analysts, clinic managers, and nurse educators) to quickly employ a series of operational and workflow changes. We implemented same-day urgent care telehealth visits as families have shown hesitation in seeking care in person. We created a dedicated COVID-19 Nurse Triage Line for families to streamline questions and provide telehealth visit referrals when indicated. To facilitate providers' adaptation to these changes, we created self-guided learning tutorials on how to conduct video visits as well as appropriate documentation and billing.

To further increase family's access to telemedicine visits, we developed a workflow to allow remote activation of patient portal accounts. Our institution utilizes the MyChart® patient portal (Epic Systems software, Verona, Wisconsin) as part of the EHR. Through MyChart®, families can participate in telemedicine visits by utilizing its video conference tool. The portal also allows patients and proxies secure access to medical records including notes and results, clinic team messaging, self-scheduling, and COVID-19 related information. Telemedicine visits for teens can be launched either through the teen or proxy portal account.

Panelist 3: Cynthia L. Kuelbs MD, CMIO
Topic/Issue discussed: Operational shift, Communication, and Community Role

Organizational Communication

As information and clinical care guidelines on COVID-19 frequently evolve, it is important for an organization to provide accurate and consistent information to staff. We follow the Healthcare Incident Command System (HICS) model[16] for emergency management planning, response, and recovery. All materials for staff and patients are approved through the HICS, and documents are archived in a central location, which can be accessed through links on the intranet and within the EHR. Emails and electronic newsletters are sent daily to update staff. Communication channels on topics such as PPE, testing, and telemedicine were created in secure chat messaging application, and messages are pushed out along with links to helpful documents and videos. Patient information is posted on the institution website, patient portal, and social media in collaboration with marketing and media affairs.

Community Involvement

Partnering with the genetic institute and public health, we implemented large-scale mass community testing and drive-through testing. Our initiative expanded to testing parents of hospitalized children, adult first responders, and health care workers.

Panelist 4: Jeannie Huang MD, Physician Informaticist, Division: Gastroenterology
Topic/Issue discussed: COVID-19 data tracking, visualization, and research.

COVID-19 Research

Dr. Huang will discuss ongoing COVID-19 research and how EHR tools play a role in data tracking and identification of subjects.

COVID-19 Dashboard

Having ready access to key and relevant data is crucial to hospital operations and resource allocation. We created a few COVID-19 dashboards with data updated daily. The dashboards contain various metrics, with each dashboard serving a different purpose. One dashboard focused on operational data and displayed COVID-19 testing volumes and results, resources (bed availability in the PICU, available ventilators (infant versus pediatric/adult), employee availability (number of staff out sick), and visit volumes at various clinical settings. By having a comprehensive, up-to-date display of organizational data, institutions can readily identify resource needs and utilization allowing for improved delivery of care.

Discussion questions to enhance audience participation

1. What barrier has your institution faced in implementing EHR-based tools or technological adaptation?
2. What types of solutions has your institution employed to overcome these barriers?
3. How have these changes affect your workflow and wellness?

All participants have agreed to take part on the panel

REFERENCES:

1. Gates B. Responding to Covid-19 — A Once-in-a-century pandemic? New England Journal of Medicine. 2020;382(18):1677-1679.
2. Hick J, Biddinger P. Novel coronavirus and old lessons — preparing the health system for the pandemic. New England Journal of Medicine. 2020;382(20):e55.

Challenges and Opportunities for Implementing Artificial Intelligence at the Speed of Technology Innovation During the COVID-19 Era

Panel Organizer/Moderator: Brett R. South, MS, PhD[1]
Panel: Wendy W. Chapman, PhD[2], Irene Dankwa-Mullan, MD[1], Michael E. Matheny, MD, MS, MPH[3], Yuri Quintana, PhD[4]
[1]IBM Watson Health, Cambridge, MA, USA; [2]The University of Melbourne, Victoria, Australia; [3]Vanderbilt University Medical Center, Nashville TN, USA; [4]Harvard Medical School, Boston, MA

Abstract
The COVID-19 pandemic has created multiple opportunities to implement Artificial Intelligence (AI) technologies in new ways that address the initial infectious curve (e.g., triaging patients and disseminating information during disease outbreaks), as well as the subsequent curves of pandemic sequelae (managing gaps in care of chronic conditions, addressing new and exacerbated mental health needs, and rectifying worsening health disparities. However, numerous challenges limit scaling development and application of AI technologies in healthcare settings, especially in the context of a rapidly evolving public health emergency. Data representing diverse patient cohorts are necessary both to train and to test systems but often are labor intensive to create and deidentify. The need for new codes and concepts can delay data availability. Biases in data must be identified, evaluated, and managed to mitigate downstream effects. System performance must be continuously monitored and validated as clinical information, such as disease transmission characteristics, become available. This panel will discuss these challenges and propose solutions that include ensuring adequate, equitable, and unbiased data sources are used for AI development, validation of AI in clinical settings, with the context of the rapidly evolving COVID-19 public health crisis as a discussion focus.

General Description

Artificial Intelligence (AI) is defined as the application of computer programs that mimic human intelligence by "learning" from available data inputs using a combination of supervised or unsupervised learning approaches[1]. AI has great potential to revolutionize healthcare particularly during times of greater need for data dissemination and introduction of efficiencies into healthcare processes. The COVID-19 pandemic has presented a global public health crisis with new opportunities and challenges in leveraging AI technologies. These challenges are related to new demands on public health, patients, and providers to manage the infectious disease itself, along with its sequelae, such as delayed care, management of chronic disease, new and exacerbated mental health problems, and worsening disparities in health and healthcare delivery. Each of these challenges represents a unique curve of the pandemic where implementation of AI technologies could help address needs.

The first of these challenges involves ensuring that adequate data representing diverse patient cohorts are available and used for system training and testing and, in turn, to generate clinical evidence for AI. In most health care settings, data are siloed and exist in aggregations that are not readily accessible due to patient privacy and confidentiality regulations, hence data sharing is difficult without significant effort. Limited or delayed access to patient data act as temporal obstacles to innovation. With emergence of a new pathogen, the creation of new data elements and concepts for diagnoses, tests, and treatments may further delay availability of data. Moreover, a second challenge includes addressing biases present in existing data used for training and testing of AI systems especially with regard to health disparities or lack of representation of social determinants of health that must be mitigated in real world application of AI. Another growing challenge is how to manage the care of a rapidly growing elder population that has an existing burden of chronic disease management and which has experienced disproportionate infection rates and complications of COVID-19[2-5].

During the COVID-19 pandemic and with the health system response, early successes will likely be focused in the area of implementing AI systems where the tools, for patient and provider end-users, have the potential to augment human decision making and help manage overwhelming information volume and demand. Furthermore, developed and deployed tools must adapt with the rapid growth of real-world data generation and prospective innovations, along

with a rapidly changing healthcare landscape and policy environment. Requirements for success rely on early and continuous engagement of end users and key downstream stakeholders, considerations for how the tools will be integrated into healthcare workflows, and implementing common frameworks to support real-world application of AI technologies that scale to increasing and ever-changing health system demands.

This panel session will address these challenges, opportunities and potential solutions in the current landscape of AI applications in healthcare, and present practical real-world examples of implementation and innovation using the COVID-19 public health crisis as the discussion focus.

Description of Panelists & Presentations

This panel, comprised of two women and two men from unique racial /ethnic backgrounds and geographic locations, as well as both industry and academic positions, seeks to provide diverse perspectives on addressing important challenges for AI. The moderator will open the panel discussion with a 10 minute presentation on the motivation for convening the panel and the significance of the subject matter. Panelists will then each give a presentation of 10 minutes providing examples and perspectives to the challenges identified above. The final 40 minutes of the panel will be available for questions from the audience and discussion between the panelists and audience.

Brett R. South, MS, PhD is a Biomedical Informatician with IBM Watson Health and the Center for AI, Research, and Evaluation (CARE). He was previously Senior Scientist in the Department of Biomedical Informatics, University of Utah and the Veterans Informatics and Computing Infrastructure (VINCI), Prior to completing his PhD, he worked as a Senior NLP Research Engineer for the Nuance Clinical Language Understanding group where he helped lead a group of 75 clinical language analysts tasked with large-scale semantic annotation of clinical corpora.

Dr. South will serve as the moderator of the panel. He will introduce each panelist with a brief statement about their experience and background, initiate the panel discussion, and facilitate the question and answer segment with the audience.

Wendy W. Chapman, PhD is Associate Dean for Digital Health and Informatics at the University of Melbourne, Australia. She leads the Centre for Digital Transformation of Health. Previous to her current role, she was chair of the Department of Biomedical Informatics at the University of Utah. She is a member of the National Academy of Medicine and was a co-author with Dr. Matheny and others on the recently released report on AI in Healthcare. Her research expertise is natural language processing applied to clinical notes.

Dr. Chapman will speak on the challenge of generating clinical evidence for AI in real-world health settings. She will speak to the gap between innovation and clinical validation and the need for infrastructure to support both innovators and health systems reach digital maturity to implement and validate AI and other digital health interventions. She will use examples related to the COVID-19 pandemic to describe some necessary elements of that infrastructure.

Irene Dankwa-Mullan, MD is Deputy Chief Health Officer and Chief Health Equity Officer at IBM Watson Health. She is nationally known for her contributions to population health and health disparities science, community-based participatory research, implementation and translation science. In her role, she assists with the global efforts of the Center for AI, Research and Evaluation to promote scientific evidence for Watson's technology. She also leads efforts to increase awareness, and transparency around the importance of diversity of identity, inclusion and thought in the industry to better represent the populations being served with cutting-edge technologies.

Dr. Dankwa-Mullan will discuss the ethical and social implications of rapid development of AI applications in healthcare settings, particularly in the arenas of clinical decision support, care management planning, risk stratification and prediction, and precision medicine. The COVID pandemic presents a number of important ethical and social issues that need to be addressed including resource allocation and priority-setting, public health surveillance and contact tracing, patient privacy and frontline or healthcare worker rights. In addition, there is the obligation from industry and researchers to ensure optimal clinical trials and vaccine development are being conducted in rapid time and ethically acceptable. She will describe the value issues and propositions reflected in various AI systems and the need to identify drivers of unwarranted outcomes, medical explainability and risk mitigation. She will summarize the various Ethical, Legal and Social Implications (ELSI) statements, established by governmental, global organizations and various industry stakeholders, and provide a framework for harmonizing rapidly changing technology innovation and with need for ethical and socially accountable AI.

Michael E. Matheny, MD, MS, MPH is the Co-Director of the Center for Improving the Public's Health Using Informatics, an Associate Professor of Biomedical Informatics, Biostatistics, and Medicine at Vanderbilt University Medical Center, and Associate Director of VINCI, HS&RD, at the Tennessee Valley Healthcare System VA. His area

of expertise is in the development and evaluation of machine learning risk prediction models in inpatient settings for a variety of clinical use cases as well as natural language processing.

Dr. Matheny will discuss the overall framework for the lifecycle of AI development that the National Academy of Medicine has proposed, including needs assessment, workflow mapping, target state definition, model development, implementation, surveillance, and de-implementation in healthcare applications (Chapter 6 of the publication), with examples of successes, failures, and challenges from his work and others that relate to elements of that framework[9]. He will also highlight some of the challenges faced due to tremendous data shifts that have occurred during the COVID pandemic and initial small volumes of training data.

Yuri Quintana, Ph.D. is Chief of the Division of Clinical Informatics, Beth Israel Deaconess Medical Center, and Assistant Professor of Medicine at the Harvard Medical School. His research is focused on developing innovative technologies and systems that empower collaborative care between healthcare professionals, patients, and families. Quintana and colleagues have created InfoSAGE Health for caring for frail older adults at home, the European Union's UNICOM Consortium for standardization of data collected via mobile apps, and Alicanto™, a global online collaboration platform for health professionals that supports virtual tumor boards, standardization of care treatment guidelines, and collaborations for international bioinformatics studies.

Dr. Quintana will speak on approaches to collecting patient-reported data to support big-data analytics for improving symptom management. Early data shows significant differences in COVID-19 cases and mortality based on race and ethnicity, but it is unclear to what extent this is related to genetics, clinical history, social determinants of health, or other factors. To understand this problem and perform data analytics at a global scale requires the standardization of symptom representation, social determinants of health, and ways to cross-link international databases for drug identifiers that vary by country.

Participation Statement:

All proposed panelists have agreed to participate in the panel.

Discussion Questions

1. What are the various ethical and social implications of rapid development and implementation of AI in healthcare particularly in the context of COVID-19?

2. What opportunities for innovation can be leveraged to ensure scalable, equitable, adequate and reproducible application of AI in the healthcare domains? How will these innovations affect the pandemic response in terms of information dissemination and information volume?

3. What community activities can be used as a model for development and implementation of AI in healthcare?

4. Are there community accepted recommendations or guidelines for evaluation of AI systems and what types of best practices ensure transparent and reproducible reporting of system performance?

5. What types of collaborations between industry/academia could be pursued to address the challenges in AI development and implementation addressed by this panel?

References

1. Benjamins JW et al. A primer in artificial intelligence in cardiovascular medicine. Neth Heart J. 2019; 27(9): 392-402.
2. Quintana Y, Fahy D, Crotty B, Jain R, Kaldany E, Gorenberg M, Lipsitz L, Engorn D, Rodriguez J, Orfanos A, Bajracharya A, Henao J, Adra M, Skerry D, Slack WV, Safran C. InfoSAGE: Supporting Elders and Families through Online Family Networks. AMIA Annu Symp Proc. 2018 Dec 5; 2018: 932-941.
3. Suzman R, Beard J. Global health and aging. National Institute on Aging. Published October 2011. Available from: https://www.nia.nih.gov/sites/default/files/2017-06/global_health_aging.pdf.
4. Agency for Healthcare Research and Quality. Multiple Chronic Conditions [Internet]. Rockville, MD: AHRQ. 2014 May [last reviewed 2017 December; cited 2020 March 10]. Available from: https://www.ahrq.gov/patient-safety/settings/long-term-care/resource/multichronic/mcc.html.
5. Matheny M, Israni ST, Ahmed M, Whicher D, editors. Artificial Intelligence in Healthcare: The Hope, The Hype, The Promise, The Peril. NAM Special Publication. Washington, DC: 2019, National Academy of Medicine.

The Digital Recipe for Success: Conducting Clinical Research Using Mobile Apps

Alexander Turchin, MD, MS[1,2], Mihir M. Kamdar, MD[2,3], Jukka-Pekka Onnela, DSc[4], Timothy R. Smith, MD, PhD, MPH[1,2]
[1]Brigham and Women's Hospital; [2]Harvard Medical School; [3]Massachusetts General Hospital; [4]Harvard T. H. Chan School of Public Health; all in Boston, MA

Abstract

Digital revolution is coming to clinical research. Mobile applications can both be used to study traditional – medication- and device-based – therapeutic interventions, and can also form the basis of medical interventions themselves. However, using mobile applications as the tools of clinical research is often different in many ways from using pen and paper or even online tools to record participants' data. Evaluating the effectiveness of mobile applications for diagnosis and treatment of medical conditions difference also presents many new opportunities and challenges compared to the conventional clinical trials. To help researchers, clinicians and developers of mHealth applications learn how to take advantage of the strengths and navigate the pitfalls of using mobile applications in clinical research, the panelists will discuss their experience studying mHealth apps in special patient populations, using them as innovative tools for clinical research and recruiting participants for mHealth clinical trials.

Introduction

As we begin the 3rd decade of the 21st century, mobile applications are ubiquitous in our life. They help us buy groceries, get from place to place and brag about our children's extraordinary achievements. But can they help us get healthier? Even though Apple App Store and Google Play Store are bursting with health and wellness-related apps, this fundamental question is surprisingly seldom answered in a rigorous fashion we are accustomed to when evaluating health benefits of medications, procedures and devices. Well-conducted clinical research studies assessing the effects of mobile are few and far between.

It is important to recognize that, just as mobile apps represent a paradigm shift from desktop PCs and flip phones of the last century, clinical research evaluating health mobile apps presents its own challenges that require novel solutions. In the proposed panel the speakers will share their experience and lessons learned on several important aspects of digital clinical research: a) studying mobile apps for serious ill patients; b) digital phenotyping using mobile apps; c) leveraging mobile apps to study quality of life and d) participant recruitment for clinical trials of mobile apps.

Lessons Learned from Studying Mobile Apps for Seriously Ill Patients

The digital health landscape is rapidly expanding, with over $7 billion invested in these technologies in 2019. While many of these applications are focused on wellness, very few digital therapeutics are aimed at addressing the needs of patients with serious illness. Patients with serious and terminal illness not only have symptoms that markedly impair quality of life, they also account for a significant portion of healthcare utilization and associated costs in the U.S. healthcare system). Furthermore, very few randomized controlled trials (RCTs) of digital therapeutic technologies demonstrating benefit for patients and on healthcare utilization outcomes exist.

ePAL is a digital therapeutic smartphone application designed to improve the management of cancer-related pain. Pain affects up to 66% of patients with advanced cancer, severely impacts patient quality of life, and is one of the leading causes of hospitalizations in oncology populations. In a randomized controlled trial of 112 patients with stage IV cancer with moderate to severe cancer pain, ePAL demonstrated significant improvements in pain severity, a significant reduction in negative attitudes towards cancer pain management, and a 69% reduction in the risk of hospital admission due to cancer-pain.

While ePAL is one of the first digital therapeutics to demonstrate improvements in both symptom management and healthcare utilization in patients with serious illness, studying a digital therapeutic in a trial of complex and ill patients was by no means a straightforward undertaking. In this session, the panelist will discuss lessons learned along the journey of ePAL from its conception to completion of its RCT.

Digital Phenotyping Using Smartphones

The phenotype of an organism is a collection of traits, such as enzyme activity, hormone levels, and behavior. Given the increasing availability of high-throughput genotyping technologies, it is clear that phenotyping is one of the key rate-limiting and cost-limiting factors in human genetics and in our understanding of disease more broadly. The call for large-scale phenotyping, in particular that of behavior–which presents special challenges for phenomics because of its temporal nature and context dependence–has been growing in the literature as efforts in precision medicine are becoming more concrete. The advancement of behavioral phenotypes is especially welcome because conventional laboratory-based methods are expensive, subjective, and do not scale well.

The ubiquity of smartphones presents an opportunity to measure different social and behavioral markers, offering a scalable solution to the phenotyping problem. We have defined and operationalized the concept of *digital phenotyping* as the "moment-by-moment quantification of the individual-level human phenotype *in situ* using data from personal digital devices". This research is continuation of our previous work on using cell phone data to study social networks and communication dynamics started in 2005.

As part of our efforts in this area, we have developed the Beiwe research platform (https://github.com/onnela-lab) for smartphone-based high-throughput digital phenotyping. The Android and iOS smartphone apps that constitute the front-end of the configurable platform collect various types of active data, such as surveys and audio diary entries, and passive data, such as GPS and accelerometer data in their raw (unprocessed) form, and anonymized phone call and text messages logs. The Beiwe back-end, which is based on Amazon cloud computing infrastructure–making it both scalable and globally accessible–collects, stores, and analyzes the collected data. These data enable us to study behavioral patterns, social interactions, physical mobility, gross motor activity, and speech production among other phenotypes.

Beiwe was designed to provide data collection and analysis tools for research settings to take advantage of new opportunities created by the prevalence of smartphone users. When we searched for data collection platforms in 2012, we discovered that no such tool existed for research. This was our primary motivation for developing our own platform. Some of the key design considerations were the following: (i) the platform should collect and store raw data; (ii) the platform should have strong encryption to protect the privacy of subjects; and (iii) the platform should be easily customizable to accommodate the needs of different studies.

Although collection of high-grade data is important, the main intellectual challenge in smartphone-based digital phenotyping is arguably moving from data collection to data analysis. Use of passive data, data from smartphone sensors and logs that can be collected unobtrusively without any burden on the subject, enables long-term follow-up of subjects, but it also presents formidable data analysis challenges. Because smartphones are consumer grade devices and their use patterns vary from person to person, analysis methods have to be robust and accommodate a wide range of use cases. For example, to learn about physical activity and mobility one has to account for different placements and orientations of the phone.

The panelist will discuss some of the opportunities and challenges that are present in this line of work, both from the point of view of data collection and data analysis. The panelist will also highlight some key results from areas where this approach has been successful and has yielded new insights.

Using Mobile Apps to Measure Quality of Life

The World Health Organization defines quality of life (QOL) as "A state of complete physical, mental, and social well-being not merely the absence of disease . . .". QOL has become increasingly important in the evaluation of a patients' overall health, especially those suffering from chronic diseases. Historically, QOL has been measured through medically validated instruments such as the Quality of Life Scale (QOLS) and administered via paper at in-person medical appointments. These instruments traditionally provide a composite score by assessing a patient's physical, social, emotional, and cognitive capacity as it relates to their subjective wants in life. While significant progress has been made in how QOL is evaluated, many limitations remain. However, many of these instruments evaluate "causal indicators" of QOL rather than QOL itself. Traditional QOL instruments also rely on active input from patients, making them highly susceptible to response, acquiescence, and recall bias amongst others. Furthermore, obtaining repeated measurements at regular intervals can be difficult and a burden to patients with chronic illnesses and sporadic medical appointments. As mobile technology is embraced by the medical community, new opportunities are arising for how, when, and where QOL can be measured.

Many of the domains evaluated through current QOL instruments can be measured passively from the sensors within mobile devices. A patient's communication patterns (e.g., quantity and length of calls and texts) can serve as a proxy measure for sociability. On-board gyroscopes and accelerometers can help determine a patient's mobility and activity level. Screen on and off time can be used to compute a patient's sleep patterns. Voice recordings offer the potential to gauge affect, hydration, pain, and fatigue. Developing novel ways to examine how an individual interacts with their phone, such as text response time, text length, and number of typos or "autocorrects" could help clinicians measure cognitive function, mood, fine motor control, and overall digital health. Finally, an index which summarizes these measurements into a single composite score can help easily provide a quick and comprehensive assessment of a patients' quality of life without any active input from the patient themselves.

Mobile devices can not only be used a surrogate for in-person paper instruments, but also afford significant advantages over traditional methods. The increasing ubiquity of these devices allows for cheap and effective scalability. They are present at nearly all times and are often the most important everyday carry item for patients allowing for real-time monitoring of a patients' QOL at shorter intervals. The various built-in sensors offer reliable passive methods for measuring QOL with a higher degree of precision and specificity than paper instruments by eliminating the need for active patient input and subsequently survey biases. In the panel, we will discuss how these advantages can be leveraged to create novel, scalable and user-friendly approaches to measuring QOL.

D^2R^2: Digital Recruitment for Digital Research

Anyone who has ever conducted a clinical trial knows: recruiting participants is one of its greatest challenges. Less than 20% of clinical trials finish enrollment on time and 50% never reach target enrollment. Could mHealth clinical trials be different?

Recruiting participants for studies of mobile health apps can be both easier and harder. Unconstrained by having to physically dispense medications or perform procedures, studies can be conducted across the country or even internationally with a push of a button. At the same time, effectively identifying potential participants hundreds of miles away mandates a completely different approach from the traditional outreach to the investigators' and their colleagues' patients. It can require skillful navigation of social media, understanding how different generations and social groups conduct their online lives and designing ads that will be a clickbait rather than an eyesore.

This part of the panel will share the lessons we have learned on all of the above and more while conducting an international clinical trial of a mobile diabetes app Control:Diabetes (http://controldiabetes.bwh.harvard.edu). We will discuss our experience designing Facebook ads and identifying the ones that were more effective; exploring different outreach channels and user targeting within Facebook and other online forums; and navigating both institutional and Facebook regulations.

No matter how good a novel medication or a snazzy mobile app is, it is impossible to conclusively prove it without having enough people agree to try it out. Our experience and the pitfalls we have traversed will serve as a guide for future mHealth innovators as they get ready to upend modern healthcare in a whirlwind of digital revolution.

Anticipated Audience

Mobile health and wellness apps and their benefits are an increasingly popular topic of conversation among both healthcare professionals and patients and their caregivers. We therefore expect that the proposed panel will attract strong interest from a broad swathe of attendees who use, design, implement and study mobile health apps. These will include clinicians, mHealth developers, trialists and many others who have ever wondered looking at their device: will this app actually help me?

Discussion Questions

1. Have you used mobile applications to gather data for clinical research? What were the advantages and disadvantages of that approach?
2. Have you been involved in a project where you would have needed access to more granular device data (from smartphones or wearables) than what was made available?
3. What challenges have you faced when recruiting participants for an mHealth clinical trial? How did you overcome them?

Statement of Agreement to Participate

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

Machine Learning to improve care delivery: Opportunities and Challenges

Kavishwar B. Wagholikar, MBBS PhD [1,2], Jyotishman Pathak PhD[3],
Hongfang Liu PhD [4], Nigam Shah MBBS PhD [5], Shawn N. Murphy, MD PhD [1,6]

[1] Massachusetts General Hospital, Boston, MA; [2] Harvard Medical School, Boston, MA;
[3] Cornell University, New York, NY; [4] Mayo Clinic, Rochester MN;
[5] Stanford University, Palo Alto, CA; [6] Mass General Brigham, Boston, MA;

Abstract

Stratification of patients into clinical cohorts is a critical task for academic research and clinical operations— accurate cohorts facilitate epidemiological analysis and enable efficient population programs for clinical interventions, respectively. Cohort identification technologies based on machine learning (ML)— referred to as phenotyping – have been pioneered by the academic community, and they are being increasingly used for optimizing clinical operations. The resulting data-infrastructure offers an unique opportunity to drive the optimal utilization of existing therapies. However, successful adoption of phenotyping approaches requires advances in methodology as well as technological and cultural changes in the healthcare ecosystem. This panel brings together experts from a diverse set of health systems and aims to provide insights for using ML in the clinical setting. Each of the panelist will describe specific projects at their health-system that either facilitates or directly use phenotyping in the clinical setting.

General description of the panel and issues that will be examined

A. Medicine is undergoing more rapid change than at any time in the last 200 years. The costs of delivering care under the current model and the opportunity from new technologies, promise to deliver a revolution in care. Molecular insights in chronic disease highlight the need for improved resolution in diagnosis and management. The advent of massive streams of structured data has open health care to rigorous analytics and predictive quantitative biology.[1] Practice of medicine is evolving towards integrating population analysis for efficiently discovering and managing patient sub-groups who will benefit from evidence-based therapies.

B. Research consortiums such as eMERGE (gwas.org), PGRN (pgrn.org), OHDSI (ohdsi.org), NIH collaborator, ACT (actnetwork.us), and PCORnet, have pioneered the development and sharing of EHR-driven phenotyping algorithms. The majority of these algorithms consist of collections of manually constructed rules using mostly structured EHR data (billing codes, medication prescriptions, and laboratory and test data). More recently, a greater number of the phenotype algorithms, many of which are being shared on PheKB (PheKB.org), are now using Natural Language Processing (NLP). There is need for research on rapid development, dissemination and deployment of such algorithms.

C. Even after phenotyping algorithms show strong performance in research settings, they fail to deliver meaningful benefit in terms of improved care for patients. Currently, estimating the net benefit of an algorithm prior to deployment is difficult and rarely done. This is in part because there is no well-established framework to quantify the implementation costs and benefits of machine-learning enabled care workflows. Components of this framework include computational costs, such as an evaluation of the accuracy of the model on live data and monitoring for performance degradation over time as well as the organizational effort required to integrate the output of the model into the appropriate workflow, including training of personnel, altering workflows, changing culture, and responding to unintended consequences.[2-4]

D. Informatics for Integrating Biology and the Bedside (i2b2) is an open source clinical data analytics platform deployed at over 200 sites for querying patient data to identify research cohorts. i2b2 has been adapted to build multi-institutional networks to allow federated querying across institutions like the PCORnet network. I2b2 and other such research data repositories have a central role to play in leveraging population analysis to drive clinical intervention. Several novel extensions are being developed to use i2b2 for supporting clinical operations. [5-7]

E. This panel will bring together experts from a diverse range of health systems to examine the challenges posed for applying phenotyping techniques in the clinical setting. The panelists will describe their experience with projects involving these challenges and share how their audience can leverage the lessons learnt. The resulting discussion

will serve as guidance to the audience for evolving their institutional infrastructure for applying phenotyping for clinical use-cases.

Panelists' Presentations

Jyotishman Pathak
Dr. Jyotishman Pathak is a Professor and Chief of Health Informatics at Weill Cornell Medical College, Cornell University. He has extensive experience in biomedical ontologies, vocabularies and terminology standards. Dr. Pathak's group focuses on developing novel informatics methods and their applications in scalable, robust and high-throughput phenotyping from electronic health records (EHRs). Dr. Pathak will present his research on novel computational techniques in ontology-based data encoding, semantic inferencing and standards for sharing computable definitions of phenotype representations. [8]

Hongfang Liu
Dr Liu is Professor of Biomedical Informatics, and Chair, Division of Digital Health Sciences at Mayo Clinic. She leads the Biomedical Informatics group at Mayo Clinic and has published extensively on secondary use of EHR for translational science research and health care delivery improvement through clinical NLP. In this presentation Dr. Liu will specifically present work on two algorithmic challenges for bringing phenotyping into clinical practice: i) given medical history of a patient, how to efficiently retrieve a list of genetic tests to be recommended, and ii) given genetic or genomic information of a patient, how to retrieve their clinical implication information and summarize it for physicians and counselors? Dr. Liu will also discuss the impact of data fragmentation related to phenotyping using a record linkage system in the Rochester Epidemiology Project. [9,10]

Nigam Shah
Dr. Shah is Associate Professor of Medicine (Biomedical Informatics Research), Stanford University. His research combines machine learning, text-mining, and prior knowledge in medical ontologies to enable the learning health system. Dr. Shah's research group is part of the Center for Biomedical Informatics Research at Stanford. Dr. Shah will discuss his group's experience in synthesizing factors that affect implementation costs and benefits of machine-learning enabled care workflows into a framework for performing utility assessment of ML models prior to their clinical deployment. He will describe another project for improving advanced care planning using phenotyping.

Shawn Murphy
Dr. Shawn Murphy is Professor of Neurology at Harvard Medical School and is Chief Research Information Officer at Mass General Brigham (MGB). He is recipient of the Donald A.B. Lindberg Award for Innovation in Informatics, and is the chief architect of i2b2. Dr. Murphy will describe the high-throughput phenotyping projects at MGB. He will describe the developed methods and speak on the evolution of use of population research data to drive clinical innovation. Finally, he will describe the roadmap for i2b2 to support high-throughput phenotyping and patient stratification for clinical care, and will describe the ML extensions to i2b2 that are under development. [11-14]

Kavishwar Wagholikar
Dr. Wagholikar is Assistant Professor of Medicine and Assistant in Laboratory of Computer Science at Massachusetts General Hospital. His research focuses on development of methodology and tooling to facilitate the use of ML in the clinical setting. [15,16] Dr. Wagholikar will serve as moderator on the panel.

Need for the panel

The movement towards bundled payment model, has created the incentive to stratify patients into cohorts to increase the efficiency of healthcare. This is leading to the increasing use of phenotyping for cohort identification technologies for optimizing clinical operations. Such a data driven ecosystem has presented a unique opportunity for healthcare institutions to characterize their clinical population to facilitate utilization of well-established as well as novel therapies.

Plan for the integration of content presented by each panelist

Dr. Wagholikar will serve as moderator for the panel. He will initiate the panel with a brief introduction of the topic, and will introduce the speakers. Dr Pathak will present efforts of his group at Cornell on implementing a standard for sharing phenotyping models. Next, Dr. Liu will present efforts at Mayo Clinic on specific algorithms to facilitate use of ML for clinical applications and also present her work on data fragmentation. Dr. Shah from Stanford will outline his lab's research on a framework to project the implementation costs and benefits of machine-learning for enabling clinical workflows. The final presenter, Dr. Murphy will describe efforts on patient stratification at MGB

biobank and how this is being utilized to drive care-coordination projects. Each of the panelist will describe specific projects at their institution that either facilitates or directly uses phenotyping in the clinical setting.

List of discussion questions to enhance audience participation

- How is clinical cohort identification different from population management?
- What are the key changes in the field that have enabled the current advances in phenotyping?
- What are the primary gaps that need to be resolved?
- Who are the key players in this field?
- What is the role of a data framework for clinical programs?
- What advances are necessary to facilitate use of Artificial Intelligence for Cohort Identification?
- What is the effect of data fragmentation on phenotyping algorithms?
- What role does IOT/digital therapeutics play?
- How do we scale and standardize these solutions across providers?
- How do we move from a reactive to a predictive stratification model given current data fragmentation?

Plan for interaction between panelists and the audience

The panelists will interact with audience through questions and commentaries discussing and reflecting on challenges that current clinical-informatics infrastructure faces to implement patient stratification. The panelist will take audience questions at the end of their individual presentations and towards the end, after all the panelist presentations. More than a third of the session will be reserved for audience participation. All the presenters have agreed to take part on the panel.

References

1. Macrae CA. Discovering New Diseases to Accelerate Precision Medicine. Transactions of the American Clinical and Climatological Association 2017;128:83-9.
2. Shah NH, Milstein A, Bagley Ph DS. Making Machine Learning Models Clinically Useful. JAMA 2019.
3. Jung K, Kashyap S, Avati A, et al. A framework for making predictive models useful in practice. medRxiv 2020:2020.07.10.20149419.
4. Liu VX, Bates DW, Wiens J, Shah NH. The number needed to benefit: estimating the value of predictive analytics in healthcare. J Am Med Inform Assoc 2019;26:1655-9.
5. Wagholikar KB, Mandel JC, Klann JG, et al. SMART-on-FHIR implemented over i2b2. Journal of the American Medical Informatics Association : JAMIA 2017;24:398-402.
6. Wagholikar KB, Dessai P, Sanz J, Mendis ME, Bell DS, Murphy SN. Implementation of informatics for integrating biology and the bedside platform as Docker containers. BMC Med Inform Decis Mak 2018;18:66.
7. Wagholikar KB, Ainsworth L, Vernekar VP, et al. Extending i2b2 into a framework for semantic abstraction of EHR to facilitate rapid development and portability of Health IT applications. AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science 2019;2019:370-8.
8. Jiang G, Solbrig HR, Kiefer R, et al. A Standards-based Semantic Metadata Repository to Support EHR-driven Phenotype Authoring and Execution. Stud Health Technol Inform 2015;216:1098.
9. Ravikumar KE, Wagholikar KB, Li D, Kocher JP, Liu H. Text mining facilitates database curation - extraction of mutation-disease associations from Bio-medical literature. BMC Bioinformatics 2015;16:185.
10. Wang L, Olson JE, Bielinski SJ, et al. Impact of Diverse Data Sources on Computational Phenotyping. Front Genet 2020;11:556.
11. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association : JAMIA 2010;17:124-30.
12. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. Journal of the American Medical Informatics Association 2016;23:909-15.
13. Murphy S, Wilcox A. Mission and Sustainability of Informatics for Integrating Biology and the Bedside (i2b2). EGEMS 2014;2:1074.
14. Yu S, Ma Y, Gronsbell J, et al. Enabling phenotypic big data with PheNorm. Journal of the American Medical Informatics Association : JAMIA 2017.
15. Wagholikar KB, Fischer CM, Goodson AP, et al. Phenotyping to Facilitate Accrual for a Cardiovascular Intervention. Journal of clinical medicine research 2019;11:458-63.
16. Wagholikar KB, Estiri H, Murphy M, Murphy SN. Polar Labeling: Silver standard algorithm for training disease classifiers. Bioinformatics 2020.

Consortium for Clinical Characterization of COVID-19 by EHR (4CE)

Griffin M Weber, MD, PhD[1], Gabriel Brat, MD[2], Shawn Murphy, MD, PhD[3], Diane Keogh[4]

[1]Harvard Medical School, Boston, MA; [2]Beth Israel Deaconess Medical Center, Boston, MA; [3]Masscusetts General Hospital, Boston, MA; [4]i2b2 tranSMART Foundation, Boston, MA

Abstract

There are several large, national and international projects to build informatics infrastructure to analyze the electronic health record (EHR) data of patients with COVID-19. However, aggregating data from multiple EHRs only works if you can trust the final results. This means being able to talk to the people at each site who know the data best, to understand the local clinical guidelines, coding practices, data quality problems, and other factors that affect the data. In March, 2020, we launched an international effort called the Consortium for Clinical Characterization of COVID-19 by EHR (4CE), which brings together more than 100 informatics experts, statisticians, and physicians representing 200+ hospitals around the world. We run analyses locally within sites and share aggregate results centrally, where we review the data together and iteratively fix any issues. Through this process, we have identified key laboratory tests associated with COVID-19 disease severity.

Description of Panel

The COVID-19 pandemic has caught the world off guard, reshaping ways of life, the economy, and healthcare delivery. Data in electronic health records (EHRs) should be widely available to study COVID-19 but have not yet been effectively shared across clinical sites, with public health agencies, or with policy makers. To address this problem, in March 2020, the i2b2 tranSMART Foundation launched 4CE (pronounced "foresee"), an international consortium representing 200+ hospitals[1]. The novel aspect of 4CE is that we recognize the complexities of EHR data and the need to directly involve the local data experts, not only in the data collection, but also in the development of research questions and the data analyses. We try to move fast, believing that early intelligence is worth more than complete intelligence later. To do this, we avoid roadblocks that typically slow down informatics projects, such as building or installing new software, or the regulatory hurdles involved in sharing patient-level data. Instead, we ask participating sites to run analyses locally, using simple existing tools, like SQL and R scripts, and only share aggregate counts and statistics centrally with the rest of the 4CE consortium (Figure 1). Our secret sauce is that we review and validate the data as a group, identify and fix data quality problems, and ask sites to repeat the analyses until everything is right. Through multiple cycles of data verification, we iteratively clean up the data and gain confidence that the findings we are seeing are real. Because we can do this quickly, we go from research question to results in just a few weeks.

Figure 1. The 4CE workflow includes: (1) weekly Zoom meetings and thousands of Slack messages to coordinate activities across 200+ institutions worldwide, (2) sites running analyses locally and only sharing aggregate counts and statistics centrally through a data upload and validation tool, and (3) iteratively reviewing the data through interactive visualizations and going back to informatics experts at each site to fix data quality problems.

In this panel, we will provide an introduction to the 4CE, describe its architecture, discuss challenges and lessons learned, and explain how other institution can participate in the effort.

Panelists

Griffin Weber (Moderator) – 4CE Technical Architecture

With 4CE Phase 1, a consortium of international hospital systems of different sizes utilizing mostly i2b2 [Murphy 2010], but also OMOP and other platforms, was convened to address the COVID-19 pandemic. The group, coordinated through the i2b2 tranSMART Foundation, initially focused on admission comorbidities and temporal changes in key laboratory values during infection. After establishing a common data model that leverages a COVID-19 ontology developed by the ACT Network[2], each site generated four data tables of aggregate data as comma-separated values (CSV) files. Database scripts were provided to sites to help them create these files (https://github.com/CovidClinical). These non-interlinked files encompassed COVID-19 positive patients: daily case counts, demographic breakdown, daily laboratory trajectories for 14 labs, and recorded diagnoses by ICD code (Figure 2). These are shared centrally, merged, and displayed on the https://covidclinical.net.

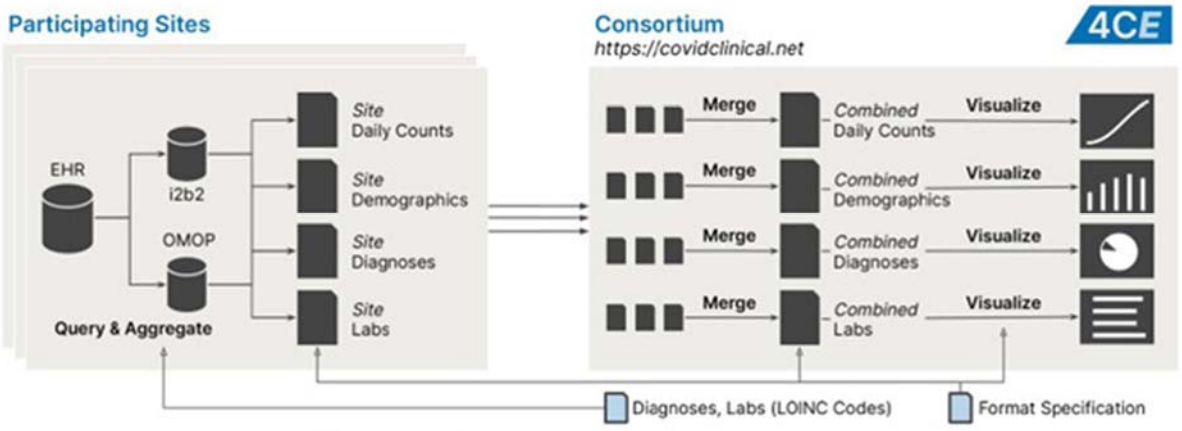


Figure 2. Starting from i2b2, OMOP, or other data models, sites generate four CSV files that are centrally merged.

Although asking sites to create four CSV files sounds simple, in Phase 1 we learned there are many complexities. Sites have different IRB and aggregate count obfuscation policies; sites that received COVID-19 patients as referrals do not know when they originally tested positive; most COVID-19 patients are sent home right away and therefore have little data; determining which patients were in the ICU is challenging for many sites, especially since parts of the hospital were repurposed as temporary ICUs; deaths are reported in different ways across sites, and date of death is often not available; there are numerous problems with combining laboratory tests from sites, with LOINC code mapping problems, incorrect units, and different forms of the same test; date formats vary by site and country (e.g., mm/dd/yyyy vs dd/mm/yy); and, sites modified their CSV files by changing the order of columns or including additional columns that were not part of the file definitions. We addressed formatting problems in "Phase 1.1" by asking sites to upload their CSV files to a Data Upload and Validation website, but other data problems are due to limitations of the source systems, which we figured out through our iterative approach of work with local data experts. We have recently begun 4CE Phase 2, which enables more sophisticated data quality checks and local analyses (e.g., machine learning models) by distributing R scripts to sites. Though, still only aggregate data are shared centrally in Phase 2.

Shawn Murphy – Validating a COVID-19 Disease Severity Algorithm

From 4CE Phase 1, we learned that direct outcome measures, including ICU and death, ventilator settings, and vitals signs are difficult for sites to collect since they are not coded in standard ways across hospitals. Instead, for Phase 1.1, we created a computational phenotype for severe COVID-19 diseases that uses more commonly available codes that can serve as proxies for intubation/non-invasive ventilation, shock (hypotension), and ARDS (ventilatory dyssynchrony and mismatch). Specifically, we categorize patients with severe disease if they were admitted for COVID-19 and have ANY of the following: (1) Lab Test (LOINC): PaCO2 or PaO2 (regardless of the result); (2) Medication (RxNorm, ATC): sedation/anesthesia or shock/severe cardiac disease; (3) Diagnosis (ICD-10): ARDS, ventilator-associated-pneumonia; (4) Procedure (ICD-10): endotracheal tube insertion or invasive mechanical ventilation. Our hypothesis is that each of these codes are highly specific for severe disease, meaning that they would only be given to very ill patients. While the sensitivity of each individual code might be low (a given patient won't have all the codes in our list), the overall sensitivity of the algorithm is high because it searches for many codes. At sites with coded ICU or death data, we use these as a "silver standard" to validate our algorithm. However, manual chart review appears to be the only way of obtaining a gold standard at most sites.

Gabriel Brat – 4CE Results to Date

In 4CE Phase 1, 96 hospitals in the US, France, Italy, Germany, and Singapore contributed data to the consortium for a total of 27,584 COVID-19 positive cases and 187,802 collected lab values. Case counts and laboratory trajectories were concordant with existing literature. Laboratory test values at the time of viral diagnosis showed hospital-level differences that were equivalent to country-level variation across the consortium partners. Our first preprint was available by mid-April, a mere four weeks after our first 4CE Zoom meeting and subsequently published in *Nature Digital Medicine*. We created interactive visualizations of the data (Figure 1, right) and made all the aggregate data publicly available on our 4CE website (https://covidclinical.net). These visualizations and the ability of anyone to look at and comment on the data were essential in helping us identify and correct data quality problems and interpret the findings. The website visualization are built using open tools, including Jupyter Notebooks (https://jupyter.org) and Altair (http://altair-viz.github.io/) to enable reproducible research. In 4CE Phase 1.1, we focused on patients who were admitted to a hospital for COVID-19 and compared patients who progressed to "severe" COVID-19 disease to patients who were never severe.

Diane Keogh – Outreach and Sustainability

Many of the people and sites that are part of 4CE are also involved in other large informatics efforts to study COVID-19, including All of Us, N3C, PCORNet, ACT, and TriNetX[2,3]. 4CE is complementary to these in that we are working with the local data experts to iteratively understand the differences between sites and cleanup data quality problems. The knowledge we are generating in 4CE includes our computational phenotype for severe COVID-19 disease; important details on key lab tests (e.g., variations we have found in how sites measure troponin); and site differences in data collection and coding practices (e.g., how death and race are reported). The i2b2 tranSMART Foundation is using several methods to disseminate this knowledge to the broader community and encourage additional institutions to participate in 4CE, including promoting the efforts and progress of 4CE on the Foundation website (https://transmartfoundation.org) and through social media; proving updates about 4CE in the monthly i2b2 tranSMART newsletter and monthly community meetings; and, developing and promoting a sponsorship program for 4CE through the Foundation.

Impact

4CE is complementary to other large ongoing efforts to collect and analyze clinical data on COVID-19. In 4CE we are taking a deep dive into the data quality and unique aspects of more than 200 hospitals worldwide representing tens of thousands of patients hospitalized for COVID-19. We are learning how to stratify patients into levels of disease severity, how to identify data elements are most robust across sites, and how to leverage local expertise within sites to study COVID-19 or future health crises. It is critical that others involved in multi-institution COVID-19 research projects know about 4CE, our approach, our findings, and the lessons we have learned, so that they can better understand the data coming from different sites and use the data most effectively.

Discussion Questions

1. In what ways are 4CE working with other national/international efforts to study COVID-19?
2. How much effort is required from institutions to participate in 4CE?
3. To what extent is 4CE scalable? What are its limitations?
4. How does 4CE's definition of COVID-19 severity compare to the many others that have been developed?
5. How can sites improve their data quality, documentation, coding practices, etc., to make it easier to use EHRs to study future pandemics?

References

1. Brat GA, Weber GM, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. Nature Digital Medicine. 2020. 3, 109 (2020). https://doi.org/10.1038/s41746-020-00308-0
2. Shyam Visweswaran, Michael J Becich, Vincent D'Itri, et al. Accrual to clinical trials (ACT): a clinical and translational science award consortium network. JAMIA open. 2018 Oct;1(2):147-52. PMID: 30474072
3. Haendel M, Chute C, Gersing K, The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. Journal of the American Medical Informatics Association. ocaa196. August 17, 2020. https://doi.org/10.1093/jamia/ocaa196

Computable Phenotypes :
A Primer on Creation, Evaluation and Applications

[1]Mark Weiner, MD, [2]Jeffrey Brown, PhD, [3]Nicholas Tatonetti, PhD, [4]Lisa Bastarche, MS

1. Weill Cornell Medicine, New York, NY, 2. Harvard Medical School, Boston, MA
3. Columbia University Irving Medical Center, New York, NY,
4. Vanderbilt University, Nashville, TN

Computable phenotypes (CPs) are programmable specifications of patient characteristics that enable identification of cohorts of similar individuals from electronic health data. The appropriate development, evaluation and application of CPs are all active areas of informatics research that are critical to our understanding of the scope and course of disease and effective therapeutics. This panel will describe the idiosyncrasies of clinical data that need to be considered in defining a CP. We will discuss the variety of methods used to evaluate the accuracy of a CP and address the implications of these differences. The panel will also provide practical examples of applications of CPs in clinical research, GWAS and PheWAS studies, and show how different decisions in developing and evaluating the CP can impact the results and interpretation of research findings.

Learning Objectives:

By the end of this panel, attendees will be able to:

1. Explain the impact of clinical data idiosyncrasies on the generalizability and use of computable phenotypes
2. Identify the presence of bias in computable phenotype development and evaluation
3. Describe how differences in computable phenotype specification can impact outcomes of prospective and retrospective clinical research as well as PheWAS and GWAS studies

A computable phenotype (CP) is a rigorous specification of one or more patient characteristics or conditions using data derived from an electronic health record (EHR). The CP is typically programmed as a query of a clinical data warehouse that identifies cohorts of patients meeting the specification. The components of a computable phenotype include demographic characteristics, the presence or absence of diseases, procedures, patterns of medication use, laboratory results, location and provider of care characteristics, and clinical observations and events. CPs may be used to assess quality of care, identify eligible patients for prospective clinical trials, compare the effectiveness and safety of therapeutics through retrospective analysis, analyze genomic data, and develop predictive algorithms. A great deal of published research exists on defining and evaluating computable phenotypes across a broad spectrum of clinical conditions, and resources such as eMerge have emerged that help collate and organize the breadth of computable phenotypes.

While investigators agree on the value and importance of computable phenotypes, there is less consensus on the approach to their development, assessment of fitness-for-purpose, evaluation methods and applications. This panel brings together experts who will discuss CPs and invite dialog from the audience regarding their experiences in developing and using CPs. The panelists will use real

world examples to discuss and dissect assumptions made when creating and interpreting CPs. At a time when EHR based research is more popular than ever, the presentations and discussions prompted by this panel are vital to the achieving consensus on the appropriate development and use of CPs and challenging many assumptions about CPs.

The topics to be covered by the panelists are as follows:

Constructing the phenotype - Nicholas Tatonetti

Computable phenotypes are used frequently to define patient cohorts. The task of identifying patients affected by a particular condition using EHR data often requires confronting unexpected complexities that arise from fragmentation of care, health care processes, and health IT implementation, as well as varying documentation and coding styles. These complexities can often cloud the meaning of diagnosis codes that would otherwise seem straightforward.

As a result, CP development requires a rigorous analysis of a set of related diagnosis codes, along with other clinical data such as the presence of relevant medications, labs, procedures and clinical observations. CP developers must be mindful that all of these data types have their own idiosyncrasies in their timing, frequency, biases and regional variation in their use. This part of the panel will highlight the variety of idiosyncrasies in these data that complicate simple queries of discrete data and can lead to different interpretation of results at different sites despite the use of the same computable phenotype.

Evaluating the phenotype – Mark Weiner

Evaluation of computable phenotypes typically involves comparisons of patients who meet and do not meet the specification of the CP against a gold standard set of patients who are known to definitively have, or not have a condition of interest. There are several commonly used metrics used to measure and compare CP performance, including reporting on the proportion of patients meeting the CP specification who truly have the disease of interest (positive predictive value). A related approach explores a set of patients known to have disease, often from a registry, and reports on the proportion of those patients who meet the CP specification (Sensitivity). The harmonic mean of the positive predictive value and the sensitivity is reported as another evaluation metric, the F-score. While these evaluations produce quantitative values, they can be subject to bias based on the characteristics of the data source in which the CP is evaluated. Bias can be introduced if the computable phenotype is applied to a cohort with a high prevalence of disease, or if the cohort of patients known to have a disease was drawn from a registry that included patients that were known to meet the CP specification, or applied to a data source with incomplete data capture of CP parameters. Furthermore, the F score and its components do not provide any information on the false negative rate and the negative predictive value, so the ability of the CP to rule out a diagnosis, especially for patients with related but different diagnoses is less certain.

This part of the panel will explore the idiosyncrasies of CP evaluation and discuss the implications for their application in clinical operations and research.

Applications of the phenotype

Retrospective and prospective analysis and QI type issues – Jeffrey Brown

Once a computable phenotype is defined, it is often used to support clinical operations and research. This is where the idiosyncrasies in CP definition and evaluation may have significant impact on the interpretability of the results. In this part of the panel, we will consider the complexities of creating interpretable CPs, with a focus on issues that arise when CPs are intended to be used at multiple sites. A syntactically sound CP will reliably produce results when executed across different datasets, but the interpretation of results may vary between sites in important ways that impact the reproducibility of findings. A definition of a diagnosis like diabetes may allow heterogeneity in the disease severity at different sites, affecting apparent quality assessments. Attempts to create more homogeneous cohorts may have the unintended effect of narrowing the cohorts more than is desirable. For prospective or retrospective research studies, the computable phenotype may have a high positive predictive value, but the patients identifiable through the CP may not be representative of all patients with the condition.

This part of the panel will provide real world examples of CP applications in prospective and retrospective research and describe some of the analytical and interpretive issues in using CPs.

Computable Phenotypes and genomic analysis – Lisa Bastarche

Computable phenotypes are now widely used in GWAS and PheWAS studies, supported by the proliferation of EHR-linked biobanks. Analysis of EHR data has been shown to reproduce genetic associations found in more traditional case/control studies. CPs have enabled the development of novel methods to explore the relationship between phenotypes and genotypes. However, the integration of genomic and EHR data has also given rise to new analytic challenges. Issues of bias may impact the generalizability of the findings when applied in a real-world environment.

This part of the panel will provide real-world examples of the application of computable phenotypes in GWAS and PheWAS studies.

Questions for discussion:

What considerations are necessary when applying a Computable Phenotype at a new clinical site that was developed and tested at one group of clinical sites?

Under what circumstances can an application of the same computable phenotype produce non-comparable results across the institutions?

Is the same computable phenotype appropriate for defining cohorts for Quality Measurement as it is for comparative effectiveness or safety research?

How do biases in the evaluation of a computable phenotype affect analyses of real-world data?

As we develop new research methods and applications with EHR data, how do we ensure that they produce interpretable and reproducible results?

Statement on participation

All participants have agreed to take part on the panel.

Using Synthetic Data to Enhance Antimicrobial Use Data Quality within the National Healthcare Safety Network

Wendy L. Wise, MPH, PMP[1,2]; Senthil K. Nachimuthu, MD, PhD, FAMIA[3,4]; Joseph Chuong, MS[5]; Jay P. Kim, RPh, MBA[6]; Shuai Zheng, PhD[1]

[1]Division of Healthcare Quality Promotion, Centers for Disease Control and Prevention, Atlanta, GA; [2]Lantana Consulting Group, East Thetford, VT; [3]IDEAS Center, VA Salt Lake City Healthcare System, Salt Lake City, UT; [4]Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT; [5]Asolva, Pasadena, CA; [6]Cerner Corporation, Kansas City, MO

Abstract

The Centers for Disease Control and Prevention's (CDC) National Healthcare Safety Network (NHSN) Team partnered with the Veterans Affairs (VA) Informatics, Decision-Enhancement and Analytic Sciences (IDEAS 2.0) Center at the VA Salt Lake City Health Care System to create an Antimicrobial Use (AU) synthetic data set (SDS) and a vendor software validation process. NHSN will require all vendor systems electronically submitting AU data to NHSN be validated by January 2021 using the AU SDS.

The panel will provide an overview of the AU SDS and validation process along with a discussion of challenges faced including lessons learned during the implementation of this unique validation effort for NHSN data. Panelists will speak about the motivation, insights and methods behind the AU SDS and the validation process. Additionally, two vendors that have successfully completed the AU SDS validation process will share their experiences completing the validation process.

Learning Objectives

1. Understand how synthetic data is used to test external vendor systems and ensure receipt of high-quality data.

2. Describe how statistical time-series data simulation techniques are used to create a data set based on an absorbing Markov model, including calibrating the model to over-represent rare use cases.

3. Compare and contrast different vendors' approaches for implementing synthetic data set validation.

4. Identify lessons learned by all stakeholders that were involved in the AU SDS initiative.

General Description of Panel and Issue(s) that will be Examined

NHSN recently added the AU SDS as a tool that provides a new way for vendors to validate the AU data aggregation logic used in their software solutions, relieving some of the burden usually placed on hospital antimicrobial stewards to manually validate data. The AU SDS validates the aggregation of both the Days of Therapy (DOT) numerator and the Days Present denominator according to the NHSN AU Option protocol definitions to ensure the data quality of both counts. To use the data set, vendors load the AU SDS into their production systems and process it just as they would with their clients' real data. The vendor then uploads the resulting tabular format file to a CDC hosted web application for validation. The web application checks the uploaded file against the correct answer keys and returns the validation results, which consist of descriptive error feedback on incorrect rows. Once the vendor has a successful file, they send it to the CDC Team for verification.

This panel includes representatives from the CDC, the VA and vendors that successfully completed the AU SDS validation process. The panel will provide an overview of the AU SDS and validation process along with a discussion of challenges faced including lessons learned during the implementation of this unique validation effort for NHSN data. Panelists will address the motivation, insights and methods behind the AU SDS and the validation process. Additionally, two vendors, Asolva and Cerner, will share their experiences completing the validation process. A brief description of each panelist's presentation is described below:

- CDC will discuss the impetus for the first ever required use of synthetic data for the NHSN application. Further, CDC will describe the SDS validation initiative process from inception to implementation including

the importance of creating a vendor neutral solution, the pilot process, the validation mechanism, NHSN application changes, and documentation (instructions, FAQs, tracking).

- VA partnered with the CDC to support the NHSN's AU SDS project. The IDEAS Center at the VA Salt Lake City Health Care System used an absorbing Markov model to simulate the admission, transfer and discharge of patients in various wards of a fictitious hospital as well as their medication administration for all the 91 drugs and 4 routes of administration included in the NHSN Antimicrobial Use and Resistance – AU Option Protocol. The simulated data included multiple inpatient and outpatient units and covered several positive and negative test cases. These test cases cover a variety of scenarios, both simple and complex, described in the NHSN AU Option Protocol. The VA team created detailed instructions and packaged the simulated data and instructions for release to the vendors. The instructions also specified the tabular format in which the monthly summary results should be submitted for validation. The drug administration table used local codes along with a mapping table that provided mappings to RxNorm to mimic the use of local and standard terminologies in real world EHR systems.

  The VA team also generated the monthly numerator and denominator summary reports for every month of the year covered in the simulated data set. These summaries serve as the 'answer keys' used to validate the results submitted by the participating vendors. The VA and CDC teams jointly defined error conditions and descriptions to explain an incorrectly large or small number, or a row that must or must not be present in the summary reports uploaded by the vendors. The answer keys and error descriptions were then provided to the CDC team for their use in a web application, built by the CDC team, to validate the uploaded results and provide a report for incorrect values and corresponding explanations. The VA and CDC teams worked with vendors to pilot the data set and the validation system before general release.

- Asolva was invited to participate in the CDC AU SDS validation process to help establish a method by which all future AU submissions into NHSN from any vendor could be properly vetted for accuracy. Prior to this effort, there was only HL7 Clinical Document Architecture (CDA) syntax validation with the CDA Validator. While the validator was effective in validating the technical grammar of the CDA file, it did nothing to ensure the quality of the summarized data and whether the DOT and denominator formulas were being applied correctly. An answer key was needed.

  To begin the SDS validation process, the CDC provided data files simulating extracts from a fictitious hospital. Like taking a test, Asolva was asked to respond with our answers, which in this case would be our DOT and denominator calculations. Since Asolva's Medici® software product was designed to universally accept data from any EHR system, data were easily loaded, mappings were set up, and required output was produced within one or two days. As an extra validation step, Asolva was able to graphically see the calculated DOT, days present, and admission metrics through the Medici® Analyzer screen. Detection of data anomalies was followed by drilling down into a specific data point and viewing the underlying raw data. Once initial validation was complete, Asolva generated the required output and submitted it to NHSN's SDS validation site for judging. The validation site would then respond back with a clear list of any errors that it found, and these discrepancies could be further investigated and resolved.

  Overall, the SDS validation process was straightforward and not time-consuming. Further, the process supported data quality assurance and it has been incorporated into Asolva's software release cycle and routine testing.

- Cerner will share our experience from completing the validation process of the Antimicrobial Usage reporting including successes and opportunities for future certifications. Additionally, Cerner will share clients' experience submitting to NHSN.

Topic Relevance

Analyzing antimicrobial use over time and between contributors requires semantic interoperability, i.e., that the datasets contributed by hospitals across the country all have the same meaning. With over 1900 hospitals now contributing to the AU Option, it has become increasingly important to ensure compliance with the NHSN AU Option Protocol. NHSN has provided adjustment for ward patient mix and hospital characteristics from the beginning but this validation will also ensure that its prescribed methods for tabulation are being observed by each contributor. For example, common mistakes can lead to downward bias in reported numerators and denominators. The standard is updateable such that implementation of new standards can be confirmed by new releases of the validation protocol.

Conclusion

The AU SDS and associated validation process has provided a method for vendors to test their software against an expected result based on the NHSN AU Option Protocol thereby providing them greater confidence in their products. Additionally, this has provided NHSN additional assurance in the quality of AU data received from vendor systems which are ultimately used to produce risk-adjusted AU metrics and inform hospital stewardship programs. Similar initiatives could be broadly implemented to improve data quality obtained from external systems for healthcare data and beyond.

Discussion Questions

1. Why did the CDC ask the participants to upload the monthly summary reports in tabular format rather than as the CDA files that were being submitted to the CDC? Are there any plans for validation using the CDA?
2. How much effort (e.g. estimated number of hours) did it take each vendor to pass the validation?
3. What were the common reasons that caused submissions to fail validation? How did the CDC and vendors work together to overcome it?
4. Is the current test design complete? Should the validation process add any additional test cases?
5. In order to facilitate the debugging process, what additional feedback information should the validation website provide?
6. Could vendors list the software updates which may invalidate or outdate the validation? How often do they happen?
7. Is vocabulary maintenance error-prone? Do we need to add procedures to validate the vocabulary table used in the vendor system?
8. What issues did the panelists face or what lessons were learned from implementing the AU SDS and the answer keys?
9. How did CDC/VA test and validate the synthetic data and the answer keys before they were released?
10. Do vendors need to repeat the validation? How often will a 'pass' result be recognized?
11. Thoughts on moving towards FHIR for both validation as well as data submission in production?
12. Will CDC implement synthetic data sets and data validation for all protocols that are being reported by healthcare organizations?
13. Can a similar process be used to ensure validity and consistency of data being gathered from multiple reporting organizations around COVID-19 or other emerging health conditions?
14. Does this have any implications beyond CDC for other types of data that are being submitted to the government? How about non-healthcare data?
15. What led the CDC to create a synthetic data set for validating AU? Were there any prior work that gave the idea? Was it based on a need that was recognized?
16. How was the validation process received by the vendors? How did they respond and how was the working relation between the vendors and the CDC?
17. How can this help other areas of health informatics, such as repeatable and reproducible science?
18. What will happen to the vendors who do not validate their system by the end of 2020?
19. Do the CDC or vendors have any results on the improvement in data quality or any useful findings from the data submitted to CDC before and after the AU SDS validation was implemented?

Statement from Panel Organizer

All panel participants have agreed to take part in this panel and have contributed to the submission of this abstract.

Health Information Technology Interoperability Standards to Advance the Precision Medicine Initiative

Teresa Zayas-Cabán, PhD[1], Robert R. Freimuth, PhD[2], Ida Sim, PhD, MD[3], Stephanie Devaney, PhD[4], Tracy H Okubo, CSM, PMP[1]
[1]Office of the National Coordinator for Health Information Technology, Washington, DC; [2]Mayo Clinic, Rochester, MN; [3]University of California, San Francisco, San Francisco, CA; [4]National Institutes of Health, Bethesda, MD

Abstract

The Precision Medicine Initiative (PMI) ushers in a new area of health care delivery centered on tailoring prevention and treatment to an individual's unique characteristics. Success of the PMI hinges on the ability to collect and analyze large electronic datasets. PMI projects led by the Office of the National Coordinator for Health Information Technology (ONC) are accelerating standards development for health information technology (IT) for research; specifically, standards needed to collect relevant data for the PMI. This panel will provide timely information about these projects, being conducted in close collaboration with the All of Us Research Program, a foundational component of the PMI led by the National Institutes of Health. The panel will describe other innovative collaborations to realize the PMI's potential—Sync for Science, Sync for Genes, and Advancing Standards for Precision Medicine. The panel will actively encourage participant interaction and feedback, essential to informing and strengthening future priorities.

Introduction

The Precision Medicine Initiative (PMI) is a nationwide initiative that moves health care delivery treatment and prevention strategies away from a "one-size-fits-all" approach.[1] The PMI vision is to individually tailor healthcare delivery by considering characteristics that make each person unique, including factors such as environment, lifestyle, and biology. The PMI's bold mission is to "enable a new era of medicine through research, technology, and policies that empower patients, researchers, and providers to work together toward development of individualized care."

Launched in 2015 with the National Institutes of Health (NIH), National Cancer Institute,[2] Food and Drug Administration,[3] and Office of the National Coordinator for Health Information Technology (ONC) as participating agencies, the PMI has expanded to include the U.S. Department of Health and Human Services Office for Civil Rights and Health Resources and Services Administration, the Department of Veterans Affairs, the Department of Defense, and the Department of Energy. The PMI leverages "the unique expertise and history of each agency" to advance its ambitious vision.[1] The breadth of mission represented by these eight agencies and their research partners ensures that collaborative efforts consider, among others, important factors of diversity.

Advancing Standards that Support Health IT Interoperability for Precision Medicine Research

Involved with the PMI before its official launch, ONC's continuing role is to:

- Accelerate innovative collaboration around pilots and testing of standards that support health IT interoperability for research;
- Adopt policies and standards to support privacy and security of cohort participant data; and
- Advance standards that support a participant-driven approach to patient data contribution.[4]

The 21st Century Cures Act (Cures Act) (2016) directed ONC to advance nationwide interoperability and enable sharing of electronic health information, as well as advance health IT solutions for the PMI.[5] The Cures Act galvanized ONC efforts to make needed data interoperable for clinical care and precision medicine research, to accelerate evidence generation to improve health.[6] Data standards and quality, in particular, have been highlighted as areas that need to be addressed to advance precision medicine.[7]

Accelerating Health Research through Comprehensive Data Collection and Dissemination

The *All of Us Research Program* (*All of Us*), led by NIH, is a key component of the PMI. *All of Us* is developing a national research cohort of at least a million volunteers to "accelerate research and improve health."[8,9] The greatest

value of this groundbreaking initiative, unprecedented in size and scope, is the volume and richness of data collected and made available directly to participants and to researchers since May 2020 via a web-based data repository. The repository enables researchers to learn about "more concise diagnosis, prevention, and treatment" by considering "individual differences in lifestyle, socioeconomic factors, environment, and biologic characteristics."[10] *All of Us* enrollment began in May 2018; by August 2020, more than 277,000 participants had contributed biospecimens. More than 80% of participants are from groups that are historically underrepresented in biomedical research. Electronic health record (EHR) data on more than 225,000 participants from 34 sites have been collected. Repository data include participant-provided information (e.g., demographic or lifestyle surveys), EHR data, baseline physical measurements, biospecimens (for whole genome sequencing and other clinical and environmental assays), and some mobile health data.[9] Effective and efficient collection and analysis of such data, particularly EHR data, genetic test results from biospecimens, and mobile health data, as well as emerging data types (e.g., environmental data or social determinants of health) will be realized with the judicious use of appropriate standards.

Advancing Health Information Technology as a Foundation for Precision Medicine Research

ONC also partnered with NIH to launch several research projects that also support the PMI's ambitious vision.

- *Sync for Science* enables patients to more easily share health records with researchers through an open, standardized, application programming interface (API).
- *Sync for Genes* seeks to seamlessly share patients' genomics test results by expediting the use of standards at the point of care and for research.[10]
- *Advancing Standards for Precision Medicine*, launched in 2018, focuses on improving interoperability by developing, testing, and balloting health IT standards in identified priority areas, including mHealth; wearables and sensors; and social determinants of health.[11]

ONC and NIH recognize the importance of broad informatics and research community involvement in this effort and value the opportunity to disseminate current efforts through panel presentations, reflect on the challenges that remain and the path forward, and invite feedback and consideration from all participants in the session.

Panel Objectives and Presenters

The aim of this panel is to provide an overview of the PMI and ONC's current portfolio of work and priorities, share preliminary results and findings from initiatives supporting PMI, discuss how findings from these projects inform real-world standards development and adoption, discuss the data needs of the *All of Us* Research Program, summarize additional work that remains, and invite participants to provide feedback and reflections on the work conducted to date, and help identify priorities for future work. The panel brings together the informatics and research experts leading this work at ONC and NIH, respectively.

Ms. Tracy Okubo (moderator and organizer) is a Senior Program Analyst/Project Manager at ONC. She will introduce the session and moderate the discussion.

Dr. Teresa Zayas-Cabán (panelist) is the Chief Scientist at ONC; her division leads ONC's participation in the PMI. She will describe the PMI and provide an overview of ONC's relevant portfolio of work and priorities, including how this work is advancing the availability of data needed for precision medicine research.

Dr. Robert Freimuth (panelist) leads a variety of research studies, including *Sync for Genes*, focused on the intersection of medical informatics, and genomics to speed the translation of advances in genomics to clinical practice. He will discuss current efforts to share genomic testing results in a standardized way using Health Level Seven International's (HL7®) Fast Healthcare Interoperability Resources (FHIR®) genomic resource.

Dr. Ida Sim (panelist) is a primary care physician, informatics researcher, and entrepreneur. She leads research on the use of mobile apps and sensors to improve health and manage disease for populations and individuals and to make clinical research faster and less expensive. She will present on emerging work regarding development and piloting of Open mHealth and FHIR standards for wearables and sensors.

Dr. Stephanie Devaney (panelist) is the Chief Operating Officer of the *All of Us* Research Program and previously led the coordination of the PMI from the Office of the Chief of Staff at the White House. She will discuss *All of Us* data collection objectives, issues encountered to date with collecting data of interest, and goals moving forward.

Panel Discussion Questions

- What are other standards testing efforts currently underway that are relevant to the PMI and *All of Us*?
- What types of organizations are involved in those efforts?
- What needs or gaps have those efforts addressed and what remains?
- A high priority of *All of Us* is to enroll a diverse group of participants and return value to them; how have standards helped to increase the value of data in research studies aimed at diversity and improving health disparities?
- What are some of the steps that can be taken to address those gaps?
- What other data types relevant to the PMI, *All of Us*, and precision medicine require additional standards development work? What are the existing or anticipated barriers to this work?
- What barriers to broader implementation and adoption of standards are addressed through these projects?

Panel Learning Objectives

1. Participants will understand ONC's PMI standards development and testing pilot projects, and their accomplishments.
2. Participants will learn the objectives and current status of the *All of Us* Research Program as well at the program's current and emerging data needs.
3. Participants will understand the gaps in health data standards that currently obstruct effective support for research.
4. Participants will learn how to address these gaps in health data standards in the coming years.

Conclusion

This panel will address progress and impediments to the current health IT interoperability standards work being led by ONC in collaboration with *All of Us* in support of the PMI. The panelists seek to stimulate a rich discussion and gather participant input that will inform and strengthen the work at ONC and NIH. Discussion will also include the current and future data needs of *All of Us* as well as the actions required to address precision medicine research data standards-related challenges from multiple perspectives.

Statement of Participation

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

References

1. The White House. The precision medicine initiative. 2015. https://obamawhitehouse.archives.gov/precision-medicine Accessed March 11, 2020.
2. National Cancer Institute. NCI and the Precision Medicine Initiative®. 2017. https://www.cancer.gov/research/areas/treatment/pmi-oncology Accessed October 4, 2019.
3. Altman RB, Prabhu S, Sidow A, et al. A research roadmap for next-generation sequencing informatics. *Sci Transl Med.* 2016;8(335):335ps10.
4. Office of the National Coordinator for Health Information Technology. Precision medicine. 2020. https://www.healthit.gov/topic/scientific-initiatives/precision-medicine Accessed March 11, 2020.
5. 21st Century Cures Act, Pub. L. No. 114-225, 130 Stat. 1034. 2016.
6. 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program, 45 C.F.R. § 170 and 171. 2020.
7. Adamo JE, Bienvenue Li RV, Dolz F, et al. Translation of digital health technologies to advance precision medicine: informing regulatory science. *Digit Biomark*. 2020;4(1):1-12
8. *All of Us* Research Program Investigators. The "All of Us" Research Program. *N Engl J Med*. 2019;381(7):668–76.
9. National Institute of Health. *All of Us* Research Program. 2020. https://allofus.nih.gov/ Accessed August 5, 2020.
10. Williams MS, Taylor CO, Walton NA, et al. Genomic information for clinicians in the electronic health record: lessons learned from ClinGen and eMERGE. *Front Genet*. 2019;10:1059.
11. Garcia SJ, Zayas-Cabán T, Freimuth, RR. Sync for Genes: Making clinical genomics available for precision medicine at the point-of-care. *Appl Clin Inform*. 2020;11(02):295-302.
12. Sim I. Mobile devices and health. *N Engl J Med*. 2019;381(10):956-68.

Data-level Linkage of Multiple Surveys for Improved Understanding of Global Health Challenges

Girmaw Abebe Tadesse, PhD[1], Celia Cintas, PhD[1], William Ogallo, RPh, PhD[1], Skyler Speakman, PhD[1], Aisha Walcott, PhD[1], Komminist Weldemariam, PhD[1]
[1]IBM Research — Africa, Nairobi, Kenya

Abstract

Data-driven approaches can provide more enhanced insights for domain experts in addressing critical global health challenges, such as newborn and child health, using surveys (e.g., Demographic Health Survey). Though there are multiple surveys on the topic, data-driven insight extraction and analysis are often applied on these surveys separately, with limited efforts to exploit them jointly, and hence results in poor prediction performance of critical events, such as neonatal death. Existing machine learning approaches to utilise multiple data sources are not directly applicable to surveys that are disjoint on collection time and locations. In this paper, we propose, to the best of our knowledge, the first detailed work that automatically links multiple surveys for the improved predictive performance of newborn and child mortality and achieves cross-study impact analysis of covariates.

Introduction

Neonatal and child death is still a critical global health challenge. Particularly, the neonatal period represents the most vulnerable time for a child's survival. In 2016 alone, 2.6 million deaths, or roughly 46 of all under-five deaths, occurred during this period. On current trends, more than 60 countries will miss the target of a related sustainable development goal that aims to reduce neonatal death to less than 12 deaths per 1000 live births by 2030. Unfortunately, about half of those countries will not even hit the target by 2050[1]. Moreover, challenges, such as COVID19 pandemic, will also adversely impact the above target. Thus, data-driven techniques need to employed to extract and provide insights, to domain experts and policy makers, e.g., in order to facilitate the application of well-suited interventions.

To this end, large-scale cross-country surveys such as Demographic Health Survey (DHS)[2], Knowledge Integration (KI)[3], and Performance Monitoring for Action (PMA2020)[4] were collected to understand these global health challenges, particularly maternal, newborn and child health (MNCH). Though the collection of such surveys incurred significant resources (human capital, time, money, etc.), they are handy for domain experts and policymakers to extract insights for a better understanding of the problem domain and to potentially devise the right interventions. Examples of the benefits of data-driven techniques include the ability to predict the likelihood of the outcome, the detection of vulnerable groups, and the identification of informative covariates that are more predictive of the outcome. Even for a specific country, there are multiple surveys available in the MNCH domain. These surveys might differ in study samples, the information collected, and the time periods in which the surveys were conducted. However, these existing surveys are often analyzed in silos, which fails to capture distinct characteristics available across these surveys, and potentially results in poor prediction performance of critical events, such as neonatal mortality. The use of multiple surveys, with different characteristics via linkage, can improve outcome prediction capability and cross-study impact analysis of covariates. Furthermore, insight extraction from single survey is partly questionable, particularly, when the survey for a specific country suffers from "data inefficiency" challenges including small sample size and data imbalance (rare occurrence of an outcome). For example, the PMA subset of Ethiopia has less than 400 samples related to neonatal death. Thus, despite huge investments and efforts to collect these surveys, the "data inefficiency" challenges have raised several concerns about how insights, extracted in silos from different surveys, might affect decision making in population health outcomes.

Thus, we propose a novel approach for data- or record-level linkage of different surveys in order to improve the trustworthiness of insights generated from surveys. These surveys might be disjoint sets of samples and covariates collected. We show how we improve the prediction performance in each survey and achieve cross-survey (cross-study) explainability by analysing the impact of distinct covariates in predicting the outcome across studies. This provides utilisation of multiple studies to understand a specific outcome of interest, which could have required large datasets (often requiring further data collection) in the traditional setting. More recent machine learning solutions for

multiple sources are often driven by computer vision research (i.e., image modality); and are not thoroughly studied for survey data types, such as DHS and PMA. In addition, techniques developed in the global health domain need to be simple and easily interpretable for domain experts, unlike current advanced multi-domain techniques.

In our approach, first, we project disjoint surveys into equal-dimensional covariate representations so that samples in these surveys could be compared directly. Note that domains, studies, surveys, and datasets are used interchangeably in this paper. Similarly, features, covariates, and attributes are also used interchangeably. We employ different approaches to reduce the original covariate dimensions in these surveys, taking into account the degree of covariate overlapping among them. The similarity of samples across disjoint surveys is computed using a distance metric, from which close neighbors are extracted. Then, unique covariates of close neighbors are aggregated and combined with the original study, thereby augmenting the covariate representation of the original survey for better predictive performance. The proposed approach is simple, and it provides data-level integration of different surveys, which can minimise resource utilisation (time, money, computational power) compared to extra data collection or more sophisticated post-model linkage practices (e.g., transfer learning). We validate the proposed approach by using multiple surveys and linking scenarios such as linking completely disjoint surveys, linking surveys with a few common covariates. The linkage could also be cross-study, cross-country, and both cross-country & cross-country. Encouraging performance is achieved across these validation scenarios. For example, we found that vital signs, such as systolic and diastolic blood pressure in the second Alliance for Maternal and Newborn Health Improvement study (AMANHI-2), are quite predictive of samples with neonatal mortality both in its original study (i.e., AMANHI-2) and when aggregated to other studies, such as AMANHI-1 and DHS. Though DHS contains large dimensional covariate space, its linkage to other studies did not provide significant improvement, thereby suggesting the relative advantage of the AMANHI-2 survey compared to DHS. Moreover, studies with small sample sizes (e.g,. PMA of Ethiopia) were found to benefit more from the proposed linkage. Generally, the proposed framework involves the utilization of multiple surveys to: 1) maximise the efficiency of a particular study by incorporating discriminative and unique covariates from another study; 2) improve prediction performance and identify distinctively useful covariates across studies; and 3) provide domain experts and policy makers with additional insights on existing studies and further recommendations for future data collection efforts.

Motivation and Related Work

Data-driven approaches require enormous training samples for better outcome prediction. However, enough data is not often the case, and many critical problems still suffer from data scarcity, imbalance or lack of labels since collecting sufficient amounts of data in such cases might be difficult, expensive, or even sometimes not possible at all. Existing datasets for global health challenges, such as child mortality, partially share these challenges. It is straightforward to exploit two datasets via data-driven techniques when they have complete overlapping samples or features. However, it is difficult when two datasets possess partial or no overlapping of their samples or features. This is the case among existing MNCH-related surveys, which are collected at different times, with potentially different populations and a few unique covariates in each survey.

Existing approaches in the space of small/inefficient data and linkage of multiple studies incline towards data-augmentation [5–7], generation [8–10] and transfer learning [11–15] techniques. Though augmentation and generation help to create artificial samples, the intelligibility and diversity of these generated samples is still limited by the small data size in order to create samples with enough variance. Furthermore, insights derived from synthetic data is hardly convincing for domain experts in healthcare. Transfer learning is the most extensively explored solution for small data cases in the literature, and its common strategies include multitask learning [15], few-shot learning [11,14,16,17], domain adaptation [12] and a combination of these [18]. However, these transfer learning techniques are relatively complex and hence less interpretable for domain experts as interactions among multiple domains (datasets) happen at the latent level. Data-level linkage is not well exploited, and existing works focus on the context of combining same-entity records from different data sources [19–23], e.g., removing replicated samples or creating longitudinal profiles of individuals [24], which do not fit well with our prediction task. Generally, the majority of the existing solutions in the state-of-the-art multi-source modelling are tuned towards image modality and are not equally applicable to the MNCH-related tabular surveys.Thus, a novel approach needs to be designed for combining, linking, and semantically cross-referencing data from multiple surveys for a better understanding of MNCH challenges.

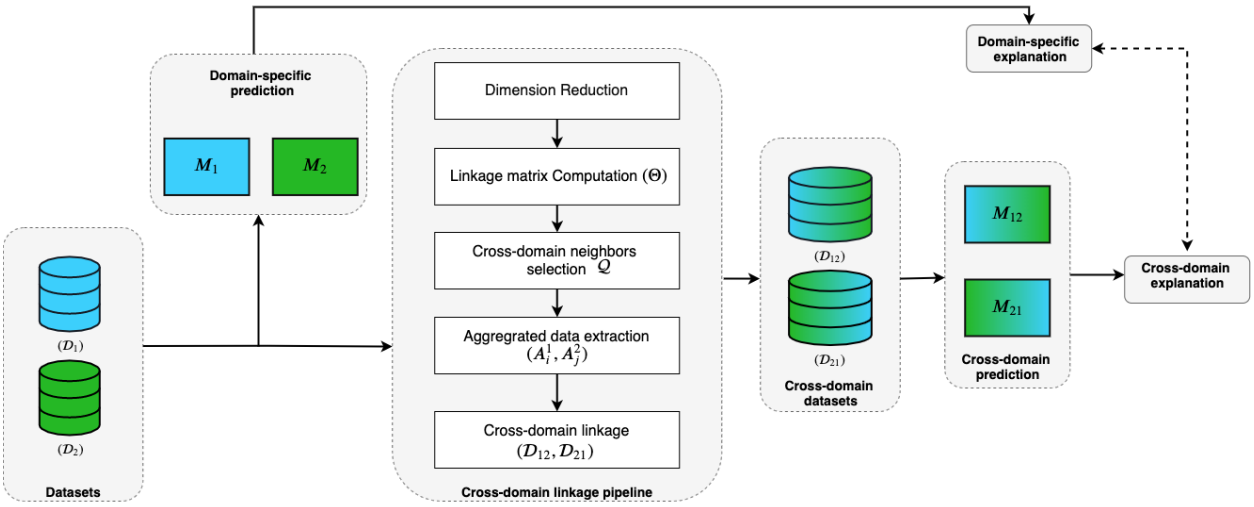


Figure 1: Block diagram of the proposed data-level linkage of disjoint surveys.

The Proposed Automated Data-level Linkage

In this section, we provide an overview of the proposed approach followed by a detailed discussion on how we achieve the data-level linkage of different surveys.

Overview

Linkage of surveys could be applied at different stages of a prediction pipeline, e.g., before modelling (pre-model) and after modelling is done on each survey (post-model). Our proposed approach employs pre-model linkage of surveys as shown in Figure 1, which provides better interpretability and enables straightforward analysis of covariates from other surveys in correctly predicting the outcome. Formally, given two surveys ($\mathcal{D}_1$ and $\mathcal{D}_2$) with similar outcomes of interest (e.g., child mortality), $\mathcal{D}_1 = (S_i^1)_{i=1}^N$ and $\mathcal{D}_2 = (S_j^2)_{j=1}^W$, where N and W represent the numbers of samples in $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. S_i^1 and S_j^2 in $\mathcal{D}_1$ and $\mathcal{D}_2$ are represented by K and L covariates, respectively, i.e., $S_i^1 = (f_{i1}^1, f_{i2}^1, \cdots, f_{iK}^1)$ and $S_j^2 = (f_{j1}^2, f_{j2}^2, \cdots, f_{jL}^2)$. Note that their covariate vectors, $\mathbf{f}_1$ and $\mathbf{f}_2$, might not have any common covariates between them. We then apply a specific model for each of $\mathcal{D}_1$ and $\mathcal{D}_2$, i.e., $M_1 : \mathcal{D}_1 \rightarrow y^1$ and $M_2 : \mathcal{D}_2 \rightarrow y^2$, where y^1 and y^2 represent the outcome labels in $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. The proposed cross-survey linkage takes $\mathcal{D}_1$ and $\mathcal{D}_2$ and provides linked datasets, $\mathcal{D}_{12}$ or $\mathcal{D}_{21}$, respectively, upon which cross-survey models, M_{12} and M_{21}, are applied and their performance is compared against the baseline models, M_1 and M_2.

Dimension Reduction for Similarity Computation

The two datasets, $\mathcal{D}_1$ and $\mathcal{D}_2$, might contain different number of covariates in their original representations. Thus, we, first, employ a dimension reduction technique in order to project $\mathcal{D}_1$ and $\mathcal{D}_2$ into equal-dimensional representations, $\hat{\mathcal{D}}_1$ and $\hat{\mathcal{D}}_2$, respectively, so that we can compare the similarity of samples across $\hat{\mathcal{D}}_1$ and $\hat{\mathcal{D}}_2$ directly using a distance metric. To this end, we employ four approaches to reduce feature dimensions (see Figure 2). The approaches are based on: common covariates, feature importance score, principal component analysis (PCA) and autoencoder (AE).

Common covariates-based dimension reduction involves $\mathcal{D}_1$ and $\mathcal{D}_2$ having a few common covariates. Examples of common covariates between DHS survey and AMANHI-2 of KI study for Ghana is shown in Table 1. Though these two surveys employed different coding techniques, common covariates could be extracted from their similar descriptions. Thus, $\hat{\mathcal{D}}_1 \in \mathbb{R}^\gamma$ and $\hat{\mathcal{D}}_2 \in \mathbb{R}^\gamma$, become the new representations of $\mathcal{D}_1$ and $\mathcal{D}_2$ to compute samples' similarity using a distance metric, where γ represents the number of common covariates between the two surveys.

PCA-based dimension reduction does not assume the need of common covariates between $\mathcal{D}_1$ and $\mathcal{D}_2$, and principal

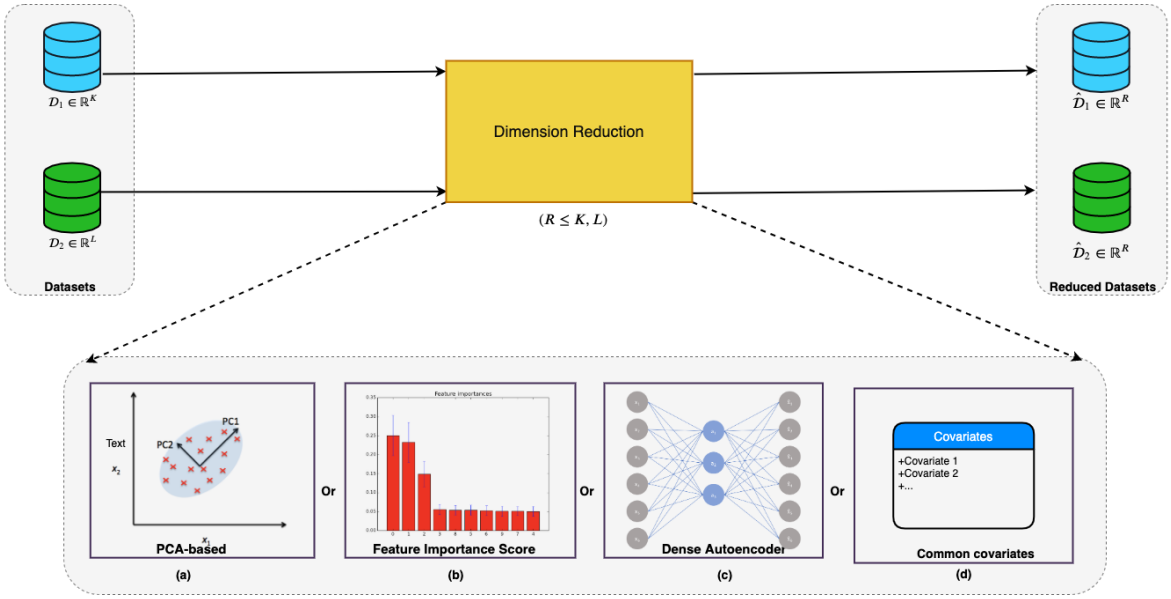


Figure 2: Dimension reduction is applied to achieve equal-dimensional feature representations of the two surveys using either of the following approaches: (a) PCA, (b) feature importance score, (c) autoencoder and (d) common covariates.

component analysis is applied to each of them independently, resulting $(P_1)^{K\times K}$ and $(P_2)^{L\times L}$, respectively. The top R Eigen vectors are then selected to project $\mathcal{D}_1$ and $\mathcal{D}_2$ into R-dimensional feature representation, $\hat{\mathcal{D}}_1 = (\mathcal{D}_1)^{N\times K} * (P_1)^{K\times R}$ and $\hat{\mathcal{D}}_2 = (\mathcal{D}_2)^{W\times L} * (P_2)^{L\times R}$, which results in equal-dimensional $\hat{\mathcal{D}}_1$ and $\hat{\mathcal{D}}_2 \in \mathbb{R}^R$.

Feature importance-based dimension reduction utilises the importance score of a feature (e.g., t-score) in the domain-specific models, M^1 and M^2. To this end, we group and sort the features in each dataset based on t-score directions and values. For $\mathcal{D}_1$, the t-score values of its features are arranged as $\mathbf{t}^1 = \{\mathbf{t}^1_+, \mathbf{t}^1_-\}$, where $\mathbf{t}^1_+ = (f^1_i)^{p_1}_{i=1}$ and $\mathbf{t}^1_- = (f^1_i)^{n_1}_{i=1}$ represent the the number of positive and negative directed features, respectively, $p_1 + n_1 = K$. Similarly, the features are grouped and sorted for $\mathcal{D}_2$, resulting $\mathbf{t}^2 = \{\mathbf{t}^2_+, \mathbf{t}^2_-\}$, $\mathbf{t}^2_+ = (f^2_i)^{p_2}_{i=1}$, $\mathbf{t}^2_- = (f^2_i)^{n_2}_{i=1}$, $p_2 + n_2 = L$. We then obtain the minimum number of positive and negative features across the two datasets, i.e., $p_{min} = \min(p_1, p_2)$ and $n_{min} = \min(n_1, n_2)$, and represent the samples $\mathcal{D}_1$ and $\mathcal{D}_1$, using the top p_{min} and n_{min} features sorted in descending order. By doing this, we reduce the feature dimension from K in $\mathcal{D}_1$ and from L in $\mathcal{D}_2$ to $p_{min} + n_{min} = R$ in $\hat{\mathcal{D}}_1$ and $\hat{\mathcal{D}}_2$, respectively.

Autoencoder-based dimension reduction utilises a dense encoder-decoder architecture to reconstruct each of $\mathcal{D}_1$ and in $\mathcal{D}_2$ separately. The output of the encoder part ($\mathcal{E}(\cdot)$) is treated as a reduced dimension and the dimension of the encoder outputs is set to be R-dimensional, resulting in $\hat{\mathcal{D}}_1 = \mathcal{E}_1(\mathcal{D}_1)$ and $\hat{\mathcal{D}}_2 = \mathcal{E}_2(\mathcal{D}_2)$.

Table 1: Examples of common covariates existing between DHS and AMANHI-2 (in KI) surveys collected in Ghana.

DHS - Ghana		**AMANHI-2 - Ghana**	
Code	Description	Code	Description
hv216	Number of rooms used for sleeping	NSLROOM	Number of sleeping rooms
hv009	Number of household members	NSLEEP	Number of persons sleeping in house
hv201	Source of drinking water	H2OSRCP	Source of drinking water
hv205	Type of toilet facility	SANITATN	Type of sanitary facility
hv226	Type of cooking fuel	COOKFUEL	Type of cooking fuel

Algorithm 1: Algorithm to obtain close neighbors in two datasets, aggregated their information and link to each other. © represents a concatenation operation.

Result: Cross-domain neighbors $\mathcal{C}$, Aggregated neighbors data A, cross-domain linked data U
Initialisation: $\hat{\mathcal{D}}_1$ and $\hat{\mathcal{D}}_2$, labels ($\mathbf{y}^1$ and $\mathbf{y}^2$), Linkage matrix (Θ), # of close neighbors (Q);
for *each* $\hat{S}_i^1$ *in* $\hat{\mathcal{D}}_1$ **do**
  $I_i^2 \leftarrow argmin(d_{i1}, d_{i2}, \cdots, d_{iM})$;
  if *linkage is supervised* **then**
    $\bar{I}_i \leftarrow (I_{ij}^2), \forall j \ni \mathbf{y}_j^2 = \mathbf{y}_i^1$;
  else
    $\bar{I}_i \leftarrow (I_{ij}^2), \forall j$;
  end
  $\mathcal{C}_i^1 \leftarrow \{S^2_{\bar{I}_i(q)_{q=1}^Q}\}$;
  $A_i^1 \leftarrow \frac{\sum(\mathcal{C}_i^1)}{Q}$;
  $\mathcal{D}_{12}^i = S_i^1 \copyright A_i^1$
end
for *each* $\hat{S}_j^2$ *in* $\hat{\mathcal{D}}_2$ **do**
  $I_j^1 \leftarrow argmin(d_{1j}, d_{2j}, \cdots, d_{Nj})$;
  if *linkage is supervised* **then**
    $\bar{I}_j \leftarrow (I_{ji}^1), \forall i \ni \mathbf{y}_i^1 = \mathbf{y}_j^2$;
  else
    $\bar{I}_j \leftarrow (I_{ji}^1), \forall i$;
  end
  $\mathcal{C}_j^2 \leftarrow \{S^1_{\bar{I}_j(q)_{q=1}^Q}\}$;
  $A_j^2 \leftarrow \frac{\sum(\mathcal{C}_j^2)}{Q}$;
  $\mathcal{D}_{21}^j = S_j^2 \copyright A_j^2$
end

Linkage Matrix Computation for Neighbors Selection

After $\hat{\mathcal{D}}_1$ and $\hat{\mathcal{D}}_2$ are obtained using either of the dimension reduction approaches discussed above, the distance d_{ij} between every pair of samples $\hat{S}_i^1 = (\hat{f}_{i1}^1, \hat{f}_{i2}^1, \cdots, \hat{f}_{iR}^1)$ in $\hat{\mathcal{D}}_1$ and $\hat{S}_j^2 = (\hat{f}_{j1}^2, \hat{f}_{j2}^2, \cdots, \hat{f}_{jR}^2)$ in $\hat{\mathcal{D}}_2$ is computed. For our implementation, we simply used the Euclidean distance metric: $d_{ij} = \sqrt{(\hat{S_{ir}^1} - \hat{S_{jr}^2})^2}$, where $r = 1, 2, \cdots, R$, $i = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, W$, resulting a linkage matrix $\Theta \in \mathbb{R}^{N \times W}$.

Cross-domain Neighbors Selection, Covariates Aggregation and Linkage

Using the linkage matrix, Θ, a set of Q close cross-survey neighbors ($\mathcal{C}^1$) in $\mathcal{D}_2$ is identified for each S_i^1 in $\mathcal{D}_1$ —i.e., the indices of top Q minimum d_{ij} values— are extracted from $\mathcal{D}_2$, and vice-versa (see Algorithm 1). We then aggregate the covariate information of the close Q neighbors to generate the aggregated neighbors datasets A_i^1 and A_j^2, respectively, for $\mathcal{D}_1$ and $\mathcal{D}_2$, using Algorithm 1. The aggregation could be a simple average for continuous covariates and median/mode for categorical covariates. To obtain the cross-survey linked dataset, $\mathcal{D}_{12}$ and $\mathcal{D}_{21}$, we simply concatenate their corresponding original feature spaces with the corresponding aggregated neighbors data A^1 or A^2 . Finally, we use the two linked datasets ($\mathcal{D}_{12}$ and $\mathcal{D}_{21}$) to train cross-domain models, $M_{12} : \mathcal{D}_{12} \rightarrow y^1$ and $M_{21} : \mathcal{D}_{21} \rightarrow y^2$ (referred as cross-domain prediction in Figure 1). We expect the prediction performance of cross-domain models, M_{12} and M_{21}, to potentially outperform domain-specific models, M_1 and M_2, as the former models acquire additional covariate information from the other dataset via the proposed linkage approach. Furthermore, the analysis of important features in the cross-domain predictions (e.g., M_{12}) can provide features from the two datasets,

$\mathcal{D}_1$ and $\mathcal{D}_2$, which provides a way to understand the impact of a feature in one dataset on the outcome prediction of the other dataset (e.g., f_i^1 on $\mathcal{D}_2$).

Experiments

In this section, we describe the datasets used for the validation of the proposed framework. In addition, we describe the experimental set-ups including classifiers used, train-test split ratio, number of principal components and network architecture of the autoencoders.

Datasets

We used multiple surveys in the global health domains, specifically related to maternal, newborn, and child health (MNCH), and their details are provided below. The DHS[2] contains representative data on population, health, HIV, and nutrition through more than 300 surveys in over 90 different countries. These nationally representative surveys are designed to collect data on monitoring and impact evaluation indicators important for individual countries and cross-country comparisons. In addition to neonatal/child mortality, a subset of DHS related to family planning is extracted to validate the proposed framework on another outcome, i.e., family planning discontinuation. The PMA[4] data include surveys related to household, service delivery point, and GPS of the area (not household level), collected using innovative mobile technology. In addition to the household, individual-level data were collected for each eligible female-identified in the household roster. The KI[3] database contain different studies some of which are controlled trials on the child growth effect of different interventions. We employed a subset of KI studies, particularly those focused on neonatal death, i.e., AMANHI-1 and AMANHI-2 surveys.

Experimental Setup

In our experimental validation, we employ appropriate pre-processing steps that clean up the raw survey and select plausible covariates to train the models. Pre-processing steps include discarding samples with large proportions of missing values and redundancy with other covariates and the outcome. The categorical variables in the survey data were also encoded into one-hot vectors during training and testings. Standard scaling of features is also applied. Only African countries are considered in our analysis.

A random linkage, resulting in $\mathcal{D}_{12R}$ and $\mathcal{D}_{21R}$, is also applied as a baseline to compare against our proposed similarity-based linkage. We set the number of principal components for dimension reduction in PCA and the size of the encoder output layer to $R = min(K, L)/0.5$, where K and L are the feature dimensions in $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. The number of common covariates between studies could vary and is not set a priori. The architecture of autoencoders is designed to contain three dense layers with dimensions 128, 64 and R, respectively, for both the encoder and decoder components. We trained the autoencoder with 100 epochs and a batch size of 256. ReLu activation is applied across the autoencoder layers except for the last encoder and decoder layers, where Sigmoid activation is used. Adam optimiser is employed during training with a loss of mean squared error between the input and the reconstructed output.

For classification, a random forest (RF) classifier was selected due to its simplicity and effectiveness across tabular survey data. Training and testing steps are repeated for 100 iterations for each experiment, and the average of the area under receiver operating characteristics (AUROC) is reported as a performance metric. In each iteration, five-fold stratified cross-validation is applied.

Results and Discussion

In this section, we discuss our experimental results conducted under different scenarios with different degrees of overlap between studies considered for linkage, different dimension reduction approaches and varying degrees of supervision during linkage (i.e., supervised vs. unsupervised).

Table 2: Performance of our proposed supervised linkage compared against separate and random linkages using DHS: Outcomes: child mortality (CM) and family planning discontinuation (FP). Burkina Faso (BF), Ghana (GH) and Nigeria (NG).

				Linked performance - AUROC (%)							
		Separate		FI		PCA		AE		Random	
Outcome	Country	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_{12}$	$\mathcal{D}_{21}$	D_{12}	D_{21}	D_{12}	D_{21}	$\mathcal{D}_{12R}$	$\mathcal{D}_{21R}$
FP	BF	54.1	55.8	97.2	97.0	97.4	96.0	98.2	97.7	53.6	54.9
	GH	55.7	55.9	97.2	97.2	95.9	95.8	98.7	98.6	54.1	54.3
	NG	57.4	58.0	98.5	98.5	96.0	96.7	99.1	98.7	56.6	57.0
	Average	**55.7**	**56.6**	**97.6**	**97.6**	**96.4**	**96.2**	**98.7**	**98.3**	**54.8**	**55.4**
CM	BF	60.4	59.8	97.6	97.7	95.0	94.7	99.6	99.7	57.7	57.5
	GH	56.4	56.4	94.9	95.2	91.5	91.9	99.2	99.2	54.8	53.9
	NG	60.6	60.1	96.4	96.5	95.4	94.0	99.1	99.2	58.3	57.9
	Average	**59.1**	**58.8**	**96.3**	**96.5**	**94.0**	**93.5**	**99.3**	**99.4**	**56.9**	**56.4**

Linkage of Completely Disjoint Datasets

Our validation of the proposed approach started with linking complete disjoints surveys, i.e., that do not have common covariates nor samples. To this end, we artificially created disjoint sub-datasets from a given survey by randomly grouping the samples and the covariates into either of the two groups. We used DHS data of three African countries: Burkina Faso (BF), Ghana (GH) and Nigeria (NG) collected in 2010, 2014 and 2013, respectively. Two different outcomes: child mortality (CM) and family planning discontinuation (FP) were validated with a supervised approach, i.e., close neighbors are selected from samples with a similar label.

Results in Table 2 show the performance of using $\mathcal{D}_1$ and $\mathcal{D}_2$, without linkage (separate) and it is comparison against linked performance using $\mathcal{D}_{12}$ & $\mathcal{D}_{21}$, across different dimension reduction techniques. It was shown that feature importance-based linkage (FI) of the two datasets improved the predictive performance, compared to the separate performance, across countries and validation outcomes (CM and FP). PCA-based linkage is shown to perform relatively below to our novel feature importance-based dimension reduction. This is partly due to the benefit of separately exploiting both positively and negatively directed features in the importance-based linkage. On the other hand, the AE-based data-level linkage is shown to perform the best among the dimension reduction techniques expectedly. Random linkage of samples ($\mathcal{D}_{12R}$ and $\mathcal{D}_{21R}$) between two datasets was applied as a baseline for the proposed linkage approaches; and they achieved much inferior results compared to our principled linkages. This validates the plausibility of our distance-based linkage approaches to improve predictive performance across different surveys.

Cross-study Linkage

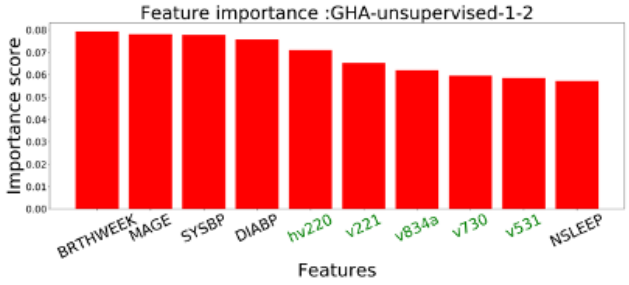
In this subsection, we discuss the results from the cross-study linkage of DHS and AMANHI-2 of KI for two African countries: Ghana and Kenya. These two studies expectedly have a few common covariates (see Table 1), and hence common covariates-based dimension reduction and unsupervised linkage is applied. The outcome of interest is specified to neonatal death, which happens in the most critical period of newborns. The results are shown in Table 3, and the linking of these two studies achieved improved performance compared to the majority of the cases. Particularly, benefited by AMANHI-2's discriminative covariates such as systolic and diastolic blood pressure (SYSBP and DIABP) (see Figure 3), AMANHI-2's linkage with DHS (DHS-AMANHI-2) improved independent DHS performance for both Ghana and Kenya. Figure 3 shows those covariates that are found to be predictive of neonatal death, both in their original study and when they are linked with other studies. Though more covariates are interpolated from DHS to AMANHI-2 studies in AMANHI-2-DHS linkage, their improvement is inferior to DHS-AMANHI-2, which partly reflects the lower discriminating potential of DHS covariates.

Cross-country and Cross-study linkage

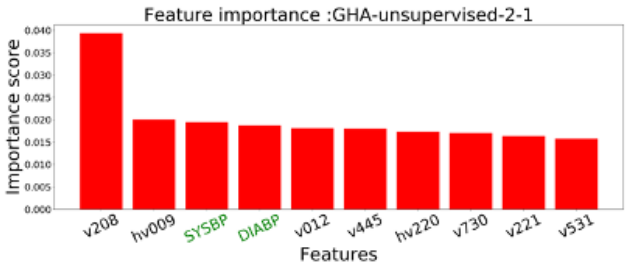
Finally, we also validated the linking of different surveys from different countries (cross-country cross-study) to evaluate the proposed framework in improving performance across these surveys and countries. To this end, we linked the PMA study of ET - Ethiopia (2017), which is characterised by a very small size (328 samples), with the DHS data

Table 3: Results of unsupervised cross-study linkage performed between KI and DHS data.

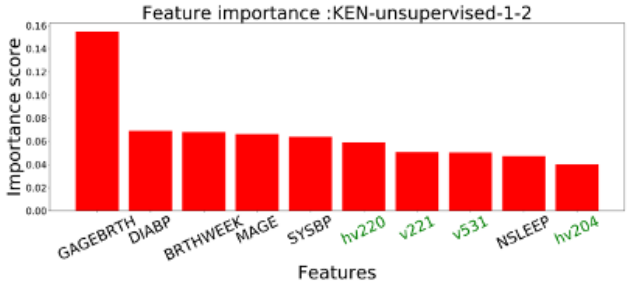
Countries	Ghana (GHA)		Kenya (KEN)	
Dataset	AMANHI-2	DHS	AMANHI-2	DHS
Population size (#)	24942	4294	14550	14949
Covariates (#)	34	341	42	355
% Death	4.73	6.24	2.80	5.47
Separate AUROC (%)	71.6	59.8	**78.9**	64.9
	AMANHI-2-DHS	DHS-AMANHI-2	AMANHI-2-DHS	DHS-AMANHI-2
Linked AUROC (%)	**73.2**	**70.3**	74.2	**77.0**



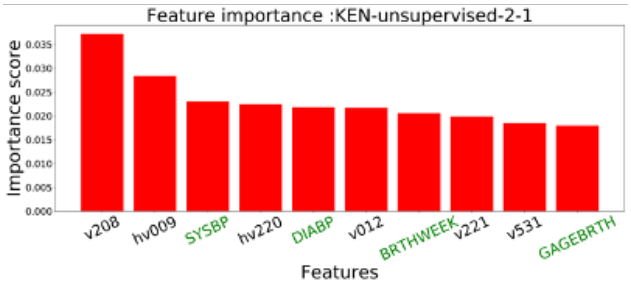
(a) DHS-AMANHI-2: Ghana



(b) AMANHI-2-DHS: Ghana



(c) DHS-AMANHI-2: Kenya



(d) AMANHI-2-DHS: Kenya

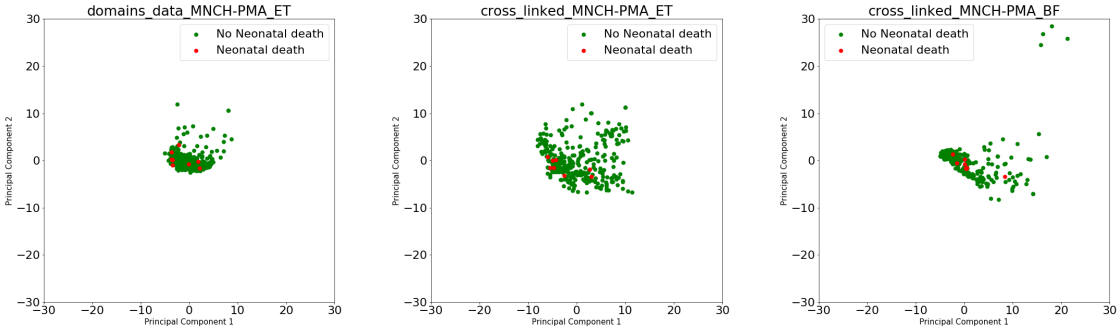
Figure 3: Important features extracted from the trained model of cross-study linkage between DHS and AMANHI-2 of KI studies for Ghana and Kenya. Covariates with green font color represent those interpolated from the other study via the proposed linkage approach.

of Ethiopia (2016), BF - Burkina Faso (2010), GH - Ghana (2014), KE - Kenya (2014), and NG - Nigeria (2013). Neonatal death was set as the outcome of interest.

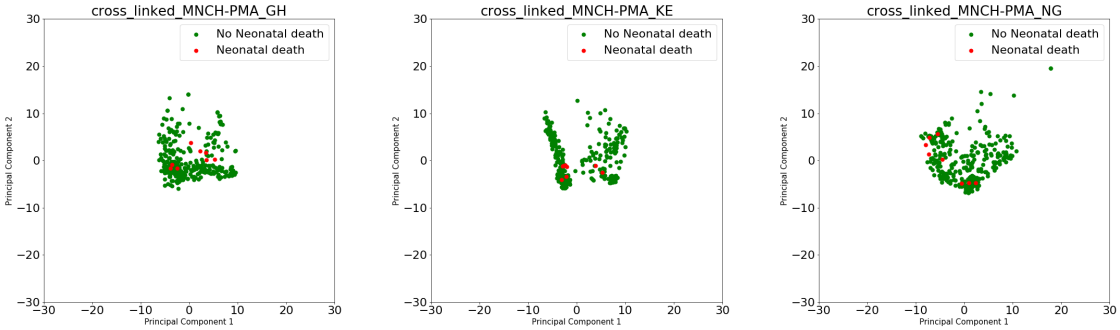
The results shown in Table 4 demonstrate that the proposed linkage helps to alleviate the prediction performance of neonatal death on PMA data of Ethiopia, when linked with DHS data of different countries, which were collected at different times and countries compared to the PMA. The PMA ET only achieved 47.3% AUROC expectedly due to its small size and high degree of imbalance. However, this performance is improved to 85.0% using the DHS of ET, and improved further to 89.1% by using the DHS data of Ghana. The 2-D PCA projections in Figure 4 demonstrate that the proposed linkage makes the original data sparse and hence eases prediction.

Table 4: Increase in the neonatal death prediction on PMA data of Ethiopia (from 47.3% AUROC) when linked with DHS data of other African countries.

	Linked with DHS of:				
Linking method	ET	BF	GH	KE	NG
Feature importance	61.2	62.2	59.7	55.8.	58.3
Principal component analysis	63.5	64.9	74.1	56.2	68.9
Autoencoder	**85.0**	**86.2**	**89.1**	**81.3**	**87.2**



(a) PMA (ET) **(b)** PMA (ET) - DHS (ET) **(c)** PMA (ET) - DHS (BF)



(d) PMA (ET) - DHS (GH) **(e)** PMA (ET) - DHS (KE) **(f)** PMA (ET) - DHS (NG)

Figure 4: PCA projections of (a) PMA data from Ethiopia (ET) before linkage is applied followed by the projections after PMA ET is linked with DHS of (b) Ethiopia and other African countries: (c) Burkina Faso, (d) Ghana, (e) Kenya and (f) Nigeria.

Conclusion

Neonatal and child death is still a critical global health challenge. Most developing countries are behind the 2030 target of reducing neonatal death. To this end, we proposed a machine learning technique to exploit multiple surveys to improve the understanding of such critical global health challenges. The proposed approach provides data-level integration of these surveys, which might have different study populations and covariate profiles. First, dimension reduction is employed to project these datasets to equal-dimensional representation; for which four different alternatives are employed: common covariates, feature importance score, principal component analysis, and autoencoders. Second, close neighbors of samples across datasets are identified via distance metric computation, lastly, the information of these closes neighbors is aggregated and linked to the original samples. To the best of our knowledge, this is the first study to aggregate multiple existing surveys in order to improve the predictive performance of neonatal and child mortality. Our approach has the potential of enabling domain experts and policy makers to evaluate the intelligibility of existing surveys and identify informative covariates to be included in future data collection efforts. Limitation of the current work includes computational complexity associated with the linkage matrix computation in large surveys, and future work aims to address this issue in addition to scaling up the proposed work to other data types and problem domains.

Acknowledgement

This work is done in collaboration with Bill & Melinda Gates Foundation.

References

1. World Health Organisation: Global Health Observatory (GHO) data. https://www.who.int/gho/child_health/mortality/neonatal_text/en/. Last accessed on December 18, 2020.;.
2. ICF International: Demographic and Health Surveys (DHS). Funded by USAID. Rockville, Maryland; 2004-2017. .
3. Data Store Explorer: Knowledge Integration (KI) - Africa. http://africa.studyexplorer.io/. Last accessed on December 18, 2020.;.
4. Performance Monitoring and Accountability 2020 (PMA2020) Project. Bill & Melinda Gates Institute for Population and Reproductive Health, Johns Hopkins Bloomberg School of Public Health.;. .
5. Wong SC, Gatt A, Stamatescu V, McDonnell MD. Understanding data augmentation for classification: when to warp? In: International Conference on Digital Image Computing: Techniques and Applications (DICTA); 2016. p. 1–6.
6. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:171204621. 2017;.
7. Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters. 2017;24(3):279–283.
8. Antoniou A, Storkey A, Edwards H. Data augmentation generative adversarial networks. arXiv preprint arXiv:171104340. 2017;.
9. Douzas G, Bacao F. Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Systems with Applications. 2018;91:464–471.
10. Yu L, Zhang W, Wang J, Yu Y. Seqgan: Sequence generative adversarial nets with policy gradient. In: Thirty-First AAAI Conference on Artificial Intelligence; 2017. .
11. Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006;28(4):594–611.
12. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering. 2009;22(10):1345–1359.
13. Vinyals O, Blundell C, Lillicrap T, Wierstra D, et al. Matching networks for one shot learning. In: Advances in Neural Information Processing Systems; 2016. p. 3630–3638.
14. Sung F, Yang Y, Zhang L, Xiang T, Torr PH, Hospedales TM. Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 1199–1208.
15. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data. 2018;5:180161.
16. Mensink T, Verbeek J, Perronnin F, Csurka G. Distance-based image classification: Generalizing to new classes at near-zero cost. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013;35(11):2624–2637.
17. Guo Y, Codella NC, Karlinsky L, Smith JR, Rosing T, Feris R. A New Benchmark for Evaluation of Cross-Domain Few-Shot Learning. arXiv preprint arXiv:191207200. 2019;.
18. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015;350(6266):1332–1338.
19. Doan A, Halevy A, Ives Z. Principles of data integration; 2012.
20. Jurczyk P, Lu JJ, Xiong L, Cragan JD, Correa A. Fine-grained record integration and linkage tool. Birth Defects Research Part A: Clinical and Molecular Teratology. 2008;82(11):822–829.
21. Winkler WE. Overview of record linkage and current research directions. In: Bureau of the Census; 2006. .
22. Ali MS, Ichihara MY, Lopes LC, Barbosa GC, Pita R, Carreiro RP, et al. Administrative data linkage in Brazil: potentials for health technology assessment. Frontiers in Pharmacology. 2019;10.
23. Boratto M, Alonso P, Pinto C, Melo P, Barreto M, Denaxas S. Exploring hybrid parallel systems for probabilistic record linkage. The Journal of Supercomputing;75(3):1137–1149.
24. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. International Journal of Epidemiology. 2016;45(3):954–964.

Benchmarking Modern Named Entity Recognition Techniques for Free-text Health Record Deidentification

Abdullah Ahmed, B.S.[1], Adeel Abbasi, M.D.[1], Carsten Eickhoff, Ph.D.[1]
[1]Brown University, Providence, RI, United States

Abstract

Electronic Health Records (EHRs) have become the primary form of medical data-keeping across the United States. Federal law restricts the sharing of any EHR data that contains protected health information (PHI). De-identification, the process of identifying and removing all PHI, is crucial for making EHR data publicly available for scientific research. This project explores several deep learning-based named entity recognition (NER) methods to determine which method(s) perform better on the de-identification task. We trained and tested our models on the i2b2 training dataset, and qualitatively assessed their performance using EHR data collected from a local hospital. We found that 1) Bi-LSTM-CRF represents the best-performing encoder/decoder combination, 2) character-embeddings tend to improve precision at the price of recall, and 3) transformers alone under-perform as context encoders. Future work focused on structuring medical text may improve the extraction of semantic and syntactic information for the purposes of EHR deidentification.

Introduction

A majority of medical practices across the United States have adopted Electronic Health Records (EHRs). Between 2008 and 2016, EHR use by office-based physicians has nearly doubled from 42% to 86%[1] – an increase largely attributable to the Federal Health Information Technology (IT) Strategic Plan of 2011[2,3]. One of the goals of this plan is to allow data within EHRs to be leveraged for scientific research. The use of EHR data continues to be restricted by the Health Insurance Portability and Accountability Act (HIPAA), whose Privacy Rule limits the distribution of patients' *protected health information* (PHI). Unrestricted research use of EHR data is only permissible once it is *de-identified* - all PHI has been removed. Per the HIPAA Privacy Rule, health information may be deemed de-identified through one of two methods: 1) "Expert Determination," a formal conclusion by a qualified expert that the risk of re-identification is very small, and 2) "Safe Harbor," the removal of 18 specified individual identifiers (names; geographic subdivisions; dates; telephone numbers; vehicle identifiers; fax numbers; device identifiers and serial numbers; emails; URLs; Social Security Numbers; medical record numbers; IP addresses; biometric identifiers; health plan beneficiary numbers; full-face images; account numbers; certificate or license numbers; any other identifier, code, or characteristic).

Manual de-identification is tedious and time-consuming[4]. Researchers in the Natural Language Processing (NLP) community have developed systems to automate "Safe Harbor" de-identification processes by scanning medical free text for PHI identifiers. End-to-end de-identification involves three steps: 1) locating PHI in free text, 2) classifying the PHI correctly, and 3) replacing the original PHI with realistic surrogates. Step (3) is beyond the scope of this study; for simplicity, we will use the term "de-identification" to refer only to steps (1) and (2). De-identification can be framed as a named entity recognition (NER) problem. Formally, given a sequence of input tokens $s = \{w_i\}_{i=1}^{n}$, an NER system outputs a list of tuples $< I_s, I_e, t >$, each of which is a named entity in s[5]. I_s represents the start token, I_e represents the end token, and t is the entity type. t is drawn from the 18 HIPAA PHI identifiers.

Automatic de-identification methods fall into four broad categories: rule-based, machine-learning, hybrid, and deep learning. Rule-based systems rely on pattern-matching of textual elements[6]. They are simple to implement, interpret, and modify, but they require laborious construction, lack generalizability to unseen data, and cannot handle slight variations in language or word forms (*e.g.*, misspellings, abbreviations). Machine learning systems model the de-identification task as a sequence labeling problem: given an input of tokens $w_1, w_2, \ldots, w_n$, the system outputs label predictions $y_1, y_2, \ldots, y_n$. Traditional machine-learning algorithms can recognize complex patterns in the data not evident to the human reader[7]. However, they require an input of handcrafted numerical "features" that are often time-consuming to engineer, and not guaranteed to be generalizable to other medical corpora. Hybrid methods combine elements of machine learning and rule-based systems[8]. Although they outperform their constituent parts, they still suffer from a lack of generalizability and a need for manual feature engineering.

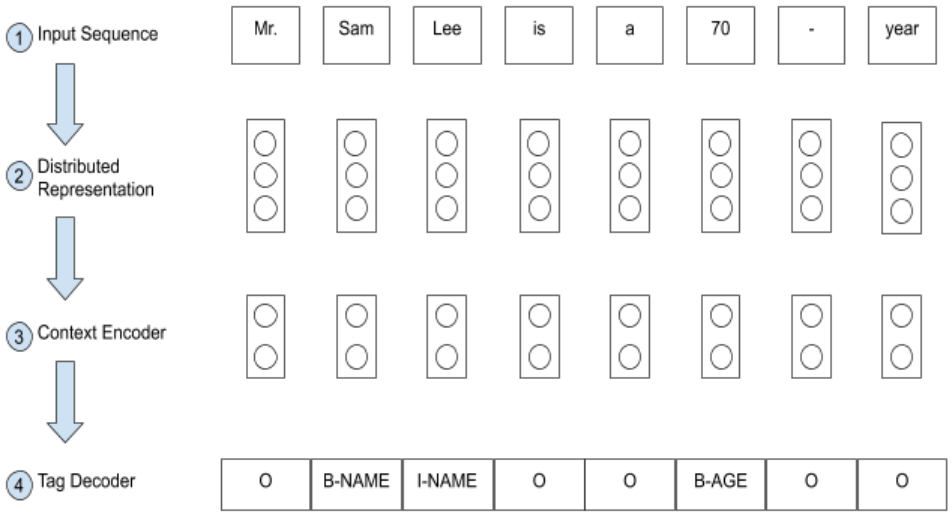


Figure 1: Pipeline taxonomy of deep-learning based NER systems.

Deep learning – a subset of machine learning based on artificial neural networks (ANNs) – circumvents these problems. ANNs are capable of representation learning *(i.e.*, automatically discovering useful features for a given task). In supervised learning, features are learned by training on a large set of labeled data of the form $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the vector representations of the inputs and labels, respectively. Deep learning-based models can learn complex representations of token sequences through a series of non-linear transformations. Li *et al.*[5] outline the general structure of deep learning methods for NER, displayed in Figure 1. Once the tokenized sentence is passed into the model, it undergoes three stages of processing. The distributed representation stage converts every token to a numeric vector. The context encoder then processes these vectors to capture the contextual dependencies across the entire sentence, outputting a new sequence of vectors (not necessarily in the same dimensionality as the embeddings). Finally, the tag decoder uses the output of the context encoder to predict the label for each token. All deep learning-based NER systems can be characterized by the concrete design decisions made for each of these stages of processing.

Recently, deep learning-based NER has been applied to de-identification[9–11]. Pre-trained word- and character-level embeddings have been employed in the first stage to form distributed representations of medical text. Recurrent Neural Networks (RNNs), specifically Long-Short-Term-Memory Networks (LSTMs), have demonstrated success in incorporating contextual information. Conditional Random Fields (CRFs) have gained popularity as a means of decoding the tags and predicting PHI labels in the final stage. In this study, we aimed to determine which NER design combinations perform better when tackling the de-identification task. Additionally, we aimed to extend the work of Yang *et al.*[11] and Yogarajan *et al.*[12] by evaluating the performance of our models on real EHR data collected from a local hospital.

Methods

Figure 2 summarizes the end-to-end structure of our system. The details are described in the following sections.

Data Collection

Upon request, the Blavatnik Institute of Biomedical Informatics at Harvard University granted us access to the de-identification corpus created by the Center for Informatics for Integrating Biology and Bedside (i2b2) in 2014. The dataset contains 1,304 free-text medical records of patients with diabetes for which all PHI was manually annotated and replaced with surrogate PHI. This corpus was used for model training and quantitative evaluation.

Additionally, 25 health record notes collected between March and May of 2020 from Rhode Island Hospital (RIH)

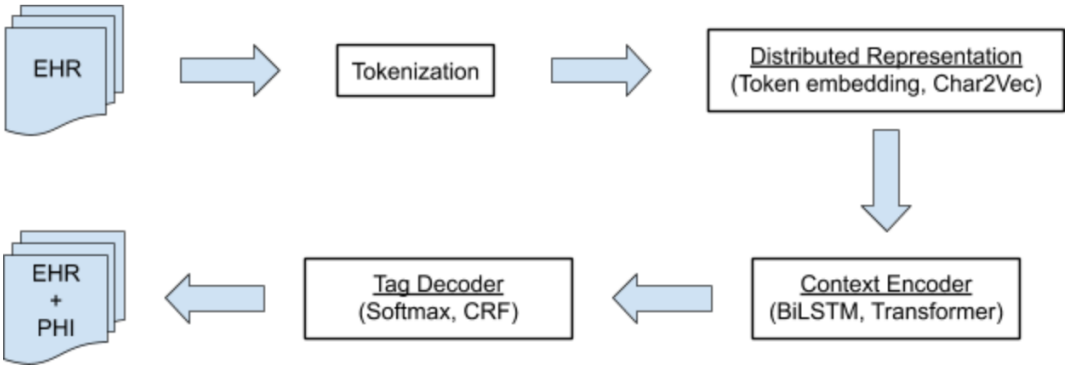


Figure 2: End-to-end pipeline of the NER design combinations we tested.

were included for qualitative manual performance inspection.

Pre-processing

Before feeding the medical text into our deep learning-based algorithms, we pre-processed it to take the form of sequences of sentences and tokens for each document. Formally, for every document $d \in D$, where D is the set of all medical documents, we split d into sentences s_i and tokens $t_{i,j}$.

Tokenization, the process of splitting sentences into tokens ($s_i \rightarrow t_{i,j}$), is a critical and highly customizable step for NLP systems. For many forms of free text, a tokenization scheme that splits based on whitespace and punctuation may suffice. However, the highly unstructured text in EHRs demands a more refined approach. We emulate the work of Liu *et al.*[10], which proposed a tokenization module that first splits on blank spaces, then recursively on other characters, words connected without a space, and numbers that appear adjacent to letters. For example, the EHR sentence "Mr. SamLee is a 70yo man" would be tokenized as ['Mr.', 'Sam', 'Lee', 'is', 'a', '70', 'yo', 'man']. This module preserves normally-occurring words and numbers whilst avoiding several pre-processing errors, such as "SamLee" and "70yo" in the example sentence above.

The set of all unique tokens in the training set is known as the "vocabulary." The training and testing sets have equivalent vocabularies because the models are not permitted to incorporate any testing words into their training vocabulary. Any out-of-vocabulary (OOV) tokens – words that appear in the testing set but not the training set – were replaced with the `UNK` token. This method allows models to generalize to tokens never encountered during training.

For each token $t_{i,j}$, we also stored the characters it spans, $c_{i,j,s}$ and $c_{i,j,e}$, so that our results coincide with the i2b2 label format. Each sentence s_i functions as a single training instance for our algorithms. Sentences were padded with `PAD` tokens so that every sentence had the same length m. `PAD` tokens are masked during training loss calculation so that the model focused on predicting actual tokens correctly.

The last pre-processing step generated a sequence of labels for every sentence s_i, such that every token $t_{i,j}$ has a corresponding label $l_{i,j}$. We employed the popular BIO scheme to create the label sequence. Let $L_{i,s:e}$ be a PHI in sentence i that starts at token number s and ends at e. The BIO scheme prepends `B-` (for beginning) to $l_{i,s}$ and `I-` (for inside) to $l_{i,s+1:e}$. For instance, if "Rhode Island Hospital" appeared in sentence 2 and spanned tokens 14-16, the corresponding labels $l_{2,14:16}$ would be `B-HOSPITAL, I-HOSPITAL, I-HOSPITAL`. Any tokens that do not qualify as PHI are assigned the label `O` (for "outside"). Figure 1 demonstrates BIO tagging for a sample sequence.

Distributed Representation

We leveraged information about the input across two levels (word and character) to form the embeddings for each token. At the word level, a token is viewed as a standalone unit. Every token in the training vocabulary was mapped to a unique vector in $\mathbb{R}^d$. The vectors were initialized to random values, and through training converged to useful represen-

tations. A random seed was set at the beginning of the program to ensure that the random vectors were initialized to the same values for every model we tested. At the character level, a token is viewed as a sequence of characters, allowing the model to incorporate sub-token patterns into the representation and thus capture additional semantic information from OOV tokens. Because it was trained on substantially more text than is available in the i2b2 dataset, we utilized a pre-trained character embedding layer, char2vec, to generate character-based embeddings. char2vec, trained using a Bidirectional Long-Short-Term-Memory (Bi-LSTM) to detect similar words based on character information[13], outputs vectors in $\mathbb{R}^{50}$, which we concatenated with the token-level vectors in $\mathbb{R}^d$ to form a new distributed representation of each token.

The use of pretrained word embeddings has led to dramatic successes in a wide range of NLP tasks. Pretrained word embeddings are embeddings learned through one task – generally one that requires no labeled data – and applied to solve a different task. While pretrained embeddings would likely have increased our models' performances, our study was focused on the foundational architectural components of the NER pipeline. We refer the reader to other work that focuses specifically on how pre-trained word embeddings improve performance on the de-identification task[14].

Context Encoder

Recurrent Neural Networks (RNNs) have demonstrated success in capturing contextual information from variable-length sequential data. Let $x_1, x_2, \ldots, x_m$ be a sequence of vectors at steps $t = 1, \ldots, m$. In our task, $x_1, \ldots, x_m$ correspond to the distributed representations of each token in a sentence. Unlike a normal feed-forward network, RNNs maintain a hidden state h_t that is fed as input into the model at time $t+1$ along with x_{t+1}. This way, the model is able to propagate prior information forward as the embeddings are sequentially processed.

RNNs lack the ability to capture long-term dependencies and suffer from the vanishing gradient problem. LSTMs attempt to alleviate these issues by incorporating a "cell state" c_t. c_t serves as a memory block that retains relevant information and discards irrelevant information collected up to time t. It does so through the use of forget and input gates; the forget gate controls what information from previous timesteps should be removed from the cell state, while the input gate controls what information from the current timestep should be added to the cell state. Furthermore, an output gate combines information from the current cell state c_t and the previous hidden state h_{t-1} to calculate a new hidden state h_t (recall that RNNs only utilize h_{t-1}). Both c_t and h_t are transmitted to the next timestep for use in processing input x_{t+1}.

Bi-LSTMs improve upon LSTMs by performing the same calculations in the reverse direction (with different parameters), thereby propagating contextual information in both directions. The final output of a Bi-LSTM is the concatenation of the hidden states from both the forward and backward passes. Bi-LSTMs are widely used in state-of-the-art NER systems, including those designed for de-identification[9–11]. Still, sustaining long-range dependencies is a challenge for sequential models such as Bi-LSTMs. In addition, because Bi-LSTMs perform sequential operations, they cannot be parallelized. These issues can be addressed by an alternative context encoder: a transformer. Transformer models have been adopted as context encoders for many NLP tasks, including NER[15], but are yet to be tested on the de-identification task.

Transformers gained immense popularity in the NLP community following recognition of the power of self-attention in sequence-to-sequence (seq2seq) modeling (*e.g.*, machine translation)[16]. Self-attention mechanisms simultaneously relate elements in a sequence to each other. Formally, attention is mapping of a query and a set of key-value pairs to an output, calculated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where Q, K, V are matrices of the query, key, and value vectors packed together, and d_k is the dimension of the key vectors. Q, K, V are calculated by multiplying the input sequence $x_1, x_2, \ldots, x_m$ by weight matrices W_Q, W_K, W_V that are learned through training.

Since transformers do not rely on sequential processing in their calculations, they have no inherent notion of token order. In order for the model to leverage positional information of the input tokens, positional encodings are added to the embedding of each element in the input sequence. The encoding function is designed such that the same token will

have slightly different embeddings depending on where it appears in the sentence, thereby "encoding" its position. We employed the same positional encoding function used by Vaswani *et al.*[16].

We employed Multi-Head Attention by computing attention multiple times (with different W_Q, W_K, W_V), concatenating the results, and multiplying by another weight matrix W_O. The output of Multi-Head Attention is passed through a feed-forward network to retrieve the final output sequence. The transformer model proposed by Vaswani *et al.*[16] includes an encoder and a decoder, each consisting of several Multi-Head Attention "blocks." Because the model was designed for tasks such as translation between languages, the decoder does not necessarily output sequences of length m as necessitated by NER (one label for each token). Therefore, only the encoder portion of the model was used to encode context, retaining the benefits of self-attention and parallelization.

Tag Decoder

The tag decoder takes the output of the context encoder as input and produces a final sequence of tags. Sequence labeling can be cast as a multi-class classification problem; that is, for every token, output a probability distribution over all possible PHI (after BIO conversion). This can be achieved using a time-distributed dense layer with softmax activation. The dense layer is applied to each token, and the softmax activation creates a probability distribution over all the PHI for that token. For a vector x, softmax is calculated as follows:

$$p(y = j|x) = \frac{e^{(w_j^T x + b_j)}}{\sum_{k \in K} e^{(w_k^T x + b_k)}}$$

where w, b denote the weights and biases of the dense layer, j is the index of one label, and K is the set of all labels. To find the most probable label, we take the $argmax$ of the above equation. Because softmax assumes the tags to be independent, the probability of an entire sequence of tags $y_1, \ldots, y_m$ is given by

$$p(y_1, \ldots, y_m|x) = p(y_1|x) \cdot \ldots \cdot p(y_m|x)$$

A shortcoming of using the softmax approach is that every token and label is decoded independently, rendering it unable to capture patterns in the sequence of tags (*e.g.*, `I-HOSPITAL` is likely to follow `B-HOSPITAL`). CRFs improve this by modeling dependencies between labels through graphical connections. In particular, linear-chain CRFs implement strictly sequential dependencies, as is the case in NER. Linear-chain CRFs define a global score for a sequence of tags as

$$C(y_1, \ldots, y_m|s_1, \ldots, s_m) = \\ b[y_1] + \sum_{t=1}^{m} s_t[y_t] + \sum_{t=1}^{m} T[y_t, y_{t+1}] + e[y_m]$$

where m is the length of the sequence, T is a transition matrix between all tags, and b, e are vectors that indicate the cost of beginning or ending on a given tag. The scores $s_1, \ldots, s_m$ are obtained by passing the output of the context encoder through a linear dense layer of size $|K|$. T contains parameters that encode how likely it is transition from one tag to the next, thereby capturing common sequences of tags that appear in the training data. Similar to softmax, CRFs model the posterior probability of a tag sequence using the following equation:

$$p(y_1, \ldots, y_m = j_1, \ldots, j_m|s_1, \ldots, s_m) = \\ \frac{e^{C(j_1, \ldots, j_m|s_1, \ldots, s_m)}}{\sum_{k_1, \ldots, k_m \in K^m} e^{C(k_1 \ldots, k_m|s_1, \ldots, s_m)}}$$

Linear-chain CRFs satisfy the optimal substructure property. Consequently, the calculations over possible sequences of tags can be completed efficiently via dynamic programming. The optimal sequence can be calculated using the Viterbi algorithm.

Training

Our networks were trained using cross-entropy loss, defined as

$$L = -\sum_{i} log(P(y_i))$$

where $y_i = y_{1i}, \ldots y_{mi}$ is the correct sequence of tags for sentence i. The probability P is given by the outputs of the softmax and CRF decoders.

Adam has been shown to yield the highest performance and fastest convergence on sequence labeling tasks[17]. Thus, we used the Adam optimizer with a learning rate of 0.001 to update the network weights in batches of size 32 for 10 epochs.

Experiments and evaluation

Table 1 lists the combinations of model components we tested. Model hyperparameters were selected according to the literature and constrained by GPU memory allocation. We trained each model independently on the official i2b2 training set and subsequently tested it on the official test set. All models were built using Tensorflow, a deep learning framework developed by Google.

Table 1: List of the tested models with their combination of representation, context encoder, and tag decoder.

Model Name	Distributed Repr.		Context Encoder		Tage Decoder	
	Token	Char2Vec	BiLSTM	Transformer	Softmax	CRF
BiLSTM	✓		✓		✓	
BiLSTM-CRF	✓		✓			✓
C2V-BiLSTM-CRF	✓	✓	✓			✓
Transformer	✓			✓	✓	
Transformer-CRF	✓			✓		✓
Transformer-BiLSTM	✓		✓	✓	✓	

To assess the performance of our models, we computed precision (PPV), recall (sensitivity), and F_1 of the PHI *entities*. We evaluated entities rather than tokens because unidentified tokens represent an infringement of the HIPAA Privacy Rule. For the same reason, we used the i2b2 "strict" measure that only takes a prediction to be correct if the entire entity is matched exactly.

Let TP stand for true positives, FP stand for false positives, and FN stand for false negatives. Precision calculates the proportion of correctly labeled PHI entities in the set of all PHI entities returned by the system (i.e. $\frac{TP}{TP+FP}$). Recall calculates the proportion of correctly labeled PHI entities in the set of all PHI entities in the test set (i.e. $\frac{TP}{TP+FN}$). F_1 is the harmonic mean of precision and recall (i.e. $2 \cdot \frac{precision \cdot recall}{precision+recall}$). The overall performance of each system was evaluated using "micro" and "macro" versions of these metrics. Micro-average calculates metrics at the corpus level, whereas macro calculates them at the document level and averages the result over all documents. Furthermore, we calculated precision, recall, and F_1 for each HIPAA-PHI type; these metrics are reported on the token level to offer a more detailed insight into performance variation across different PHI types. All of these calculations were executed using the official i2b2 evaluation script.

To evaluate the generalizability of our model, we qualitatively inspected the results of our best system on EHR data collected from RIH between March and May of 2020. We were unable to perform quantitative analyses of the RIH data because it was not accompanied by any true PHI labels.

Results

Table 2 shows descriptive statistics about the dataset after pre-processing. Table 3 displays the global results of our systems, evaluated at the macro-average level. Bi-LSTM-CRF is the best-performing system according to all three metrics (0.8391, 0.818, 0.8284), followed closely by Bi-LSTM (0.8154, 0.7949, 0.805).

Table 2: Summary statistics of the data after pre-processing

	Training	Testing
Sentences	31,535	21,670
Vocab Size	23,905	23,905
Tokens	627,208	421,839
PHIs	15,953	10,834
PHI Tokens	44,298	30,006

Table 3: Global performance (all PHI categories) on the test set. Metrics are reported on the macro-average level.

Model	Precision	Recall	F_1
BiLSTM	0.8154	0.7949	0.805
BiLSTM-CRF	**0.8391**	**0.818**	**0.8284**
C2V-BiLSTM-CRF	0.7925	0.3183	0.4542
Transformer	0.5027	0.6345	0.561
Transformer-CRF	0.6068	0.5843	0.5953
Transformer-BiLSTM	0.7259	0.6865	0.7056

Table 4 lists the performances of our models on HIPAA-PHI categories, evaluated at the micro-level. Bi-LSTM-CRF has the highest F_1 score in all categories. The addition of a transformer to the context encoder in Transformer-Bi-LSTM improved precision in both the AGE and CONTACT categories. Char2vec significantly improved the precision of LOCATION and slightly improved the precision of DATE, yet suffered tremendously in recall as a consequence. Bi-LSTM had higher recall than Bi-LSTM-CRF in three categories. DATE yielded the highest scores in all model combinations, owing to the constant, structured format in the i2b2 dataset (MM/DD/YYYY).

Table 4: Performance per category. Metrics are reported on the micro-average level.

PHI Category	BiLSTM			BiLSTM-CRF			C2V-BiLSTM-CRF		
	P	R	F_1	P	R	F_1	P	R	F_1
NAME	0.9012	0.7238	0.8028	**0.9268**	**0.7266**	**0.8146**	0.8992	0.2482	0.389
PROFESSION	0.7331	0.5389	0.6212	**0.8148**	0.5483	**0.6555**	0	0	0
LOCATION	0.7792	**0.6181**	0.6894	0.7975	0.6174	**0.696**	**0.8814**	0.2681	0.4112
AGE	0.8863	**0.9241**	0.9048	0.9407	0.8868	**0.913**	0.8843	0.2543	0.395
DATE	0.9798	0.9675	0.9736	0.9703	**0.9839**	**0.9771**	**0.9914**	0.228	0.3707
CONTACT	0.6416	**0.619**	0.6301	0.8226	0.5113	**0.6306**	0.7619	0.401	0.5255
ID	**0.8943**	0.7099	0.7915	0.8398	**0.7847**	**0.8113**	0.8631	0.4142	0.5598
	Transformer			Transformer-CRF			Transformer-BiLSTM		
	P	R	F_1	P	R	F_1	P	R	F_1
NAME	0.6922	0.7117	0.7018	0.7401	0.6074	0.6672	0.8171	0.5914	0.6862
PROFESSION	0.4403	**0.5514**	0.4896	0.6959	0.4704	0.5613	0.6307	0.4735	0.5409
LOCATION	0.6772	0.5137	0.5842	0.5987	0.4925	0.5404	0.6189	0.5571	0.5863
AGE	0.7779	0.8628	0.8182	0.8176	0.8176	0.8176	**0.9508**	0.7976	0.8675
DATE	0.8083	0.8564	0.8317	0.8563	0.824	0.8398	0.9603	0.9805	0.9703
CONTACT	0.3953	0.2932	0.3367	0.4854	0.1253	0.1992	**0.8525**	0.2607	0.3992
ID	0.6443	0.5967	0.6196	0.6751	0.4626	0.549	0.8887	0.4443	0.5925

Table 5 highlights one example in which Char2vec improved the ability to predict the label for OOV tokens. Table 6 shows the results of our BiLSTM-CRF model on ten samples of EHR data collected from RIH. The samples show that although our model generalizes to some pieces of PHI, it struggles with others that are unlike the ones present in the i2b2 dataset (*e.g.*, signature formats).

Table 5: Output of BiLSTM-CRF *vs.* C2V-BiLSTM-CRF on a sentence that contains several OOV terms. Character embeddings are able to identify "HESS" and "CLARENCE" as PATIENT tokens, whereas BiLSTM-CRF is not.

Original Tokens	HESS	,	CLARENCE	64365595
Test Tokens	UNK	,	UNK	UNK
BiLSTM-CRF	B-HOSPITAL	O	O	O
C2V-BiLSTM-CRF	B-PATIENT	O	I-PATIENT	O
True Labels	B-PATIENT	I-PATIENT	I-PATIENT	B-MEDICALRECORD

Discussion

In this study, we tested several different combinations of NER components – distributed representations, context encoders, and tag decoders – for EHR de-identification. We found that Bi-LSTM-CRF, introduced by Huang *et al.*[18] for general NER outside of the clinical domain, is the best overall encoder/decoder combination for de-identification. Our results are in agreement with Dernoncourt *et al.*[9], Liu *et al.*[10], and Yang *et al.*[11].

Despite Bi-LSTMS-CRF's overall superior performance, Table 4 shows that other configurations can locally outperform Bi-LSTM-CRF for some of the HIPAA-PHI categories. We attribute these findings to the distribution patterns of tokens in each category. LOCATION, for example, includes ZIP codes, sequences of five numbers. Char2vec is able to recognize that ZIP codes are consistently tokens with a length of five and composed only of numbers. Therefore, C2V-BiLSTM-CRF is the model most equipped to classify ZIP codes, contributing in part to its nearly 10% increase in LOCATION precision.

Furthermore, we found that character embeddings improved precision in several categories, yet decreased recall. This implies that morphological information captured by character embeddings increased the model's accuracy in identifying PHI type, yet decreased its sensitivity in detecting PHI. Disambiguation of PHI type is especially difficult for OOV tokens without the use of character embeddings, as evidenced in Table 5. The decline in recall is likely because the char2vec embeddings were not fine-tuned during the training process. Thus, the character embeddings remained static, unable to adapt to the distribution of medical text. Alternative character embeddings, such as those that utilize Convolution Neural Networks (CNNs)[19], could also improve performance.

In our study, transformers were less effective than Bi-LSTMs at encoding context. This may be accounted for by the uncontrolled sentence lengths in EHRs. Due to the transcriptional style of medical text, there are "sentences" that contain over a thousand tokens (m = 1567). As a result, the transformer model may try to capture long-term dependencies via self-attention in the absence of meaningful relationships. Moreover, in an analysis of encoder representations in transformers, Raganato *et al.*[20] show that syntactic information is captured in the first 3 layers of the encoder, while semantic information is captured later. The transformer we used, which only had two layers of multi-headed attention, may have only partially captured the syntactic information of a distribution of medical text that conformed to limited syntactic rules. Performance improved when a Bi-LSTM was stacked on top of the transformer, potentially having compensated for the lack of captured semantic information. Future research may explore adding more transformer layers to the context encoder to extract more semantic information.

While CRFs as tag decoders generally improve the F_1 score, our results show that they can decrease recall for both Bi-LSTM and transformer context encoders. We hypothesize that certain sequences of tags seen in the training set became favored by the model, leading to unseen sequences in the testing set receiving low likelihoods.

A hybrid method that leverages the strengths of each model – based on its performance in individual PHI categories – may function best in practice. For instance, Bi-LSTM-CRF could be used to output an initial set of candidate PHI's because it has the highest F_1 score in all categories. The candidates could then be filtered using models with high specificity scores, such as Transformer-BiLSTM for AGE and CONTACT predictions and C2V-BiLSTM-CRF for LOCATION and DATE predictions.

Qualitative assessment of our top model with the EHR data collected from RIH indicates that it somewhat generalizes beyond the i2b2 dataset (Table 6). It was still able to classify crucial PHI such as Medical Record Numbers (MRNs), account numbers, and dates. However, it failed with sentences and phrases whose formatting significantly differs from

Table 6: Sample output of BiLSTM-CRF for phrases in the RIH EHR dataset. Original tokens have been manually replaced. Green entities are PHI correctly identified (TP), red entities are PHI that went unidentified (FN), and orange entities were incorrectly identified as PHI (FP).

Mr. Smith is a 200-year-old gentleman
Admitted to Rhode Island Hospital for COVID-19
travel to YZ from 8/20 - 8/26 who presented to RIH ER on 8/28/60
Signature: Sam Lee, MD Electronic Signature
I communicated with this patient's father John Smith at 123-456-7890
Sunday will be the last day of therapy
Vent Mode: PC FiO2 (%) [50% - 100%]
until Sunday, as discussed with dr. Lee
Social work will continue to follow. LICSW 456-7890
MR #: 00000000000 Account #: 111111111

i2b2 (*e.g.*, signature formats, incomplete phone numbers), as well as with tokens it never encountered (e.g. "COVID-19"). One particularly revealing example is the classification of "Rhode Island Hospital" vs. "RIH ER." Our model could correctly classify the former because it extrapolated from similar hospital names it encountered during training. On the other hand, it was unable to extract any semantic information from the abbreviated form and thus misclassified it.

To alleviate the problem of model portability, Yang *et al.*[11] show that fine-tuning their model on labeled data from local hospital EHRs improves their performance. We were unable to do the same because the EHR data we received contained no PHI labels. Future research might explore the utilization of local EHR data to fine-tune a language model that is independent of the de-identification pipeline, drawing inspiration from models like ClinicalBERT that are fine-tuned on clinical text[21]. That said, recent research has shown that it is possible to extract personally identifiable information from large language models through adversarial attacks[22]. More work must be done to protect against these attacks before safely incorporating PHI into training data.

The underlying problem remains that medical text is highly unstructured and non-standardized, resulting in sentences that lack syntactic and semantic cohesiveness. Without structured information, it becomes near impossible to automatically achieve results that fully satisfy the HIPAA Privacy Rule and are portable to multiple hospital systems. At the lowest level, the text must be tokenized in a way that permits inference. Dedicated medical tokenizers like Medex exploit domain knowledge to extract information about medications from medical narratives[23]. However, this does not resolve the long and disorganized nature of medical text. Recent efforts to enforce structure upon notes using NLP may help in downstream tasks like de-identification that rely on extracting very specific information[24]. Uniformity in note structure will not only improve the model's performance but will also increase its ability to generalize beyond the data used in training.

Conclusions

This study gives a comprehensive review of wide-ranging information extraction techniques on the de-identification of EHRs. Through empirical testing of different NER design combinations, we found that Bi-LSTM-CRF is the best-performing encoder/decoder combination for the de-identification task. Character-embeddings tend to improve precision at the cost of recall, while the opposite is true for CRFs. Meanwhile, transformers alone underperformed as context encoders. Qualitative assessment of Bi-LSTM-CRF on local EHR data showed some success, yet the issue of model portability remains. Future work lies in automatically structuring medical text such that semantic and syntactic information can more easily be extracted and models become more generalizable.

References

1. Myrick K, Ogburn D, Ward B. Percentage of office-based physicians using any electronic health record (EHR)/electronic medical record (EMR) system and physicians that have a certified EHR/EMR system, by U.S. state. National Center for Health Statistics; 2019.

2. Adler-Milstein J, Jha AK. HITECH act drove large gains in hospital electronic health record adoption. Health Affairs. 2017;36(8):1416–1422.
3. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015. ONC Data Brief. 2016;35.
4. Dorr D, Phillips W, Phansalkar S, Sims S, Hurdle J. Assessing the difficulty and time cost of De-identification in clinical narratives. Methods Inf Med. 2018;45:246–52.
5. Li J, Sun A, Han J, Li C. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering. 2020:1–1.
6. Friedlin J, McDonald CJ. A software tool for removing patient identifying information from clinical documents. Journal of the American Medical Informatics Association. 2008;15(5):601–610.
7. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC medical research methodology. 2010;10(1):1.
8. Yang H, Garibaldi JM. Automatic detection of protected health information from clinic narratives. Journal of Biomedical Informatics. 2015;58:30–8.
9. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2016;24(3):596–606.
10. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. Journal of biomedical informatics. 2017;75S:34–42.
11. Yang X, Lyu T, Li Q, et al. A study of deep learning methods for deidentification of clinical notes in crossinstitute settings. Yang et al BMC Medical Informatics and Decision Making. 2019;19(5):32.
12. Yogarajan V, Pfahringer MMB. A survey of automatic de-identification of longitudinal clinical narratives. CoRR. 2018:1810.
13. Hussain A, Moosavinasab S, Sezgin E, Huang Y, Lin S. Char2Vec: Learning the Semantic Embedding of Rare and Unseen Words in the Biomedical Literature; 2018.
14. Johnson AEW, Bulgarelli L, Pollard TJ. Deidentification of Free-Text Medical Records Using Pre-Trained Bidirectional Transformers. New York, NY, USA: Association for Computing Machinery; 2020. .
15. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:181004805. 2018.
16. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. Advances in Neural Information Processing Systems 30. 2017:5998–6008.
17. Reimers N, Gurevych I. Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. CoRR. 2017.
18. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. CoRR. 2015.
19. Chen H, Lin Z, Ding G, Lou J, Zhang Y, Karlsson B. GRN: Gated relation network to enhance convolutional neural network for named entity recognition. AAAI. 2019.
20. Raganato A, Tiedemann J. An analysis of encoder representations in transformer-based machine translation. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; 2018. p. 287–297.
21. Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72–78.
22. Carlini N, Tramer F, Wallace E. Extracting Training Data from Large Language Models. arXiv:201207805. 2020.
23. Xu H, Stenner S, Doan S, et al. MedEx: A medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association : JAMIA. 2010 01;17:19–24.
24. Steinkamp JM, Chambers C, Lalevic D, Zafar HM, Cook TS. Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. Journal of Digital Imaging. 2019;32:554–64.

Heterogeneity in COVID-19 Patients at Multiple Levels of Granularity: From Biclusters to Clinical Interventions

Suresh K. Bhavnani PhD[1,2], Erich Kummerfeld PhD[8], Weibin Zhang PhD[1], Yong-Fang Kuo PhD[1], Nisha Garg PhD[3], Shyam Visweswaran MD PhD[10], Mukaila Raji MD MS FACP[4], Ravi Radhakrishnan MD MBA[5], Georgiy Golvoko PhD[6], Sandra Hatch MD FAACS FACCS FACR[7], Michael Usher MD PhD[8], Genevieve Melton-Meaux MD PhD[8,9], Christopher Tignanelli MD MS[9]

[1]Preventive Medicine and Population Health, [2]Inst. for Translational Sciences, [3]Depts. of Microbiology & Immunology and Pathology, [4]Div. of Geriatrics, Internal Medicine, [5]Depts. of Surgery & Pediatrics, [6]Depts. of Pharm. & Toxicology, [7]Cancer Center, Univ. of Texas Medical Branch, Galveston TX; [8]Inst. for Health Informatics, [9]Dept. of Surgery, Univ. of Minnesota, Minneapolis, MN; [10]Dept. of BMI, Univ. of Pittsburgh, Pittsburgh, PA.

Abstract

Several studies have shown that COVID-19 patients with prior comorbidities have a higher risk for adverse outcomes, resulting in a disproportionate impact on older adults and minorities that fit that profile. However, although there is considerable heterogeneity in the comorbidity profiles of these populations, not much is known about how prior comorbidities co-occur to form COVID-19 patient subgroups, and their implications for targeted care. Here we used bipartite networks to quantitatively and visually analyze heterogeneity in the comorbidity profiles of COVID-19 inpatients, based on electronic health records from 12 hospitals and 60 clinics in the greater Minneapolis region. This approach enabled the analysis and interpretation of heterogeneity at three levels of granularity (cohort, subgroup, and patient), each of which enabled clinicians to rapidly translate the results into the design of clinical interventions. We discuss future extensions of the multigranular heterogeneity framework, and conclude by exploring how the framework could be used to analyze other biomedical phenomena including symptom clusters and molecular phenotypes, with the goal of accelerating translation to targeted clinical care.

Introduction

Despite extreme measures to contain the *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), the resulting corona virus disease 2019 (COVID-19) continues to have a devastating impact on the physical, social, cultural, and economic health of humans around the world. Although this novel corona virus is serologically close to the known SARS-CoV, it has spread more widely primarily due to virus shedding from asymptomatic patients. As of December 23rd 2020 more than 78.6 million people were infected worldwide, with more than 1.73 million dead due to fatal complications. Because many countries have yet to peak in their cases and fatalities and prepare for subsequent waves and potential reinfections, there is an urgent need to analyze and treat the causes for fatal complications in COVID-19 patients.

A key trait of COVID-19 is the high fatality rate in older adults and minorities[1-3] resulting from the following molecular and socio-demographic factors: (1) **Molecular Mechanisms Precipitating Fatal Complications.** SARS-CoV-2 has the characteristic spiked protein 3-D structure, with a strong binding affinity to the human cell receptor *angiotensin-converting enzyme 2* (ACE2).[4] Because ACE2 is expressed in many organs including the heart, lungs, and kidneys, in addition to the nervous system and skeletal muscle, SARS-CoV-2 can infect multiple sites by traversing the hematogenous or the retrograde neuronal routes.[5,6] Furthermore, laboratory tests of critical patients have shown abnormal levels for key markers including *cardiac troponin* (myocardial infarction), *D-dimer* (compromised blood clotting), *lymphocytes* (lymphopenia), *lactate dehydrogenase* (multiple organ failure), and *liver enzymes* (damage to liver cells).[7] These results suggest that SARS-CoV-2 can exacerbate already compromised organs in older patients with multiple chronic conditions, resulting in a high rate of complications, multiple organ failures, and fatalities;[4,8] Furthermore, the higher expression of ACE2 in males[9] is a critical factor in putting them at a higher risk for severe complications with COVID-19; (2) **Prevalence of Multiple Chronic Conditions in Older Adults and Minorities.** Due to a wide range of factors including improved treatments and increased life-expectancy, a growing number of older adults live with and manage multiple chronic conditions (MCCs) defined as $\geq$ 2 concomitant chronic conditions (also referred to as multimorbidities, or comorbidities when used in the context of an index condition).[10] This trend has resulted in almost 75% of Americans aged 65 years and older having more than one chronic condition, 20% having five or more comorbidities, and 50% receiving five or more medications.[11] Furthermore, due to systemic health inequities, MCCs is also growing in subgroups including women, African Americans, and non-Hispanic Whites.

The above two factors have resulted in a disproportionate number of infected older adults and minorities having adverse outcomes, including respiratory failure requiring ventilators, and multiple organ failure needing management in intensive care units (ICUs), and of mortality.[1-3] However, despite the high prevalence of MCCs in these populations, the emerging clinical practice guidelines to treat COVID-19 patients have focused on treating single conditions. For example, recommendations to treat COVID-19 patients with diabetes have little to no cross-referencing to other

comorbidities if they co-occur in the same patient.[12] This single-condition focus is not unique to COVID-19; few existing clinical practice guidelines (CPGs) to treat conditions such as congestive heart failure are designed to treat multiple co-occurring conditions.[13-17] Such condition-specific guidelines can substantially increase the burden of treatment on older adults to manage complex treatment regimes, and require constant monitoring by primary care physicians to change a treatment plan if it leads to adverse or null outcomes.[15-17]

We therefore attempted to address the above gap in understanding of how comorbidities co-occur in COVID-19 patients with the goal of designing guidelines for treating patients with multiple chronic conditions. We begin with describing the current methods used to analyze heterogeneity in the comorbidity profiles of patients, and the advantages of using bipartite networks to automatically identify patient subgroups and their most frequently co-occurring comorbidities at different levels of granularity. Next, we discuss how we used that approach to analyze COVID-19 patients at three levels of granularity, each providing direct clinical implications related to frequency, risk, and similarity of comorbidity profiles. We conclude with how the multigranular heterogeneity approach could be used to analyze other biomedical phenomena, and how it could be extended to include other analytical methods, with the goal of accelerating translation of results into clinical care.

Current Methods Used to Analyze Co-Occurrence of Multimorbidities

Because having MCCs is associated with several adverse outcomes including poor quality of life, physical disabilities, high healthcare use, drug-drug and drug-disease interactions, and mortality, several studies have attempted to analyze MCCs in older adults and minorities, with the goal of optimizing care.[13,14] These studies have used a wide range of methods to analyze multimorbidities, each with critical trade-offs. For example, many studies have attempted to identify frequently co-occurring multimorbidities using combinatorial approaches[18] (identify all pairs, all triples etc.). However, while such approaches are intuitive to understand, they lead to a combinatorial explosion (e.g., finding all combinations of the 31 Elixhauser comorbidities would lead to 2^{31} or 2147483648 combinations), but with no simple way of addressing the overlap of patients between the combinations to identify patient subgroups.

Several studies have used unipartite clustering methods[19,20] (clustering patients or comorbidities, but not both simultaneously) such as k-means, and hierarchical clustering to help identify either clusters of frequently co-occurring multimorbidities, or patients that have a high similarity in their multimorbidity profiles. Other studies have used dimensionality-reduction methods such as principal component analysis (PCA)[19] combined with k-means to identify clusters of either MCCs or patients. However, because such methods produce unipartite outputs, there is no agreed upon method to identify the patient subgroups defined by a cluster of MCCs because patients can belong to more than one MCC cluster, and vice-versa. Furthermore, such methods have well-known limitations including the requirement of user-selected parameters such as similarity measures and the number of expected clusters, in addition to the absence of a quantitative measure to describe the quality of the clustering, critical for measuring its statistical significance.

Researchers have used the above methods to analyze multimorbidities at different levels of granularities in the data. Several studies have focused on analyzing multimorbidities at the cohort-level to identify co-occurring multimorbidities in an entire dataset.[21] Other studies have focused on analyzing pre-defined patient subgroups within specific diseases such as COPD to determine their risk for adverse outcomes.[22] Finally, a few studies have analyzed multimorbidities at an individual patient-level to determine through a case study approach, their burden of treatment resulting from clinical practice guidelines focused on treating single conditions.[13]

More recently, bipartite approaches have attempted to address the limitations of unipartite methods by identifying *biclusters*[20,23] of patients and comorbidities simultaneously. For example, as shown in Fig. 1A, bipartite networks[24] can be used to represent patients as well as their comorbidities as nodes (circles and triangles respectively), and the pair-wise associations between patients and comorbidities can be represented as edges (lines). Furthermore, algorithms such as modularity maximization[24] can be used to automatically (1) identify the number and members of biclusters consisting of patient subgroups and their most frequently co-occurring comorbidities, (2) measure the quality of the biclustering through a quantity called modularity, used to measure its significance compared to random permutations of the data;[20,25] and (3) visualize the results using force-directed algorithms such as Kamada Kawai[26] and ExplodeLayout.[27]

Need to Analyze Heterogeneity at Multiple Levels of Granularity

Although bipartite networks have been effective in automatically identifying statistically significant and clinically meaningful patient subgroups, recent results have revealed that heterogeneity in patient profiles could exist at other levels of granularity. For example, while a bipartite network analysis of comorbidities in hip-fracture patients helped to reveal heterogeneities at the cohort-level, there were additional heterogeneities within patient subgroups such as a significantly different proportion of patients that had one or more comorbidities across the subgroups.[28] Such heterogeneities could impact how patients within each subgroup receive treatment. These results suggest that bipartite

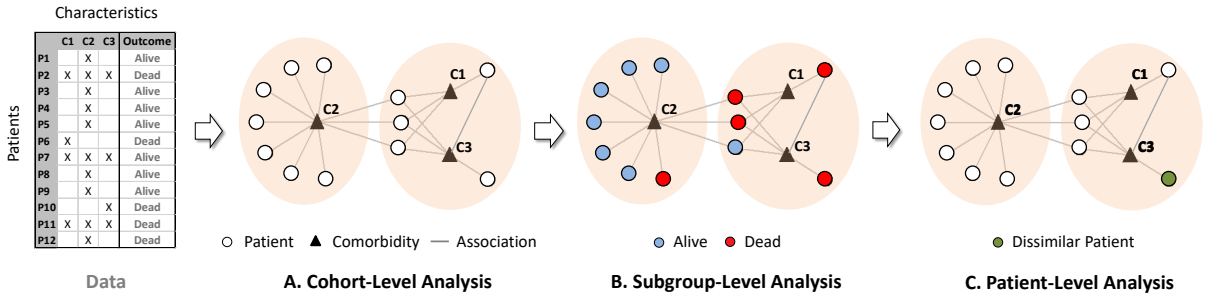


Fig. 1. Heterogeneity analysis at three levels of granularity: (A) cohort-level analysis enables the identification of frequently co-occurring comorbidities; (B) subgroup-level analysis enables the measurement of comparative risk of patient subgroups and how they differ in their characteristics; and (C) patient-level analysis helps to determine whether a specific patient is similar or dissimilar to other patients in their respective subgroup.

networks could be used to analyze heterogeneity at *multiple levels of granularity* in the data, each level providing different insights for designing clinical interventions.

For example, at the **cohort-level** (Fig. 1A) a bipartite network analysis can automatically identify biclusters consisting of patient subgroups defined by *frequently* co-occurring comorbidities. This level of analysis enables clinicians to determine which combinations of comorbidities need to be addressed in clinical practice guidelines, and identify potential underlying disease mechanisms. Furthermore, the bipartite network can be used to analyze heterogeneity at the **subgroup-level** by ranking each subgroup based on their *risk* for an outcome such as mortality (shown as red and blue nodes in Fig. 1B) enabling clinicians to design patient triage strategies in resource-constrained situations such as in COVID-19 hotspots. Finally, analysis at the **patient-level** (Fig. 1C) could help identify the degree of *similarity* of a specific patient to the rest in a subgroup, enabling clinicians to determine whether that patient should be treated using a generalized intervention designed for the entire subgroup, or requires individualized treatment.

The bipartite representation therefore provides a unified approach to quantitatively analyze and visually interpret heterogeneity at all three levels of granularity. This granularity approach enables the inspection of whether and how each level of granularity contributes to the translation of the analytical results into clinic interventions. Given the many unknowns in how multiple comorbidities impact outcomes in COVID-19 patients, and the urgent need for clinical interventions, we used the above multigranular heterogeneity analytical framework to guide our analysis of COVID-19 patients, with the explicit goal of enabling a more systematic approach for designing clinical interventions targeted to patients with multiple comorbidities.

Method

Research Questions. To analyze heterogeneity in the comorbidity profiles of COVID-19 inpatients at multiple levels of granularity, we posed the following three research questions: (1) Cohort-Level: *How do comorbidities frequently co-occur to form subgroups of COVID-19 inpatients?* (2) Subgroup-Level: *What is the risk of inpatient subgroups for adverse outcomes (ICU No-Vent, ICU With-Vent, and mortality)?* (3) Patient-Level: *What is the degree of similarity in the profile of a specific inpatient compared to the rest of the inpatients in the respective subgroup?*

Data. Using IRB (#STUDY00009771) from the University of Minnesota, we analyzed electronic health records (EHR) data from the University of Minnesota M Health Fairview COVID-19 patient registry. This registry includes patient data (spanning 135 unique zip codes) from 12 hospitals and 60 clinics in the Minneapolis-St. Paul twin-city area, and currently accounts for over 20% of inpatients in Minnesota. On August 2nd 2020, the registry contained health records of COVID-19 inpatients (n=858) with complete data for: (1) **31 comorbidities** (Elixhauser comorbidities in 2019-2020). A subset of the COVID-19 inpatients had no comorbidities (n=69), which were considered as the control group; (2) **7 complications** (acute respiratory distress (ARDS), acute kidney injury (AKI), hypotension, bleeding, delirium, and VTE/CVA/MI). Of these, 5 had <2% prevalence in the cases and controls and therefore were dropped from the analysis; (3) **55 laboratory test results** (e.g., platelet count, creatinine, hemoglobin, D-dimer, troponin, IL6) that were dichotomized into normal vs. abnormal. Of these, 13 laboratory tests had <2% prevalence in the cases and controls and were therefore dropped from the analysis; and (4) **markers of adverse outcomes** including use of an intensive care unit without a ventilator (ICU No-Vent, n=273), use of an intensive care unit with a ventilator (ICU With-Vent, n=672), and mortality (n=103). *As all our data relates to COVID-19 positive inpatients, henceforth we refer to them simply as patients.*

Analysis. We used bipartite networks to quantitatively and visually analyze the above COVID-19 patient data with the goal of enabling domain experts design targeted interventions, by using the following steps:

1. **Feature Selection.** We used the chi-squared test to measure the univariable significance (corrected for multiple testing using false discovery rate) of each comorbidity to each of three outcomes (ICU No-Vent, ICU With-Vent, and mortality), compared to the control group (patients with no comorbidities), and selected those comorbidities that were significant at the .05 level for at least one of the three adverse outcomes (ICU No-Vent, ICU With-Vent, and mortality).

2. **Bipartite Network Analysis.** Similar to Fig. 1A, we represented patients and comorbidities as nodes (circles and triangles respectively), and the pair-wise association between them as edges (lines). Patients (n=789) with at least one of the significant comorbidity were used in the bipartite network analysis, and patients with no comorbidities (n=69) were used as controls. The resulting network was analyzed at the following three levels of granularity:

    A. **Cohort-Level.** The quantitative analysis consisted of the following: (1) used a bicluster modularity maximization[24] algorithm to identify the number and boundaries of patient-comorbidity biclusters and the degree of biclustering (Q); and (2) measured the significance of Q by comparing it to a distribution of Q generated from 1000 random permutations of the network by preserving the size (number of nodes and edges in the network), and the distribution of edges for each comorbidity. The biclustering was tested for stability by measuring the Adjusted Rand Index (ARI)[19] between the comorbidity clustering in the real data, to 1000 random bootstrap resamples of patients in the data. The visual analysis of the above results consisted of the following: (1) used Kamada-Kawai[24] to layout the network; and (2) applied ExplodeLayout[27] to separate the identified biclusters for improving their interpretability.

    B. **Subgroup-Level.** The quantitative analysis consisted of the following steps applied to each patient subgroup: (1) used logistic regression to measure the odds ratio (OR) and tested its significance (corrected for multiple testing using FDR) for each of the three outcomes, in comparison to the control group; and (2) used logistic regression to measure the OR and tested its significance (corrected for multiple testing using FDR) for each demographic, complication, and laboratory test variable, compared to the control group. The visual analysis consisted of coloring the patient nodes based on mortality (as shown in Fig. 1B), and displaying in text the outcomes for each bicluster. Additionally, to increase interpretability of the results by the domain experts, a profile of each bicluster was compiled in a table (Table 4) showing the respective comorbidities, outcomes, demographics, complications, and laboratory test results.

    C. **Patient-Level.** The quantitative analysis consisted of the following: (1) defined an *augmented bicluster* as all patients within a bicluster (identified through modularity maximization in Step-2A), and all comorbidities within and outside that bicluster that were connected to at least one patient within the bicluster; (2) used the Jaccard Distance (JD),[19] defined as: 1 – (number of shared comorbidities between a patient pair / union of comorbidities in that pair), to measure the similarity in comorbidity profile of each patient to the other patients within each of the augmented biclusters, and calculated the mean JD for each patient; and (3) generated a distribution of the mean JDs for each augmented bicluster. This distribution was used to test whether the mean JD of a specific patient was significantly higher than the mean of the distribution in its augmented bicluster. A patient which had a significantly higher mean JD compared to the mean of the above distribution was considered dissimilar to other patients in the respective bicluster. Outlier patients with the minimum and the maximum number of comorbidities were selected for interpretation by the domain experts to examine whether they would require treatment that was different from their respective bicluster.

3. **Clinical Interpretation.** The results of the above quantitative and visual analysis at each level of granularity were presented to two domain experts with experience in treating COVID-19 inpatients. To guard against confirmation bias,[29] we used the following steps with each domain expert: (1) presented the network visualization generated from Step-2A (which did not contain the associations of each bicluster to any of the variables); (2) asked them to use their clinical judgement to rank the biclusters based on the risk for mortality; (3) revealed the quantitative profile of each bicluster, and asked them to discuss discrepancies between their prediction and the results, with the goal of determining which of the bicluster profiles were clinically meaningful. To translate the results into clinical interventions, we asked the two domain experts to (1) independently use their clinical judgement to design interventions to treat COVID-19 patients at each level of granularity, and (2) together arrive at a consensus.

Results

The analysis revealed statistically and clinically significant heterogeneity at all three levels of granularity. Below we present the quantitative, visual, and clinical interpretive results from each of the three levels.

1. Cohort-Level

Quantitative and Visual Results. The feature selection method identified 18 comorbidities that were significant after FDR correction for at least one of the three outcomes (13 were significant for all 3 outcomes, 1 was significant for 2

outcomes, and 4 were significant for one of the outcomes). All subsequent analyses were conducted on COVID-19 patients with these 18 comorbidities (n=789), compared to COVID-19 patients with none of these comorbidities (n=69).

The modularity maximization algorithm identified 4 biclusters consisting of patient subgroups and their most frequently co-occurring comorbidities. The biclustering was statistically significant compared to 1000 random permutations of the data (COVID-19 Q=0.22, Random Median Q=0.19, P<.001), with clusters that were stable based on 1000 random bootstrap selections of the data (Median ARI=0.76, P<.001). As shown in Fig. 2, the visualization revealed biclusters consisting of different numbers of patients (ranging from 237-158) and different numbers of comorbidities (6-3). The comorbidities within each bicluster shown in Fig. 2 are ranked by their univariable significance. Not included in the above network analysis was the control group with no comorbidities, but shown in the lower right hand-side of the figure as a dotted oval.

Qualitative Results. The following was the consensus ranking of the four biclusters, with explanations and recommended treatments for each:

i. Bicluster-4 *(hypertension complicated, renal failure, valvular disease, peripheral vascular disorders, pulmonary circulation disorders, and congestive heart failure).* The domain experts stated that this bicluster had the highest risk for mortality. They concurred that hypertension often leads to renal failure and peripheral vascular disease. Furthermore, long-standing hypertension and pulmonary circulation disorders are a common cause for congestive heart failure. Given the high risk of cardiac involvement in such patients, they recommended that treatment for COVID-19 patients in this subgroup be focused on monitoring cardiac function through an echocardiogram and cardiac telemetry (requiring ICU use), in addition to monitoring renal function through judicious use of IV fluids.

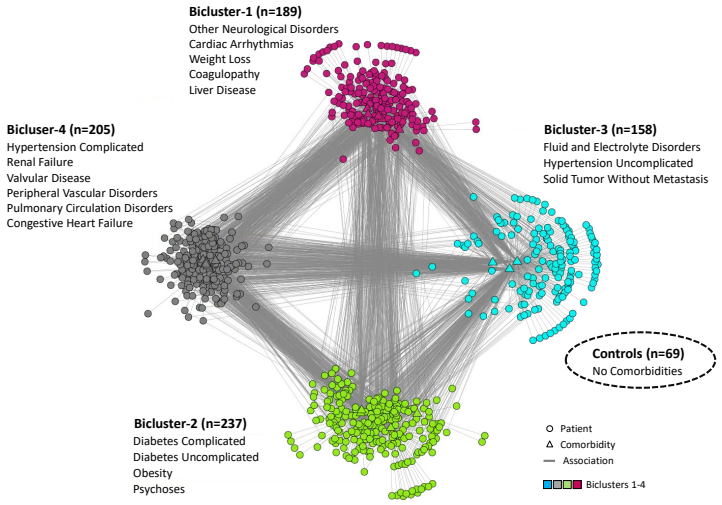


Fig. 2. Bipartite network visualization showing four biclusters each consisting of patient subgroups and their most frequently co-occurring comorbidities. Controls with no comorbidities are shown as a dotted oval.

ii. Bicluster-1 *(other neurological disorders, cardiac arrhythmias, weight loss, coagulopathy, liver disease).* This bicluster was considered the next highest risk for mortality. They concurred that liver disease and coagulopathy tend to co-occur, often accompanied by weight loss. Furthermore, several studies have shown the strong association of liver disease to neurological disorders (e.g., encephalopathy), and to cardiac arrhythmias. Given the risk of excessive bleeding through coagulopathy, they recommended that treatment for COVID-19 patients in this subgroup be focused on monitoring coagulation results (e.g., platelet count, fibrinogen, and partial thromboplastin time), and treatment through replacement therapy (e.g., transfusion through blood plasma, and fibrinogen with tranexamic acid). Furthermore, given the risk for arrhythmias, they recommended monitoring cardiac function through cardiac telemetry (requiring ICU use).

iii. Bicluster-2 *(diabetes complicated, diabetes uncomplicated, obesity, psychoses)*. This bicluster was ranked third for risk of mortality. Initially, one domain expert ranked this bicluster second for risk of mortality due to the presence of obesity, but noted that Bicluster-1 was also of high risk. They concurred that because obesity is the main cause of diabetes, the co-occurrence with uncomplicated or complicated diabetes was expected. Furthermore, medications from psychosis are a known risk for high energy consumption, leading to obesity. Given the metabolic complications arising from diabetes, they recommended that treatment for COVID-19 patients in this subgroup be focused on monitoring glucose levels, which could be done at the inpatient floor, or at home.

iv. Bicluster-3 *(fluid and electrolyte disorders, hypertension uncomplicated, solid tumor without metastasis)*. This cluster was ranked by both as the lowest risk for mortality. The co-occurrence for these comorbidities in a subgroup was probably related to medications because hypertension medication and chemotherapy are both known to cause electrolyte imbalance. Given the low risk of dying, they recommended that treatment for this COVID-19 patient subgroup should be focused on monitoring fluid and electrolytes either on the inpatient floor, or at home if their laboratory test results were normal.

	Comorbidities	Outcomes		Complications		Demographics		Abnormal Labs (ordered by signif.)
Bicluster-4	Hypertension Complicated	ICU No-Vent	44.4% (OR=9.53, CI=4.01-28.20, $P<.001$)	ARDS	41.5% (OR=2.48, CI=1.33-4.86, $P<.05$)	black	19.8% ($P=0.33$)	
	Renal Failure	ICU With-Vent	15.1% (OR=10.93, CI=2.26-196.82, $P<.05$)	AKI	11.2% (OR=4.08, CI=1.16-25.90, $P=0.10$)	white	59.9% ($P<.001$)	D_dimer, HGB, HCT, Albumin, Creatinine,
	Valvular Disease	Died	21.5% (OR=16.31, CI=3.42-292.56, $P<.05$)			asian	13.5% ($P=0.92$)	CO2, PLT, INR, CRP, Na, PTT,
	Peripheral Vascular Disorders					other	6.8%	Gamma_Gap, SBP, AST, Mg, LDH, CA,
	Pulmonary Circulation Disorders					male	51.7% ($P<.05$)	Abs_lymph_Ct, ANIONGAP, K, RDW
	Congestive Heart Failure					mean age	73.06 ($P<.001$)	
Bicluster-1	Other Neurological Disorders	ICU No-Vent	56.1% (OR=15.34, CI=6.43-45.50, $P<.001$)	ARDS	54.0% (OR=3.81, CI=2.04-7.48, $P<.001$)	black	18.3% ($P=0.24$)	Albumin, D_dimer, CA, AST, CRP, PLT,
	Cardiac Arrhythmias	ICU With-Vent	39.7% (OR=41.69, CI=8.88-744.52, $P<.01$)	AKI	11.1% (OR=3.94, CI=1.11-25.12, $P=0.11$)	white	43.3% ($P=0.09$)	LDH, Lactate, PHOS, K, Na, Creatinine,
	Weight Loss	Died	14.8% (OR=11.29, CI=2.33-203.60, $P<.05$)			asian	21.1% ($P=0.49$)	INR, Mg, HCT, Ferritin, PTT, IL6, O2SAT,
	Coagulopathy					other	17.3%	ANIONGAP, Gamma_Gap, WBC, HGB,
	Liver Disease					male	54.5% ($P<.05$)	Abs_Nphil_Ct, Procal, Triglyceride, IL8,
						mean age	58.78 ($P<.001$)	CO2, SBP, Fibrinogen, Abs_lymph_Ct
Bicluster-2	Diabetes Complicated	ICU No-Vent	39.7% (OR=7.07, CI=2.98-20.90, $P<.001$)	ARDS	38.0% (OR=2.01, CI=1.08-3.91, $P=0.06$)	black	26.3% ($P=0.95$)	
	Diabetes Uncomplicated	ICU With-Vent	20.3% (OR=15.26, CI=3.21-273.33, $P<.05$)	AKI	4.2% (OR=1.23, CI=0.30-8.28, $P=0.83$)	white	39.9% ($P=0.17$)	
	Obesity	Died	10.1% (OR=6.27, CI=1.26-113.73, $P=0.12$)			asian	17.8% ($P=0.87$)	Albumin, Creatinine, D_dimer, CA, AST,
	Psychoses					other	16.0%	CO2, CRP, Procal, HCT, Na, Lactate,
						male	47.0% ($P=0.11$)	PLT, K, Mg, INR
						mean age	54.22 ($P<.001$)	
Bicluster-3	Fluid and Electrolyte Disorders	ICU No-Vent	31.6% (OR=5.58, CI=2.29-16.77, $P<.01$)	ARDS	26.6% (OR=1.23, CI=0.63-2.50, $P=0.61$)	black	15.5% ($P=0.11$)	
	Hypertension Uncomplicated	ICU With-Vent	9.5% (OR=6.69, CI=1.30-122.49, $P=0.11$)	AKI	1.9% (OR=0.65, CI=0.11-5.04, $P=0.70$)	white	50.0% ($P<.05$)	
	Solid Tumor Without Metastasis	Died	10.1% (OR=7.76, CI=1.53-141.55, $P=0.09$)			asian	23.0% ($P=0.35$)	D_dimer, K, PLT, Na, Creatinine, CO2,
						other	11.5%	CRP, AST, Albumin, WBC, HCT, CA
						male	43.0% ($P=0.32$)	
						mean age	61.10 ($P<.001$)	
Controls	No comorbidities	ICU No-Vent	7.2%	ARDS	21.7%	black	27.7%	No significant abnormal lab tests
		ICU With-Vent	1.4%	AKI	2.9%	white	27.7%	
		Died	1.4%			asian	15.4%	
						other	29.2%	
						male	33.3%	
						mean age	34.64	

Table 1. Subgroup-level analysis showing for each of the four biclusters, their outcomes, complications, demographics, and abnormal laboratory tests (only significant laboratory test results after multiple testing correction are shown; OR and CI values are not shown due to space constraints). Cells highlighted in yellow are significant compared to the control group after multiple testing correction.

2. Subgroup-Level

Quantitative and Visual Results. The analysis of risk at the subgroup-level entailed comparing the outcomes of the four patient subgroups in each of the biclusters, to the control group. As shown in Table 1, the results showed that Bicluster-4 and Bicluster-1 had significant ORs for all three adverse outcomes (ICU No-Vent, ICU With-Vent, and mortality), whereas Bicluster-2 had a significant OR for only ICU No-Vent and ICU With-Vent, and Bicluster-3 had a significant OR for only ICU No-Vent, compared to the controls. Furthermore, Bicluster-4 and Bicluster-1 had significantly high ORs for ARDS, but neither of the other two biclusters had significant risks for any complications, compared to the controls. Finally, Bicluster-4 and Bicluster-1 both had significant ORs for males, Bicluster-4 had significantly OR for white race, and Bicluster3 had a significantly higher OR for white race, compared to the controls. Fig. 3 shows the same network as in Fig. 2, but the patient nodes were colored based on their mortality status. Furthermore, the blue text shows the percentage of patients for each of the three outcomes, and which of them had significant ORs (highlighted in yellow) compared to the controls.

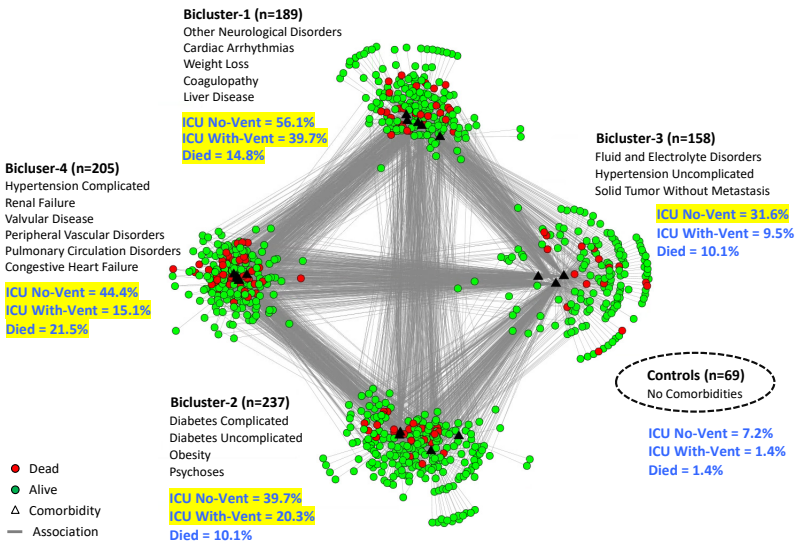


Fig. 3. Bipartite network visualization with the same 4 biclusters shown in Fig. 2, but showing patients that died (colored red) in each bicluster, in addition to percentages of all three outcomes (shown in blue text) and which of them were significant (highlighted in yellow) compared to the control group.

Qualitative Results. The domain experts used both the visualization in Fig. 2, and Table-1 to interpret the results, and to propose triage strategies for each of the biclusters. COVID-19 patients can arrive either at (1) the clinic, from where they can be triaged to home or to the ER, or (2) the emergency room (ER), from where they can be triaged to go to the ICU, inpatient floor, or home. Using this patient flow model, the domain experts recommended why and how they would triage patients from each bicluster if they arrived at the clinic, or at the ER.

i. Bicluster-4. The patients in this bicluster had a significant OR for dying from ARDS, were older (mean age 73.4), and had abnormal labs related to cardiac (systolic blood pressure), renal (creatinine), and pulmonary (CO2 and HGB) dysfunction. Given the comorbidities, it was not surprising that this bicluster had a high mean age, and a significantly higher percentage of whites and males known to have a high-risk for COVID-19 severity. The domain experts therefore recommended the following triage strategies: if such patients arrived at the clinic, they should be triaged to the ER, and if they arrived at the ER they should be triaged to the ICU.

ii. Bicluster-1. Despite having differences in comorbidities, this bicluster had a similar risk profile for adverse outcomes compared to Bicluster-4. However, the difference in comorbidity profile (liver disease, weight loss, and coagulopathy) was reflected in significant ORs for albumin, gamma gap, fibrinogen, and platelets with significantly higher AST, INR, PTT compared to the controls. In addition, indirect evidence of cardiac arrythmias is demonstrated by significant ORs for Na, K, Phos, Mg, and Ca, compared to controls. This bicluster also had a higher proportion of males known to have a high risk for COVID-19 severity. Despite these differences, the domain experts recommended the same triage strategy as for Bicluster-4: patients arriving at the clinic should be triaged to the ER, and patients arriving at the ER should be triaged to the ICU.

iii. Bicluster-2. Patients in this bicluster had a high risk for ventilator use and ICU, but not for ARDS or for dying. Furthermore, this bicluster had evidence of sequelae/complications of diabetes including renal insufficiency and electrolyte abnormalities (ketoacidosis) demonstrated in significant ORs for Cr, CO2, lactate, Na, K, and Mg, compared to the controls. Given the lower risk of dying, the domain experts recommended the following triage strategies: patients arriving at the clinic with respiratory difficulties, should be triaged to the ER, else sent home with instructions to monitor glucose levels; patients arriving at the ER, should be triaged to the inpatient floor and monitored closely for the need of a ventilator.

iv. Bicluster-3. Although the patients in this bicluster had no significant risk for ARDS or for dying, they did have a high risk for ICU use. The electrolyte imbalances in their profile is reflected in the significant ORs for Na, K, and Ca, compared to controls. However, the domain experts believed that the use of ICU may not be warranted for such patients during resource-constrained situations, as it should be reserved for more critical patients. Therefore, they recommended that patients in this bicluster be sent home if the lab values were normal, with recommendations to monitor changes in electrolytes.

3. Patient-Level

Quantitative and Visual Results. As shown in Fig. 3, the network layout revealed that the biclusters appeared to have different internal topologies. While Bicluster 1-3 had many patients on their periphery that had only one comorbidity (shown by the single edge that connects them to cluster), Bicluster-4 had very few of such patients. Furthermore, patients on the inner part of each bicluster had many comorbidities outside their bicluster (shown by the many inter-bicluster edges). These differences in topologies strongly suggested that patients in each bicluster varied in their degree of similarity *to each other*, and therefore warranted examination at the patient-level of granularity.

Fig. 4 shows the distributions of patient similarity in each bicluster. Each plot shows the distribution of the mean Jaccard Distance (JD) of each patient to the other patients in their bicluster (patients with smaller mean JD share more comorbidities with the rest of the patient in their bicluster, and are therefore are more similar). For example, Bicluster-4 (Fig. 4A) had a majority of the patients that were more similar to each other (median JD=0.59), compared to Bicluster-3 (Fig. 4C) which had a majority of the patients that were less similar (median JD=0.63) to each other in their comorbidity profiles (the pairs of biclusters had significantly different medians except for Bicluster-1 and Bicluster 3). Patients that were significantly more dissimilar compared to the rest of the patients in their bicluster

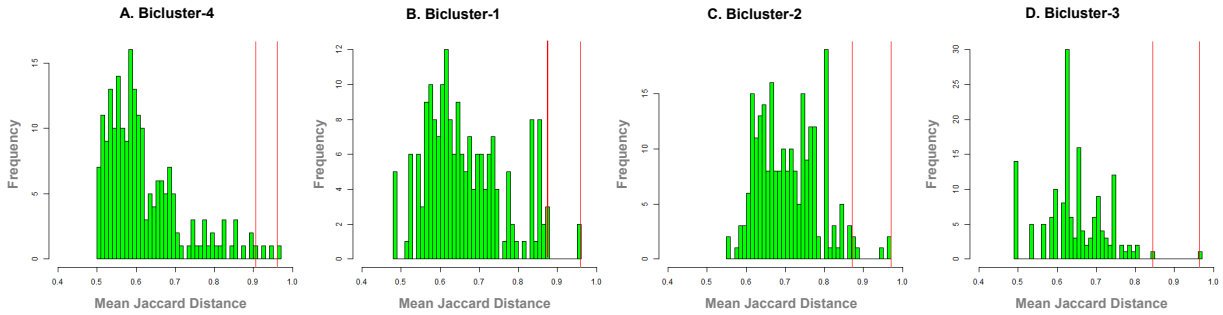


Fig. 4. Distributions of mean Jaccard Distance (JD) showing how the similarity among patients differed across the biclusters. The vertical red lines denote patients who were significantly dissimilar in their comorbidity profile compared to the rest in their bicluster, with the maximum and minimum comorbidities.

ID	Bicluster	Mean JD	Z-Score	Comorbidity Profile
P1	4	0.905338	2.882915	Valvular Disease, Peripheral Vascular Disorders, **Coagulapathy**
P2	4	0.960334	3.439748	Pulmonary Circulation Disorders
P3	1	0.958274	2.860693	Liver Disease
P4	1	0.879985	2.095412	Weight Loss
P5	2	0.871650	2.067805	Psychosis, **Other Neurological Disaeses**, **Hypertension Uncomplicated**
P6	2	0.969868	3.31486	Psychoses
P7	3	0.844305	2.49155	Hypertension Uncomplicated, Solid Tumor Without Metastasis, **Hyptertension Complicated, Liver Disease, Weight Loss**
P8	3	0.963323	3.977059	Solid Tumor Without Metastasis

Table 2. Eight outlier patients, two from each cluster that were significantly different from the rest in their bicluster, which were inspected by the domain expert to determine whether and how their treatments would be different from the bicluster to which they belonged. Colors denote biclusters shown in Fig. 2, and bolded comorbidities denote those that were outside the biclusters.

would therefore tend to be in the right tail of the respective mean JD distribution, and were considered outliers to their respective biclusters. To examine a representative sample of such patient outliers in each bicluster, we selected all patients that had a significantly higher mean JD (patients to the right of the each distribution) compared to the rest of the patients, and from those selected for examination those that had the minimum (capturing patients in the outer periphery of their bicluster), and the maximum (capturing patients in the inner periphery of their bicluster) number of comorbidities. The vertical red lines in Fig. 4 A-D shows the resulting two patients within each of the four bicluster that were selected for examination by the domain experts.

Qualitative Results. Table-2 shows the above eight outlier patients, the biclusters to which they belonged, and their comorbidity profiles. The domain experts examined these patients and their comorbidity profiles to recommend appropriate clinical interventions (to reduce the risk of reidentification, we were unable to examine and present the full profile of the patients including outcomes, complications, lab values, and demographics).

The results showed that a majority of the outlier patients selected for examination had comorbidities within their biclusters (P2, P3, P4, P6, P8), whereas a few (P1, P5, P7) had comorbidities outside their clusters (shown bolded in Table 1). However, despite having comorbidities within their bicluster, the generic treatment plans for the entire subgroup were often not appropriate. For example, while the comorbidities in Bicluster-1 overall suggest cardiac monitoring with pharmacologic control of the heart rate and correction of coagulopathy (with anticoagulation), its outliers P3 and P4 would require only replacement of clotting factors and nutritional support respectively. Similarly, while patients in Bicluster-2 overall would require glycemic control with insulin or other medications, the same treatment would cause harm to its outliers P5 and P6. Additionally, while Bicluster-3 as a whole suggests hypertension control using treatments such as diuretics, beta blockers, angiotensin converting enzyme inhibitors, its outlier P8 might be harmed with utilization of such anti-hypertensive medication. In contrast, the outliers P1 and P2 in Bicluster-4 would benefit from the same cardiac support therapy as the full bicluster. The analysis of heterogeneity at the patient-level of granularity therefore revealed that while subgroups provide efficiency in the design of treatment plans, outliers that need different treatments from their biclusters need to be flagged, underscoring the complexity of treating patients with multimorbidities.

Discussion

Several studies have shown that COVID-19 patients with prior MCCs, being older, male, and a minority are all high risk factors for having adverse outcomes. However, little is known about how prior comorbidities co-occur to form COVID-19 patient subgroups, their risks for adverse outcomes, and their implications for clinical interventions. This is particularly important because multimorbidities are common and well-studied in older adults and minorities. However, while these studies have analyzed how multimorbidities co-occur in different populations such as patients that have been readmitted to the hospital after a hip fracture,[28] they have been done using a wide range of different methods, and at different levels of granularities.

Given the critical importance of designing interventions to reduce the risk of adverse outcomes in COVID-19 patients, here we explored a *unified approach* to (1) automate and therefore accelerate the quantitative analysis of heterogeneity at different levels of granularity, and (2) visualize the results using the same representation to increase interpretability of the results with the explicit goal of designing clinical interventions. We used the bipartite network representation because it explicitly represented both patients and comorbidities simultaneously using a computable graph representation consisting of nodes and edges, which enabled their quantitative and visual analysis at different levels of granularity. The application of this approach to COVID-19 EHR patient data led to the following two insights:

Explicit and Suggestive Clinical Insights at Each Level of Granularity. Analysis of heterogeneity at each level of granularity led to explicit clinical insights that were enabled by the respective analytical methods used. At the cohort-level, modularity maximization identified biclusters which provided insights on which comorbidities frequently co-occurred. This led to inferences about the disease mechanisms that potentially connected them (e.g., hypertension →

pulmonary circulation disorders → congestive heart failure), enabling a focus on monitoring the organ or system (e.g., using telemetry to monitor the heart) that could precipitate an adverse outcome. Other applications could include the design of clinical trials. At the subgroup-level, comparative analysis between subgroups enabled ranking biclusters based on their risks for specific outcomes, resulting in the design of triage strategies critical during resource-constrained situations such as COVID-19 hotspots. Furthermore, this analysis also revealed how age and gender were stratified among the high-risk biclusters. Finally, at the patient-level, analysis of similarity helped to identify which patients were outliers to their biclusters, and whether they would require different treatments compared to their subgroups. Examining each level separately therefore helped to elucidate the contribution (frequency, risk, and similarity) each played in the design of clinical intervention for COVID-19 patients.

However, each level of granularity also provided suggestive insights for the next level. For example, at the cohort-level, the use of telemetry suggested that the patients needed to be triaged to the ICU, an insight which was more fully realized when analyzing subgroups at the next level of granularity. Similarly, analysis at the subgroup-level was suggestive that patients within each bicluster ranged in the degree to which they were similar to the rest in their bicluster, but more fully realized when analyzing individual patients at the next level of granularity. Such connections between levels could be the result of using a uniform bipartite representation to analyze heterogeneity across all levels of granularity, which needs to be further explored.

Extensions and Limitations. While using the multigranular heterogeneity framework, we realized that although the analysis at the cohort-level was within the cohort, the analysis at the subgroup and patient-levels were between subgroups and patients respectively. The framework could therefore be elaborated to include inter- and intra-level analysis. For example, at the cohort-level intra-cohort analysis would be the current modularity maximization to identify biclusters within the cohort, but inter-cohort analysis could include analyzing if the co-occurrence patterns of comorbidities in one dataset, replicate in another using methods such as the Rand Index.[19] Furthermore, the framework could include other analytical approaches such as causal modeling and association rule mining for conducting intra-subgroup analysis. Additionally, the domain experts noted that as the comorbidities were defined by disease categories such as liver disease, the data lacked details about subtypes resulting in their inability to incorporate etiology, severity, and chronicity into their design of clinical interventions. However, despite this limitation, the current dataset was sufficient to enable a provider to triage and make initial treatment decisions quickly and efficiently. Finally, the outlier detection at the patient-level analysis currently does not take into consideration the shape and size of the distributions, and our current research is exploring other statistical methods to more precisely identify those outliers.

Conclusions and Future Research

Heterogeneity and granularity are well-known and critical concepts in biomedical research. Heterogeneity embraces the notion that patients are similar and different depending on the characteristics used to describe them. This notion has led to an understanding of phenomena such as phenotypes and symptom clusters, and is a corner stone of precision medicine. Granularity embraces the notion that patients can be analyzed at different levels of detail ranging from the molecular to the environmental. This notion has led to approaches such as molecular medicine, and is a corner stone of translational science. Here we attempted to merge both concepts to analyze the co-occurrence of comorbidities in COVID-19 patients, with the explicit goal of translating heterogeneity results from each level of granularity into clinical interventions. Such an analysis was possible through the use of bipartite networks as they provided a unified quantitative and visual representation, which enabled (1) automation for the quantitative analysis of heterogeneity at each level, and (2) the rapid interpretation of that heterogeneity by domain experts through the visualization, leading to the design of clinical interventions.

The results suggest that each level of granularity can provide distinct insights into the co-occurrence of comorbidities: (1) cohort-level analysis can be used to provide insights into the *frequency* of co-occurrence patterns enabling recommendations on which co-occurrences should be included in clinical practice guidelines to address multimorbidities; (2) subgroup-level analysis can be used to provide insights into the *risk* of each subgroup, enabling the design of triage strategies critical in resource-constrained situations such as COVID-19 hotspots; (3) patient-level analysis can be used to provide insights into the *similarity* of patients in each subgroup useful to determine which patients can use interventions designed for the subgroup, and which require individualized interventions. While ultimately each patient is an individual and requires personalized care, the goal of such analysis is to enable the design of evidence-based proactive strategies, which are adaptable to specific situations with the goal of improving the quality and efficiency of care.

A critical limitation of the current research is that we analyzed only one dataset, and our current and future research will test the replicability of these results in another COVID-19 dataset. Furthermore, we will explore the use of the multigranular heterogeneity framework to analyze other phenomena such as symptom clusters and clinical phenotypes with the explicit goal of translating the analytical results from each level of granularity, to the design of clinical interventions and their evaluation.

Acknowledgements

This research was supported in part by the UTMB Clinical and Translational Science Award (UL1 TR000071) from NCATS, and the UTMB Claude D. Pepper Older Americans Independence Center Award #P30-AG024832 from NIA, and the UTMB Cancer Center.

References

1. Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: a systematic review and meta-analysis. *Int J Infect Dis.* 2020.
2. Centers for Disease Control and Prevention. Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html. Accessed April, 6, 2020.
3. Shi S, Qin M, Shen B, et al. Association of Cardiac Injury With Mortality in Hospitalized Patients With COVID-19 in Wuhan, China. *JAMA cardiology.* 2020.
4. Li Y-C, Bai W-Z, Hashikawa T. The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. *J Med Virol.* 2020.
5. Zou X, Chen K, Zou J, Han P, Hao J, Han Z. Single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to 2019-nCoV infection. *Front Med.* 2020.
6. Mao L, Wang M, Chen S, et al. Neurological Manifestations of Hospitalized Patients with COVID-19 in Wuhan, China: a retrospective case series study. *medRxiv.* 2020:2020.2002.2022.20026500.
7. McIntosh K, Hirsch M, S, Bloom A. Coronavirus disease 2019 (COVID-19). 2020; https://www.uptodate.com/contents/coronavirus-disease-2019-covid-19, 2020.
8. World Health Organization. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf. Accessed April, 6, 2020.
9. Griffith DM, Sharma G, Holliday CS, et al. Men and COVID-19: A Biopsychosocial Approach to Understanding Sex Differences in Mortality and Recommendations for Practice and Policy Interventions. *Preventing chronic disease.* 2020;17:E63.
10. Hajat C, Stein E. The global burden of multiple chronic conditions: A narrative review. *Prev Med Rep.* 2018;12:284-293.
11. HHS. About the Multiple Chronic Conditions Initiative. 2020; https://www.hhs.gov/ash/about-ash/multiple-chronic-conditions/about-mcc/index.html#:~:text=MCC%20are%20concurrent%20chronic%20conditions,both%20have%20multiple%20chronic%20conditions. Accessed 8/14, 2020.
12. Gupta A, Madhavan MV, Sehgal K, et al. Extrapulmonary manifestations of COVID-19. *Nat Med.* 2020;26(7):1017-1032.
13. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *Jama.* 2005;294(6):716-724.
14. Boyd CM, Wolff JL, Giovannetti E, et al. Healthcare task difficulty among older adults with multimorbidity. *Medical care.* 2014;52 Suppl 3(0 3):S118-S125.
15. Tinetti ME, Bogardus ST, Jr., Agostini JV. Potential pitfalls of disease-specific guidelines for patients with multiple conditions. *The New England journal of medicine.* 2004;351(27):2870-2874.
16. Muth C, Blom JW, Smith SM, et al. Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: a systematic guideline review and expert consensus. *Journal of internal medicine.* 2019;285(3):272-288.
17. Guthrie B, Payne K, Alderson P, McMurdo MET, Mercer SW. Adapting clinical guidelines to take account of multimorbidity. *BMJ : British Medical Journal.* 2012;345:e6341.
18. Lochner KA, Cox CS. Prevalence of multiple chronic conditions among Medicare beneficiaries, United States, 2010. *Preventing chronic disease.* 2013;10:E61.
19. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* New York, NY, USA: Springer New York Inc.; 2001.
20. Abu-jamous B, Fa R, Nandi AK. *Integrative Cluster Analysis in Bioinformatics.* Chichester, West Sussex, United Kingdom: John Wiley & Sons, Ltd.; 2015.
21. Violán C, Roso-Llorach A, Foguet-Boreu Q, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. *BMC Family Practice.* 2018;19(1):108.
22. Triest FJJ, Franssen FME, Reynaert N, et al. Disease-Specific Comorbidity Clusters in COPD and Accelerated Aging. *Journal of clinical medicine.* 2019;8(4).
23. Padilha VA, Campello RJGB. A systematic comparative evaluation of biclustering techniques. *BMC bioinformatics.* 2017;18(1):55.
24. Newman MEJ. *Networks: An Introduction.* Oxford, United Kingdom: Oxford University Press; 2010.
25. Chauhan R, Ravi J, Datta P, et al. Reconstruction and topological features of the sigma factor regulatory network of Mycobacterium tuberculosis. In Review.
26. Kamada T, Kawai S. An algorithm for drawing general undirected graphs. *Information Processing Letters.* 1989;31:7-15.
27. Bhavnani SK, Chen T, Ayyaswamy A, et al. Enabling Comprehension of Patient Subgroups and Characteristics in Large Bipartite Networks: Implications for Precision Medicine. *Proceedings of AMIA Joint Summits on Translational Science.* 2017:21-29.
28. Bhavnani SK, Dang B, Penton R, et al. How High-Risk Comorbidities Co-Occur in Readmitted Patients With Hip Fracture: Big Data Visual Analytical Approach. *JMIR Med Inform.* 2020;8(10):e13567.
29. Nickerson RS. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology.* 1998;2(2):175-220.

A Contextual Inquiry: FDA Investigational New Drug Clinical Review

Jonathan Bidwell PhD[a*], Kara Whelply MBA, MPH[a*], Sophia Shepard[b], John Hariadi, MD[a]

[a]U.S. Food and Drug Administration (FDA), Silver Spring, MD, US, [b]University of Maryland, College Park, MD, Denotes equal contribution*

ABSTRACT

The U.S. Food and Drug Administration (FDA) is modernizing IT infrastructure and investigating software requirements for addressing increased regulator workload and complexity requirements during Investigational New Drug (IND) reviews. We conducted a mixed-method, Contextual Inquiry (CI) study for establishing a detailed understanding of daily IND-related research, writing, and decision-making tasks. Individual reviewers faced notable challenges while attempting to search, transfer, compare, consolidate and reference content between multiple documents. The review process would likely benefit from the development of software tools for both addressing these problems and fostering existing knowledge sharing behaviors within individual and group settings.

1. Introduction

The FDA is modernizing I.T. infrastructure for enabling safer and more expedient drug approval (Administration, 2019). New data sources and the increasing volume and complexity of drug approval applications have prompted the agency to investigate existing work practices for establishing software requirements.

In this study, we investigated daily software requirements during the FDA's Investigational New Drug (IND) process. The study included review team members from the FDA Center for Drug Evaluation and Research (CDER), Department of Psychiatry (DP) and examined how work gets done within a typical group and individual workplace settings.

To gain a fresh perspective, we conducted a mixed-methods study design that included a Contextual Inquiry (CI), semi-structured interviews, and an online survey. The results highlighted important software requirements that we would likely have missed had we used traditional qualitative methods alone. Moreover, we established a user-driven consensus for prioritizing our subsequent software development efforts and showed that our study design was feasible to conduct within a large organization that handles propriety information.

1.1. FDA Review Process

The mission of the U.S. Food and Drug Administration (FDA) is to protect public health by ensuring the safety, efficacy, and security of human and veterinary drugs, biological products, and medical devices (Administration, 2018).

The FDA CDER Office of New Drugs (OND) reviews sponsor IND applications and offers guidance for encouraging safe and expedient drug approval [2]. IND applications include proposed clinical protocols for clinical testing with human subjects along with relevant animal pharmacology studies, toxicology studies, and manufacturing information.

Each IND is delegated to a specific review division such as the Department of Psychiatric Products and is assigned to a review team. The review team consists of a regulatory project manager (RPM), clinical, and non-clinical team members who are coordinated by team leaders (TL). The clinical TL assigns the IND to a clinical reviewer (CR), also known as medical officers. The non-clinical TLs assign the IND to non-clinical reviewers (NCRs) such as pharmacotoxicologists and individual discipline TLs assign chemists, statisticians, and other disciplines as needed. Meanwhile, the RPM schedules meetings, communications with the Sponsor, and organizes resources for the team, including past IND information and a SharePoint website. NCR/CRs analyze and review the submitted IND materials to write safety reports and identify areas of concern. Each review team member typically has multiple active IND applications at the same time.

After analysis, a Supplemental Release Date (SRD) meeting is held with the division head (DH) to discuss safety issues, propose safety guidance and finalize a hold/non-hold letter for the Sponsor to ensure that research subjects will not be subject to unreasonable risk (Administration, 2020).

1.2. Contextual Inquiry

Contextual inquiry (CI) is a user-centered research methodology that seeks to capture and understand user work's context by immersing researchers in the user environment through participatory observation sessions (Beyer & Holtzblatt, 1998; Wixon, 1990). CIs often require fewer resources than focus groups (Guest , et al., 2017; Smithson, 2000) and are less sensitive to peer pressure influence (Greg Guest, 2017). CIs have been used widely within industry, government, and academic organizations (Coble, et al., 1995; PRITCHARD, 2019) for developing suitable IT solutions (Beyer & Holtzblatt, 1998).

Much like an apprentice learning a skill, researchers go where the work is being conducted and ask questions to clarify what users are doing as they work (Beyer & Holtzblatt, 1998; Wixon, 1990). Instead of strictly observing as in shadowing (Daae, 2015) or asking direct questions as in interviews, researchers observe and probe at the same time to better understand how the work is accomplished. For this reason, CIs are well suited to collect *tacit knowledge* that can be difficult to ascertain with other qualitative methods.

The collection of tacit knowledge is essential because many of our daily routines have become second nature to us. Important aspects of these daily routines are often challenging for us to recall without being engaged in the work. Instead of asking users to explain a hypothetical work process, we joined them when and where they worked during the 30-day IND process to identify workflow breakdowns and problem-solving strategies.

2. Related Work

Establishing accurate user requirements is critical for developing software that successfully addresses user needs. The FDA has conducted interviews, surveys (Berndt, 2006), focus groups (Parenky, 2014), and usability tests (Fitzpatrick, 1999) in the past to identify these needs; however, these approaches require establishing questions in advance. By contrast, a CI focuses on understanding the work rather than approaching requirements gathering with an initial set of questions (Beyer & Holtzblatt, 1998).

The CI methodology offers several notable advantages over these more traditional qualitative methods. For example, CIs focus on understanding work from the ground up (Beyer & Holtzblatt, 1998) without making assumptions regarding initial questions. CI's present greater ecologically validity as researcher observations occur within the same cultural and social context as the user's everyday activities (Schmuckler, 2001). Most importantly, CI's are ideal for capturing *tacit knowledge* and other nuances that may not otherwise go unaddressed within healthcare (Coble, et al., 1995), academic (Notess, 2005), and government settings.

In our case, we selected the CI methodology for the following two reasons:

First, we needed to gain a fresh perspective. FDA review teams are structured as matrix organizations where users work within a traditional hierarchy that is overlaid by some form of lateral authority (Kuprenas, 2003). CIs are well suited for documenting this collaboration between different organizational roles. Interviews and focus groups tend to be less ecologically valid as they often do not occur in the user's environment and more susceptible to peer pressure (Greg Guest, 2017). We adopted the CI methodology, created CI flow models for documenting workflow, and used direct quotes to preserve meaning and provide a systematic, detailed, and reliable understanding of work practices.

Second, we needed to establish a broader consensus and buy-in across the review team. CIs are often supplemented with additional methods. For example, Maffitt et al. created CI models and affinity diagram sessions to identify user requirements among physicians (Coble, et al., 1995). The models were used to consolidate multiple sets of notes for understanding how the workflow occurred while the affinity diagram helped to identify user requirements (Coble, et al., 1995). In addition to conducting our CI, we also conducted semi-structured interviews (Notess, 2005) and administered an affinity ranking survey (Coble, et al., 1995) to encourage stakeholder participation and better prioritizing our subsequent design and software development efforts.

3. Methods

The study had two parts. The first part included three sessions with the entire review team during SRD meetings, which focused on understanding roles and responsibilities. The second part included six sessions with individual CRs, NCRs, and RPMs, which focused on understanding daily work practice and individual roles.

Each user research session included an observational period, a semi-structured interview, and a debrief session where we organized our notes and created CI flow models. In each case, at least two researchers were present. No recording devices were allowed due to strict confidentiality rules at the FDA. We conducted affinity diagraming sessions for establishing a broader set of user requirements themes across our user research sessions. Then, we administered an affinity ranking survey where we asked participants to vote on these themes for better prioritizing our future development efforts.

3.1. User Research Sessions

We studied participants within the context of three groups. Group #1 included three separate SRD meetings. Groups #2 and #3 included three individual sessions with a CR, NCR, and RPM, respectively. Each user research session included 60 minutes of direct observations, where we captured hand-written notes and asked clarifying questions as needed. Each set of observations lasted 60 minutes.

Next, we conducted a semi-structured interview that asked, "what applications do you use the most," "what type of resources do you use the most," "how do you collaborate with your co-workers" and "what was the most difficult aspect of your last review." Each interview lasted 15 minutes.

Then we created CI flow models for highlighting user roles, responsibilities, and how artifacts and information were exchanged between stakeholders (Beyer & Holtzblatt, 1998). Each flow modeling session was conducted within 48 hours of each user research session and lasted 1-2 hours.

3.2. Affinity Diagramming Sessions

The affinity diagramming sessions enabled us to consolidate our findings across multiple user research sessions. In total, we created three affinity diagrams following each round of observation sessions.

The first diagram included our observations from SRD meetings, Group #1, while the second and third diagrams included our observations from Group #2 and Group #3. In each case, we transcribed summarized sentences and direct quotes from our notes onto stickie note labels. We grouped our stickie notes in a "bottom-up" manner to identify overarching themes and relationships. Then we transcribed significant themes from each diagram as a list of user requirement themes that we later sent to participants during our affinity ranking survey. Figure 1 shows us creating our first affinity diagram.

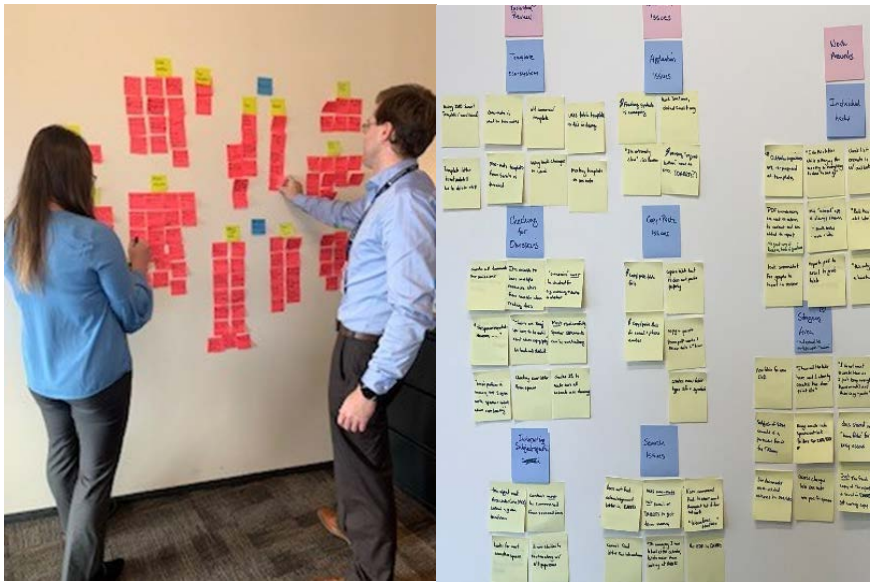


Fig. 1 – (Left) Affinity diagram creation for Group #1 (Right) Affinity diagram for Group #3

3.3. Affinity Ranking Survey

The affinity ranking survey included a list of eleven user requirement themes generated from our three affinity diagrams. The participants were asked via email to select the three most relevant themes to them to help us prioritize future development.

4. Results

The study included twenty-two rotating team members during SRD meetings and six individual team members from FDA CDER's DP. In total, we conducted nine user research sessions. Sessions with Group #1 SRD provided insight into sponsor communication at the beginning and end of the IND process as team members finalized different IND applications. By contrast, sessions with Groups #2 and #3 provided insight into daily research, writing, and scheduling tasks. The user research sessions were conducted during different stages of the FDA's IND review process, as shown in Table 1.

Table 1 - Review team roles and # user research sessions during each phase of the IND review process

	1. Receive IND application and team assignments	2. Schedule SRD meeting and create a SharePoint Site	3. Analyze and review sponsor IND materials	4. Write safety summary report for SRD meeting	5. All-hands SRD meeting to discuss safety findings and decide hold status	6. Write and send hold/non-hold letter to sponsor
CR			✓✓	✓✓		
NCR			✓✓	✓✓		
RPM	✓✓	✓✓				✓
Entire Team including DH and TLs					✓✓✓	

4.1. Models

We created nine flow models. Breakdowns are indicated with lightning bolts. Most reviewers experienced significant search and information retrieval breakdowns while using Document Archiving, Reporting and Regulatory Tracking System (DARRTS), and Mercado. DARRTS is the FDA's record-keeping system for drug applications and Mercado an analytics and visualization platform for regulatory data.

Figure 3 shows an NCR searched Mercado for an IND application, but the application contains no data. She then searches DARRTS for the same application but mistypes a number, so no results are returned. She attempts to search scanned paper documents, but the search is not available.

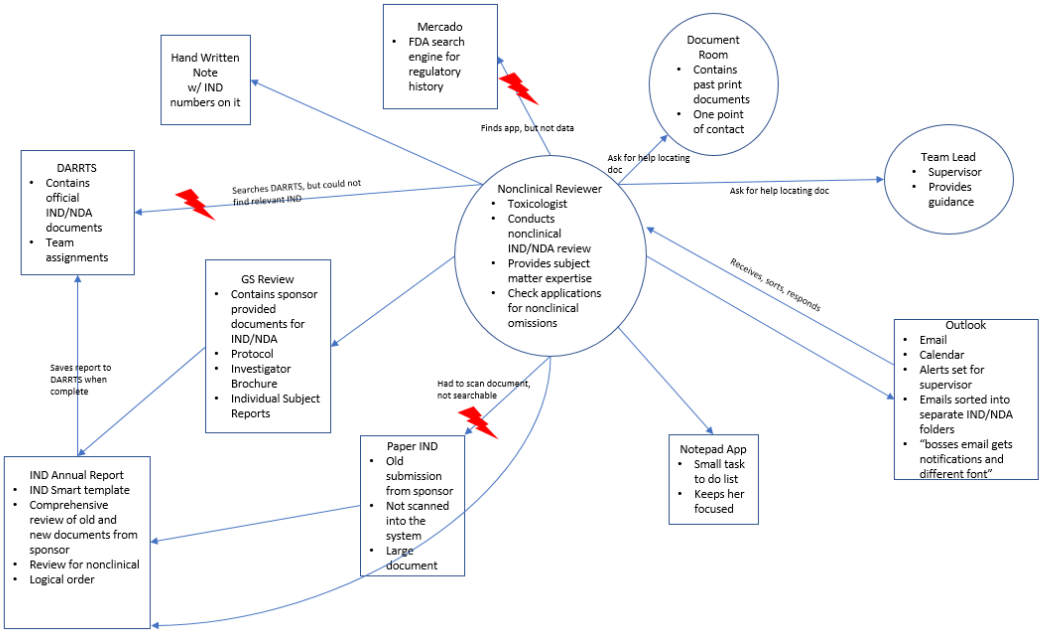


Fig. 3 - Non-clinical Reviewer Flow Model

Figure 4 shows an example of the entire group's workflow during an SRD meeting. The SRD meeting included additional roles that we could not observe during our user research sessions with the reviewer as the division head and team leader.

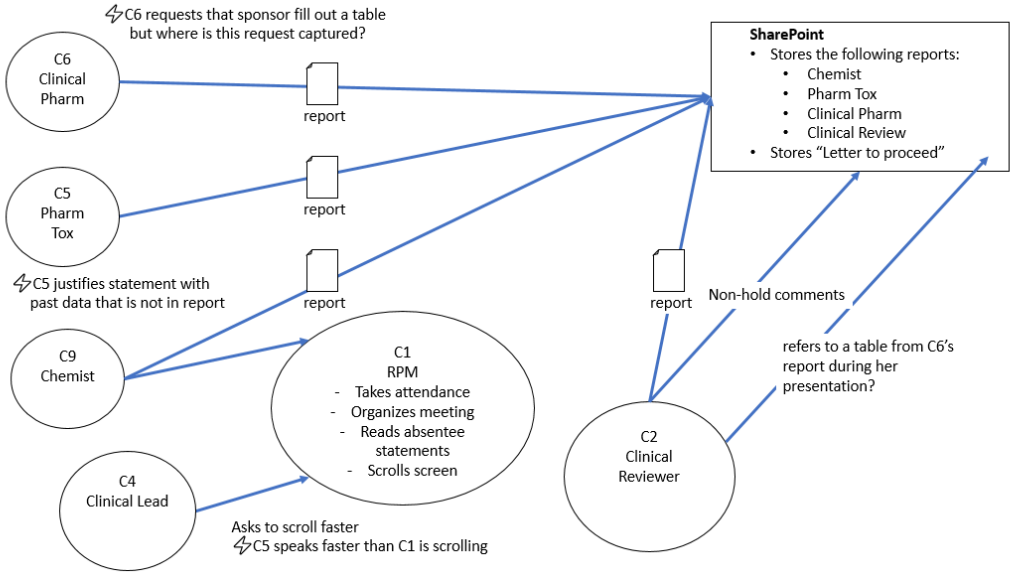


Fig. 4 – SRD Flow Model

Figure 5 shows the workflow of a CR in the role of Medical Officer. Significant breakdowns occurred during this workflow with copy and paste functionalities and difficulty with specific software. Additionally, this flow model highlights some workarounds for recall, such as bolding a stopping point and making individual folders for each IND.

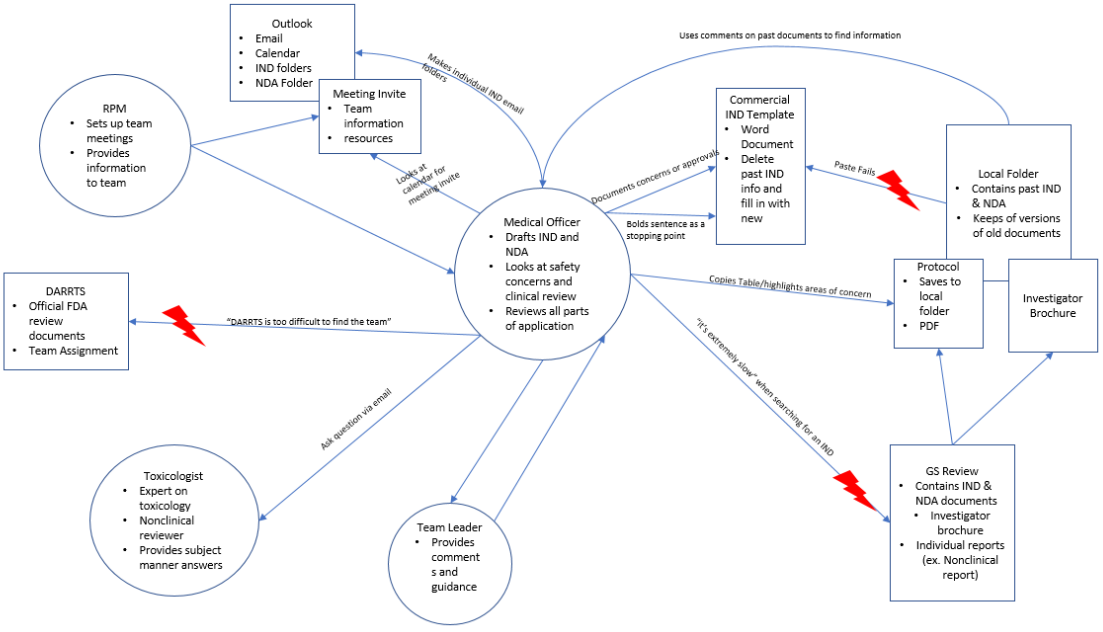


Fig. 5 – CR Flow Model

4.2. Affinity Diagrams

We created three affinity diagrams. Each affinity diagram consisted of bottom-up requirements that we structured into a hierarchy of higher-level themes. For example, we grouped "rotates through pdfs," "searches for a euthanized subject," and "checking cover letter" into the minor theme of checking for omissions, which was later organized into the theme of conducting the individual review. Table 2 highlights our affinity diagram themes. Please Appendix A for an example of one of our affinity diagrams.

Table 2 – Affinity diagrams showing major themes that were derived from bottom up requirements

	User research session	# Requirements	Major affinity diagram themes
1	Group #1-SRD meetings	117	• meeting preparation* • information retrieval • creating external deliverables
2	Group #2 - CR, NCR, and RPM	113	• workflow • writing process* • meeting preparation* • sponsor communication
3	Group #3- CR, NCR, and RPM	124	• team collaboration • workarounds • technical issues • recall issues • conducting individual review

** donates duplicate theme*

4.3. Semi-structured interviews

We conducted nine semi-structured interviews. NCR and CR respondents most often used Microsoft Word (3 of 4 responses), CRs (medical officers) most commonly used DARRTS and G.S. Review (2 of 2 responses), and Regulatory Project Managers (RPMs) most often used Outlook. Most participants used colleagues or supervisors as information resources (5 of 6 responses).

Similarly, all participants reported using email (Microsoft Outlook) or face-to-face communication for collaboration (6 of 6 responses). All NCRs and CRs noted that the most challenging part of their last review was interpreting and validating information between multiple sources. Interestingly, all RPMs reported that scheduling meetings were the *most* challenging task despite the availability of Outlook scheduling assistant.

4.4. Affinity Ranking Survey

Table 3 shows our affinity diagram user requirement themes sorted by the number of votes.

Table 3 – Affinity ranking survey themes sorted by the number of respondent votes

	Affinity ranking survey themes	**# Votes**
1	**Information Retrieval:** how you find information, i.e., command find, Google, Mercado	4
2	**Conducting Individual Review:** your individual process to write your review including research and analysis	3
3	**Team Collaboration:** how you work with your colleagues and share resources, e.g., sharing templates	2
4	**Work Arounds:** creative solutions to accomplish your work, i.e., pdf to excel for tables, linking OneNote to Outlook	2
5	**Recall Issues:** when you cannot remember where something is in a document or where you left off	1
6	**Writing Process:** how you write your review, i.e., rewording statements, proofing	1
7	**Communication with Sponsor:** creating and sending information to the Sponsor	1
8	**Meeting Preparation:** how you prepare for meetings	1
9	**Creating External Deliverables:** developing and collaborating on the documents to send to a sponsor, i.e., placing your hold or non-hold comments	0
10	**Technical Issues:** when an application does not work, i.e., when copy and paste fails	0
11	**Workflow:** how you do things, i.e., coordinating meetings, analysing sponsor data	0

5. Discussion

The findings provided us with a fresh perspective for supporting the IND review process. Being able to observe daily IND tasks first-hand, enabled us to better appreciate how the work happens and establish design requirements that we likely would not have made using traditional qualitative methods alone. Notable shortcomings and strengths can be addressed in a follow-on project. The bulk of these shortcomings directly impact review team members by causing repetition, being time-consuming, and tedious. These shortcomings can be categorized as individual tasks, collaborative work, information retrieval, data extraction, proofing, and cross-checking documents. Most notably, reviewers often needed to search, compare, transfer, consolidate and reference multiple documents while writing safety reports yet existing software required them to perform extensive workarounds to complete these tasks. Strengths included a culture of knowledge sharing and building upon established user consensus. For example, reviewers shared physical items such as templates and other general knowledge to assist others with finding documents. Software could be developed to address these shortcomings and promote these strengths.

The key findings from our user research sessions are as follows:

5.1. Information retrieval and interruptions were problematic during individual and group tasks.

Information retrieval was voted as the most important priority during our affinity ranking survey (4 out of 5 respondents). FDA's Document Archiving Reporting and Regulatory Tracking System (DARRTS) had numerous shortcomings, including slow response times when searching, limited filtering capabilities, and being unable to handle typos. Information was archived and not available in DARRTS required tracking down prior reviewers who may no longer be working in the same department and/or request paper documents. A CR stated that "DARRTS is too difficult to identify team members" and instead relied on search capabilities within her Outlook email. A CR expressed frustration after missing a digit when searching for an IND. An NCR blamed herself for these difficulties and told us that "I am new here" and that "I do not even know how to search" despite her being a reviewer for over three years. Significant breakdowns occurred in our flow models that highlighted the importance of information retrieval. These breakdowns focused on review team

members being unable to find information related to IND assignments within DARRTS, SharePoint and Mercado for information such as team assignments, sponsor documents and text content within documents. In our flow models, CRs failed to find documents in DARRTS.

Team collaboration was impacted by legacy software limitations and interruptions. Notably, certain steps in the review process were contingent on document edits; however, there was no fool-proof way to keep track of when edits occurred. Each review team member was responsible for adding comments to a hold/non-hold letter on Microsoft SharePoint 2010; however, only one person could edit the document at a time. RPMs had to first remember to enable tracked-changes before inviting reviewers to edit and then follow-up with them separately via email to confirm that they had finished editing the letter. In our flow model, we noted that this process breaks down when team members forget to enable tracked-changes within Microsoft Word. Interruptions further impacted individual tasks. Each review team member (NCR, CR, RPM) was assigned several existing IND applications. In our flow model (Figure 5), incoming emails, requests for in-person meetings often interrupt important CR writing, and research tasks. A CR marked a sentence in bold to indicate a sentence to resume work in anticipation of being pulled away. An NCR only had alerts on for emails from her superiors to help maintain her focus. Additionally, we observed reviewers using paper checklists and Microsoft OneNote checklists to keep track of specific tasks.

Improved support for information retrieval and resuming tasks would benefit most review team members. For example, showing a history of recent spreadsheet changes (Asuncion, 2011) could help reviewers to recall next steps from a previous work session. Introducing Microsoft SharePoint Online and Microsoft Teams could provide reviewers with tracked change support while also enabling real-time collaborative editing to better determine when team members have finished editing specific documents sections. Information retrieval services could index documents based on common search queries.

5.2. Individual review team members reported that data extraction, proofing, and checking for omissions were the most time-consuming and tedious for them.

Information retrieval issues further hampered efforts to compare documents when checking for omissions. A CR from Group #3 needed to check whether a euthanized animal subject was mentioned in four separate sponsor documents. Her search for the word “euthanized” failed because the PDF viewer only matched exact keywords. The synonym "terminated" was not matched. Instead she had to read several pages of text to find the animal’s subject number.

Basic copy & paste operations often failed between PDF and Word documents. Tables and other copied PDF content was transferred as raw text within Word. In three of our flow models, review team members had to stop what they were doing and recreate content from scratch before continuing. For example, an RPM copied and pasted a phone number from a PDF to paste into OneNote. The phone number appeared as symbols causing the RPM to type the number by hand. As a workaround, one NCR worked exported an entire PDF document to Excel and copying the specific tabular data that she needed. She told us that this approach only had a “50/50 chance” of working. To make matters worse, Microsoft Word often presented incorrect autocorrection suggestions after pasting. An NCR told us that “proofing” was her “most tedious task”.

Information consolidation and referencing tasks were similarly time-consuming and difficult due to the nature and length of the documents involved. Not having an easy way to keep track of references increases the risk of misrepresenting the FDA’s position and decisions when writing emails and reports. The entire review team needed to carefully review any IND sponsor's content when copying and pasting to avoid simple typographical errors. For example, a noted flow model breakdown occurred when an RPM began a scheduling email by copying and pasting a sponsor's paragraph. She would have sent the wrong information had she not caught a mistake and replaced a “type A” meeting with a “type B” meeting.

Indexing related words between documents could streamline the search for omissions between documents. Introducing the ability to paste screenshots of tables and equations could preserve formatting while transferring content between documents. Similarly, text could be pasted with formatting metadata that include a reference to the source document for keeping track of non-edited and edited content.

5.3. Existing review teams excel at knowledge sharing.

Individual reviewers were comfortable asking for help in their work environment from both colleagues and superiors. In our semi-structure interviews, participants regarded colleagues and supervisors as the best information resources. All review team members indicated that they were comfortable asking for help both in person and via email, depending on the situation or severity of the question or issue. The same was true during our sessions with individual reviewers and RPMs. In Figure 3, an NCR failed to find a document in DARRTS but succeeded after asking a team leader for help. RPMs shared resources such as email templates and acronym lists. Individual office mates shared productivity tips and Word templates.

Introducing a knowledge-sharing platform such as Microsoft Teams or Slack could help to cultivate the FDA’s knowledge-sharing culture further. For example, review team members could ask questions and receive answers from review team members with similar specializations and provide informal mentoring. FDA is currently acquiring and implementing Microsoft Teams, which could further improve the knowledge sharing capabilities.

5.4. Existing software, while not perfect, supports a broad range of needs across roles.

To date, a top-down enterprise-level adoption of tools such as DARRTS has resulted in a “one-size-fits-some” scenario where reviewers often must develop elaborate workarounds to accomplish daily tasks. For example, a CR searched Outlook for emails to find past IND information instead of internally searching through DARRTS. An RPM used the signature feature in Outlook as templates for starting emails. Electronic and paper notepads were used to assist with recalling information that was not available within the software. For example, a CR used a paper notepad to keep track of tasks related to individual INDs that she needed to complete.

The most popular software among reviewers were DARRTS, GS Review, and Microsoft Word. In our semi-structured interviews, NCRs and CRs indicated that they most often used DARRTS and GS Review; however, the software was often slow and unstable. In one case, DARRTS froze for more than 30 seconds and crashed altogether during another user research session. We consistently received requests for improved information retrieval for clinical summaries.

By contrast, the most popular software among RPMs was Outlook. RPMs used Outlook for emails and scheduling with shared calendars, SharePoint folders, and utilizing add-ins such as Cisco WebEx and Microsoft OneNote. In our affinity ranking survey, respondents indicated that information retrieval and conducting the individual review were the most important to them during the review process.

Introducing FDA-specific services and plugins could help to achieve these daily requirements while continuing to use existing software. For example, sharing search index results across DARRTS and Outlook would enable reviewers to retrieve the same search results on either platform. Email templates could be made available using a Microsoft Outlook Add-in when starting emails. Introducing Customer Relationship Management (CRM) could support role-specific needs such as accessing and annotating sponsor documents, scheduling meetings between IND applications. Multiple desktops could be used for saving and restoring workspace state. For example, reviewers could use Amazon WorkSpaces to manage all documents and browser windows associated with a given IND application.

5.5. Future work

In the future, we would like to include additional review team members for identifying a broader set of everyday needs at the FDA. Time and resource constraints limited our enrollment to twenty rotating members of DP team members. As a next step, we plan to apply Contextual Design (Beyer & Holtzblatt, 1998) for addressing multiple-document search, transfer, compare, consolidate and reference needs among review team members.

6. Conclusion

In this study, we conducted a mixed-methods study with review team members from the FDA CDER DP to investigate how work gets done during FDA's IND review process. We conducted a Contextual Inquiry (CI), semi-structured interviews, and affinity ranking survey. The results highlighted several important design challenges. Individual reviewers needed to search, transfer, compare, consolidate and reference content between multiple documents while writing safety reports yet existing software required extensive workarounds to complete these tasks. Existing knowledge sharing behaviors important for the IND review process yet are not formally supported by existing software.

Acknowledgements

Thank you to Javier Muniz, MD and Michael David, MD and the FDA DP group for your support. This project was supported in part by an appointment to the Science Education Programs at the U.S. Food and Administration, Center for Drug Evaluation and Research, administered by ORAU through the U.S. Department of Energy Oak Ridge Institute for Science and Education.

REFERENCES

Administration, U. F. &. D., 2018. *What We Do.* [Online]
Available at: https://www.fda.gov/about-fda/what-we-do
Administration, U. F. a. D., 2019. *FDA's Technology Modernization Action Plan.* [Online]
Available at: https://www.fda.gov/about-fda/reports/fdas-technology-modernization-action-plan
[Accessed 2020].
Administration, U. F. a. D., 2020. *fda.gov.* [Online]
Available at: https://www.fda.gov/drugs/types-applications/investigational-new-drug-ind-application#Introduction
Asuncion, H. U., 2011. In situ data provenance capture in spreadsheets. *IEEE Seventh International Conference on eScience,* Issue https://www.uwb.edu/getattachment/css/about/faculty/tech-reports/UWB-CSS-11-01.pdf.
Berndt, E. R. G. A. H. &. S. M. W., 2006. Opportunities for improving the drug development process: results from a survey of industry and the FDA.. *Innovation Policy and the Economy,* Volume 6, pp. 91-121.
Beyer, H. & Holtzblatt, K., 1998. *Contextual Design: Defining Customer-Centered Systems.* s.l.:Morgan Kaufmann Publishers.
Coble, J., Maffitt, J., Orland, M. & Kahn, M., 1995. Contextual Inquiry: Discovering Physicians' True Needs. *AMIA.*
Daae, J. &. B. C., 2015. A classification of user research methods for design for sustainable behaviour.. *Journal of Cleaner Production,* pp. 106, 680-689..
Fitzpatrick, R., 1999. Strategies for evaluating software usability. *School of Computing,* Volume 353.1.
Greg Guest, E. N. J. T. N. E. &. K. M., 2017. Comparing Focus Groups and Individual Interviews: Finding from a Randomized Study. *International Journal of Social Research Methodology* .

Guest , G. et al., 2017. Comparing Focus Groups and Individual Interviews: Findsing from a Randomized Study. *International Journal of Social Research Methodology*.

Kuprenas, J. A., 2003. Implementation and performance of a matrix organization structure. *International Journal of Project Management* .

Notess, M., 2005. *Understanding and Representing Learning Activity to Support Design: A Contextual Design Example*. Orlando, s.n.

Parenky, A. M. H. A. L. B.-P. K. R. A. K. S. &. Q. V., 2014. New FDA draft guidance on immunogenicity..

PRITCHARD, P. M. N. S. S. &. V. L., 2019. More Than A Robot: Designing for the Unique Advantages of Sending Humans to Mars. *Ethnographic Praxis in Industry Conference Proceedings.*

Schmuckler, M. A., 2001. What Is Ecological Validity?. In: s.l.: Lawrence Erlbaum Associates, Inc., p. 419–436.

Smithson, J., 2000. Using and analysing focus groups: limitations and possibilities. *International Journal of Social Research Methodology*.

Wixon, D. K. H. a. S. K., 1990. Contextual design: an emergent view of system design.. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

Appendix A: Affinity Diagram Example

Historical information retrieval

Past studies

- (sponsor) they have stop criteria for each cohort
- "product is off market" unable to reference during meeting
- "who is the patient population?"
- Similar to current drug but not really
- "most common reactions were…" CR
- "in European study adverse events were.." hard to compare with study findings
- "I was the reviewer on a similar drug"
- "how was the p/t to past studies"
- "what is the study protocol?" (New Zealand study)
- How to compare animal and human studies?
- Questions about precedence studies
- Chemistry communicates with pharmtox for justification
- "no data for pediatrics patients" CR
- "you guys want to see more [studies]" CSS
- Precedence to past drugs

Past meetings

- "we had a pre-IND meeting with them"
- "Who was PM from before?"
- Refers to 2018 pre-IND meeting
- Could not refer to commercial product during timeframe
- It is not clear how to find decisions from past meetings
- How will next group of reviewers know to look at long term recs
- History of important conversations is not documented
- "I thought we told them during a previous meeting"
- Refer to opioid guidance at Pre-IND

Documentation

Storage

- Email reviews
- Letter to deny in SharePoint
- Letter to proceed in SharePoint
- PM couldn't find document from day before
- Materials provided via SharePoint
- Opening topic specific word docs
- "you responded and said you put it in" "no, it was sent days ago"

Formatting

- Word format NCR
- OCP skipped most of his presentation for table
- Safety margins table brought up questions
- OCP word template w/ autobox
- Templates tweaked every 6-8 weeks
- All reviews created in word
- Used template for 7 page document
- First/last name in upper left corner OCP
- Sometimes the template changes
- "IND Quality" PDF format (chemistry)
- Conversation on non-hold meeting cancellation new rules

Presentation

Offline

- PM checks Outlook calendar event with a template
- Shares laptop via projector
- Checks Outlook calendar
- Everyone can see PM's personal email
- Absent chemist, PM reads remarks
- PM could not find CR review
- "I sent it two days ago" -CR

Real time

Remote

- "Let me bring up chemistry" –PM
- Scrolling too slow
- Controlling presentation for absent people

In-person

- Controlling presentation for absent people
- Cody asks PM to scroll down
- Remote caller on the line
- " I will not be calling in this morning" PM's email
- Chemistry presenter absent
- Shares laptop screen via Webex
- "Nobody else on Webex?" (PM attendance)
- Remote caller hard to understand
- 3 call in via Webex

Creating external deliverables

Future concerns (process related)

- "we do not want to put a hold on this study if we put a hold on another study"
- Future differences between IND approval + future NDA
- "can we refuse to file?" – future NDA
- What is "upscheduling" w/t drugs?
- "if non-clinical is ok with it, were ok with it"
- "we know what were looking for. We know we need a heart monitor"
- Long term recommendations are not documented "obviously we should monitor the dosing"

Expert concerns

- "can be abused?"
- "it's a puzzle" referring to mechanisms
- "this drug looks pretty clean"
- Hard to follow study design
- Rat/dog mg/kg bladder toxicity
- "noticed high doses"
- "migs per kig" mg/kg
- "I don't know how you'd monitor that"
- Pathologist reported "non adverse to dog"
- Brain and heart discussed not not stomach NCR
- "modest ability to penetrate brain"
- Refers to strop criteria for adverse events
- How is non-clinical risk assessed? "group strength"
- How to calculate safety margins from table
- "death is non-reversible"
- "I don't like this table"
- "put in control 2"
- "non-starter for table home" too dangerous to self
- "do we have the data?"
- Concerns about molecular weight
- "preferable to only deliver in clinic"
- Safety monitoring seems appropriate CR
- "letter should go out Dec 24"
- "Who is going to write non-hold"
- Identify one person to write non-hold

Task delegation

- "letter should go out Dec 24"
- "Who is going to write non-hold"
- Identify one person to write non-hold

No hold actions

- "did not have any hold comments" CR
- Collaborative non-hold comments
- Add non-hold comments to letter and let me know when you are done
- OCP non-hold comment, fill in P/T tables
- CR has 3 non-hold comments
- "regular request for them (sponsor) to fill out OCP tables"
- Add non-hold comments to doc in SharePoint
- "recommend that they exclude patients with heart issues and diabetes"
- Asks sponsor form more studies
- "let me know when you entered your non-hold comments"
- Non-hold comments after each topic
- P/T no "non hold comments"
- "go into SharePoint letter and populate it" templates

Blending Knowledge in Deep Recurrent Networks for Adverse Event Prediction at Hospital Discharge

Prithwish Chakraborty, PhD[1]; James Codella, PhD[1]; Piyush Madan, MS[1]; Ying Li, PhD[1]; Hu Huang, PhD[2]; Yoonyoung Park, PhD[1]; Chao Yan, MS[3]; Ziqi Zhang, BSc[3]; Cheng Gao, PhD[4]; Steve Nyemba, MS[4]; Xu Min, PhD[1]; Sanjib Basak, MS[2]; Mohamed Ghalwash, PhD[1]; Zach Shahn, PhD[1]; Parthasararathy Suryanarayanan, BSc B.Tech[1]; Italo Buleje, MS[1]; Shannon Harrer, PhD[2]; Sarah Miller, PhD[1]; Amol Rajmane, MS[2]; Colin Walsh, MD MA[3]; Jonathan Wanderer, MD MPhil[3]; Gigi Yuen Reed, PhD[2]; Kenney Ng, PhD[2]; Daby Sow, PhD[1]; Bradley A. Malin, PhD[3,4]
[1]IBM Research, USA; [2]IBM Watson Health, USA;
[3]Vanderbilt University, Nashville, TN, USA;
[4]Vanderbilt University Medical Center, Nashville, TN, USA

Abstract

Deep learning architectures have an extremely high-capacity for modeling complex data in a wide variety of domains. However, these architectures have been limited in their ability to support complex prediction problems using insurance claims data, such as readmission at 30 days, mainly due to data sparsity issue. Consequently, classical machine learning methods, especially those that embed domain knowledge in handcrafted features, are often on par with, and sometimes outperform, deep learning approaches. In this paper, we illustrate how the potential of deep learning can be achieved by blending domain knowledge within deep learning architectures to predict adverse events at hospital discharge, including readmissions. More specifically, we introduce a learning architecture that fuses a representation of patient data computed by a self-attention based recurrent neural network, with clinically relevant features. We conduct extensive experiments on a large claims dataset and show that the blended method outperforms the standard machine learning approaches.

Introduction

The digitization of health data has sparked artificial intelligence (AI) researchers to develop novel computational methods for various tasks, including the sequential prediction of clinically meaningful events, such as hospital readmissions and death. Initially, these methods were based on classical machine learning techniques, ranging from simple parametric regression to more complex non-parametric models, such as decision trees and rule learning techniques. Yet classical models are limited in their representative capacity and make certain assumptions about the data distribution (e.g., linearity assumptions) that do not always hold true. As such, deep learning architectures, including convolutional neural networks, recurrent neural networks, and, more recently, attention schemes have illustrated better performance potential for a variety of clinical prediction problems. We refer the reader to a recent review by Cao and colleagues[1] for a comprehensive survey on this topic.

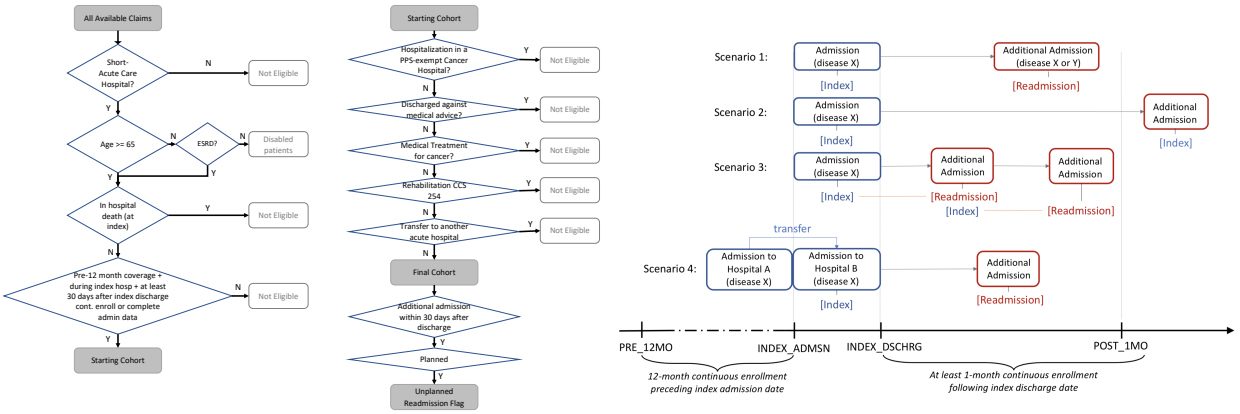
With respect to the readmission prediction problem, there have been several studies illustrating the potential for deep learning. First, a 30-day readmission prediction for patients with congestive heart failure (CHF) was achieved through the novel TopicRNN architecture, which combines global and local context[2]. Though the performance was modest (AUC in the $0.60 - 0.65$ range), this architecture outperformed state-of-the-art recurrent neural architectures, including RETAIN algorithm[3]. More recently, predictive performance was improved to over 0.70 AUC for CHF patients through a cost-sensitive formulation of a long short-term memory model that incorporates expert features and contextual embedding of clinical concepts[4].

Despite its potential, deep learning has been limited in its adoption by the healthcare community and specifically when applied to electronic health records (EHR) and claims data. As summarized by Wang and colleagues[5], while deep learning architectures have achieved success in the medical imaging domain[6,7], it has been more challenging to translate into solutions for EHR and claims data. The lack of translation stems from a number of factors, including those instigated by the data itself, such as high feature heterogeneity (e.g., a combination of discrete, continuous, and categorical features) and variability in quality, and those instigated by modeling (e.g., interpretability of learned

features and their weights). Moreover, many healthcare organizations are wary of the generalizability of the resulting models, a concern driven by challenges in semantic interoperability when these models are applied beyond the context in which they are initially trained. As an illustration of the challenge, it was recently shown that deep learning approaches were unable to outperform conventional methods in hospital readmission prediction[8]. More specifically, when only insurance claims data is available, Lasso regression appears to achieve readmission prediction performance that is comparable to deep neural network architectures[9,10].

As a consequence, though classical machine models may be suboptimal to their deep learning descendants, the former can perform well given the limited and sparse nature of medical data — especially when significant effort is applied to handcraft the features that represent domain-specific knowledge. These classical approaches also benefit from better alignment with current medical knowledge and, thus, are more readily interpretable to end users. Based on these observations, the AI community leadership has warned the community about the negative consequences that could transpire from overuse of deep approaches that tend to be less generalizable while being more difficult to interpret when more conventional rule learning methods could be applied[11].

In this paper, we illustrate that blending knowledge into deep learning frameworks can improve adverse event prediction at hospital discharge, including readmission and mortality. Specifically, we introduce a prediction method for such events from claims data that integrates 1) deep recurrent models, which capture complex temporal partners in patient data, with 2) knowledge-driven approaches, which capture medically relevant aspects of patient states. We demonstrate experimentally the benefits of this symbiosis between these data driven and knowledge driven modeling approaches. We applied this approach on a large medical claims data set, thus providing models that can be used by health organizations wishing to improve quality scores and reduce potential revenue losses resulting from these adverse events.



(a) Cohorts construction steps **(b)** Scenarios for target events

Figure 1: Cohort construction description: (a) shows the detailed steps for creating the index and target events and (b) shows the various scenarios to consider while resolving consecutive claim periods towards an index event.

Methods

Data and Study Setting

In this section, we start by describing our study setting. We primarily focused on prediction of unplanned readmission and mortality 30 days from discharge from a large scale claims dataset spanning millions of patients. Unplanned readmissions are undesirable events that can lead to increased healthcare costs and poorer health outcomes for patients. For this study, from the claims dataset we constructed the cohort by extracting the selected patients and identifying the prediction target events for 30-day unplanned readmission and also 30-day post discharge mortality. To do so, we first identify and define an *index event* from which we predict risk of each outcome as the discharge date from an inpatient admission. We begin with all claims for patients continuously enrolled for at least 12 months prior to an index event

Table 1: Descriptive statistics for the cohorts assembled for this study.

Final Cohort (2011) Total beneficiaries: 220,093	RACE	Unknown	White	Black	Other	Asian	Hispanic	North American Native	Total
	Counts	255	190,767	18,945	2,189	2,873	4,086	978	220,093
	Percentage	0.12%	86.68%	8.61%	0.99%	1.31%	1.86%	0.44%	
	Gender	Male	Female						Total
	Counts	91,389	128,704						220,093
	Percentage	41.52%	58.48%						
	Age Range	Unknown	<65	65∼69	70∼74	75∼79	80∼84	>85	Total
	Counts	7,602	0	39,843	43,374	42,735	42,154	44,385	220,093
	Percentage	3.45%	0.00%	18.10%	19.71%	19.42%	19.15%	20.17%	
30-day Readmission (2011) Total beneficiaries: 33,236	RACE	Unknown	White	Black	Other	Asian	Hispanic	North American Native	Total
	Counts	47	28,087	3,504	321	408	710	159	33,236
	Percentage	0.14%	84.51%	10.54%	0.97%	1.23%	2.14%	0.48%	
	Gender	Male	Female						Total
	Counts	14,225	19,011						33,236
	Percentage	42.80%	57.20%						
	Age Range	Unknown	<65	65∼69	70∼74	75∼79	80∼84	>85	Total
	Counts	1,268	0	5,390	6,253	6,520	6,764	7,041	33,236
	Percentage	3.82%	0.00%	16.22%	18.81%	19.62%	20.35%	21.18%	
Unplanned 30-day readmission (2011) Total beneficiaries: 18,329	RACE	Unknown	White	Black	Other	Asian	Hispanic	North American Native	Total
	Counts	28	15,200	2,148	190	261	416	86	18,329
	Percentage	0.15%	82.93%	11.72%	1.04%	1.42%	2.27%	0.47%	
	Gender	Male	Female						Total
	Counts	7,976	10,353						18,329
	Percentage	43.52%	56.48%						
	Age Range	Unknown <65	65∼69	70∼74	75∼79	80∼84	>85	Total	
	Counts	681	0	3,220	3,639	3,643	3,724	3,422	18,329
	Percentage	3.72%	0.00%	17.57%	19.85%	19.88%	20.32%	18.67%	

and for at least 30 days following discharge. Claims identified as index events must satisfy the following criteria:

• admissions must be for short stays requiring acute care.
• age of patient (except those who are eligible based due to end stage renal disease) at admission must be ≥ 65.
• patient must not expire during the inpatient stay. Note that these claims are not considered as index events but can be used to predict unexpected mortality if the previous index event was within 30 days of the death.
• patient must not have been transferred to another acute care hospital at discharge – in this case we examine the last claim in a series of one or more transfers as a possible index event.

These criteria are mainly driven by the guidelines outlined by Horowitz and colleagues.[12,13] for general cohort selection. After identifying the index events, we extract eligible *target events* (unplanned readmissions and mortality) within 30 days of each index event. Figure 1a illustrates the cohort construction process. Another important step pertains to resolving continuous periods of claims as a single claim. Figure 1b illustrates various scenarios that can arise when matching index events with readmissions. Scenario 1 shows a readmission that satisfies the 30-day readmission criteria whereas scenario 2, depicts a readmission outside the 30-day period and thus not considered to be a target event. Furthermore, we define a readmission in reference to an initial inpatient claim, such that only the first readmission within 30 days is considered as a target event. Scenario 3 shows a sequence of admissions where the second admission is considered to be a readmission for the first admission. The third admission is considered to be a readmission for the second admission only, in spite of occurring within 30 days of the first admission. It is to be noted, that all admissions including readmissions can qualify as index events if these meet the criteria for index event selection described earlier. Scenario 4 shows how we combine the transfers as a single event.

Defining the Target events: To define unplanned readmission, we first identified admissions that were considered planned. Any readmission, that was not planned or an acute event, was designated as unplanned. Planned readmissions are defined using a pre-specified list of procedure codes, or diagnostic codes for maintenance chemotherapy or rehabilitation. The list of codes were based on existing literature (See[13] Table 1 and Table 2). Even if planned proce-

dures occurred, readmissions for acute illness or for complications of care were not considered planned. The principal diagnostic code was used to identify such admissions. As a secondary target to evaluate our models, we defined unexpected 30-day mortality target. To find *unexpected mortality*, from each index claims, we identified patients who were dead within 30 days after discharge from an index admission based on the denominator table. We excluded patients who left the hospital against medical advice and died expired within 30 days of those discharge. We also excluded patients who expired after getting admitted to hospice. While the first criteria is standard and used by agencies such as the Centers for Medicare and Medicaid Services (CMS), the second criteria was included as patients in hospice care follows quite different patterns of care. Counts for the selected events are shown in Table 1.

Model Description

Claims data are typically multi-modal. Such data sets are also very high-dimensional and sparse. In addition, the historical data for each index admission is of varying temporal length. Our data set covers diagnosis, procedures, and demographic information.

In practice, methods such as LACE scores have been used to score the probability of unplanned readmissions. The LACE index[14] is computed using four dimensions to predict the risk of mortality or non-elective readmission 30 days after discharge. These variables are: length of stay (L), acuity of the admission (A), comorbidity of the patient (C), and emergency department visits (E) in the past 6 months from the index event. Using these, a 19 point scale is derived which can then be used to define the risk. While LACE provides a very interpretable and easy-to-use method to compute risks, it doesn't take into account the temporal information of patients from many different data sources.

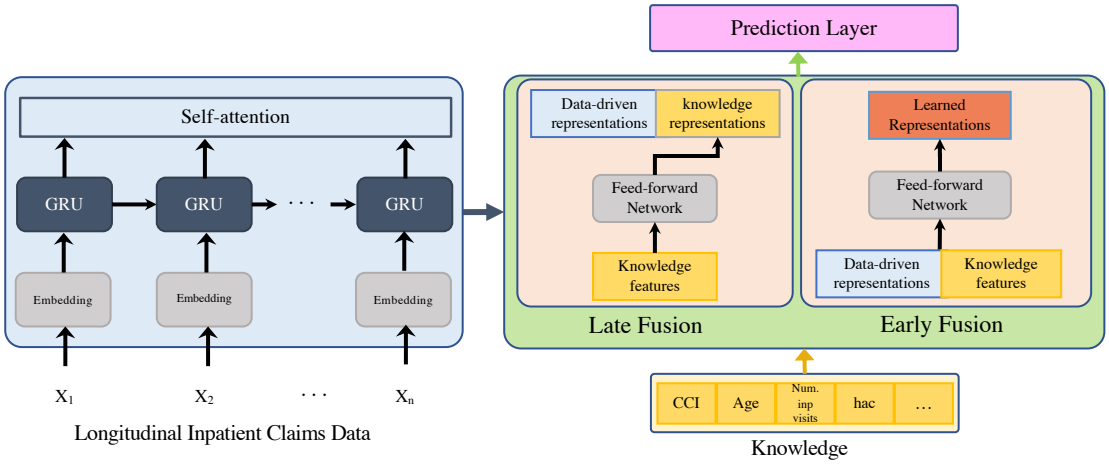


Figure 2: *Blending domain knowledge into Deep Recurrent Neural Networks*. Left (blue) box shows a self-attention based RNN that processes longitudinal inpatient claims data. The sparse data is embedded to a lower compact dimension using a Linear Embedding or Med2Vec layer. Healthcare domain-specific features (yellow) are next ingested into the model using either (a) early fusion or (b) late fusion strategy (Green box, right). In early fusion, domain-specific features are concatenated with representations from the RNN and the final model output is obtained using a deep feed-forward network. In late fusion, the domain features are transformed using a deep feed-forward network whose output is concatenated with representations from the RNN to a single shallow layer to generate model outputs.

RNNs to model longitudinal claims data: To effectively model the longitudinal claims data, we built a Recurrent Neural Network (RNN) model that ingests the temporal history of the patients and produces a risk score for the relevant target as its prediction. The architecture of the RNN is shown on the left side of Figure 2 (blue box) which we term as the data-driven model. The following is a list of three salient aspects of this RNN data-driven model.

• To account for the high dimensionality of patient records, we group together diagnosis and procedures using Clinical Classification Software (CCS) categories at their lowest level. Using this transformation we are left with more than 400 features at every time point which are one-hot encoded.

• Our dataset is highly sparse and high dimensional - to effectively model this dataset we apply an embedding layer to transform the raw one-hot encoded features at each time point to a compact representation.
• Finally, we feed the embedded vector to an RNN with self-attention to process the longitudinal temporal data.

There has been significant amount of work in literature in finding effective embedding strategies that are applicable to health data. Med2Vec[15] is one such method that is generally applicable. It fits a shallow network that aims to discover vector representations that are consistent between visits, while accounting for the apparent lack of order between features in a single visit. As an alternative mechanism we also fit a 1-layer feed-forward network to produce compact representations of the input data. While we experimented with different RNN architectures, we settled on a Gated Recurrent Unit (GRU) as our base RNN based on its performance. While such models have been successfully applied to computational health modeling scenarios[1,3], it has recently been observed that adding an attention layer can significantly improve the performance of such deep networks.

Formally, let us denote the medical history of the n^{th} patient (where $n \in 1, \ldots, N$) by $X^n = \{X_1^n, X_2^n, \cdots, X_{T^n}^n\}$, where T^n is the total number of observed time points for patient n and $X_t^n \in \mathcal{R}^M$ represents the observed features. For the sake of simplicity, we drop the superscript n in the rest of this paper. Input data X_t at each time point t is ingested by this RNN architecture (see Figure 2), starting with several layers, to produce an embedding vector e_t using either Med2Vec or a feed-forward network as $e_t = W_e e_t + b_e$. This information is combined with information from historical data points represented as a hidden vector h_t as $h_{t+1} = \text{GRU}(e_t, h_t)$. Finally, we use self-attention[16] to combine all the hidden states $H = \{h_0, h_1, ..., h_T\}$ and produce the final output y_t as the prediction for the relevant problem (readmission/mortality) as:

$$a_t = \text{Softmax}(\frac{h_T H}{\sqrt{|h_T|}})H; \qquad o_t = a_t \cdot H \qquad (1) \qquad\qquad y_t = \text{Sigmoid}\left(W_o o_T + b_o\right) \qquad (2)$$

We performed experiments that included, as well as neglected, outpatient claims. We observed that such claims had minimal influence on model performance and were excluded from further investigation.

Availability of domain knowledge: Deep learning models, such as the one described above, are non-parametric and have a high capacity to model various complicated concepts in observational data. However, in limited and sparse data settings, such as that encountered in medical data, classical methods can perform well - especially when significant effort has been applied to handcraft the features. These features are quite beneficial to end users consuming the outputs of such predictions as they are readily interpretable and transparent. In fact, it was recently reported that deep learning approaches are unable to outperform such conventional methods to predict hospital readmissions[8]. It should be noted that in well studied problems, such as readmission prediction, there is a significant literature about how to construct such features[12,13]. As such, we hypothesize that we can improve the predictive performance of deep learning models - even under such data sparsity situations - by blending such features into deep learning models. To this effect, we constructed several features from domain knowledge as listed in Table 2a. These features were identified from literature and verified by subject matter experts. They are categorized along four groups: (a) Comorbidity, (b) Clinical, (c) Demographic, and (d) Others. The latter category tracks facility and discharge disposition information that we used to derive discharge related actionable insights. A special class of the features constructed from Clinical data elements are *hospital acquired conditions* (HAC) or *presence of hospital acquired complications during index admission*. HAC are undesirable events and can be indicative of future complications. The full list of HAC is shown in Table 2b which we constructed by leveraging the HAC as published by CMS.

Blending domain knowledge in an RNN: To effectively use such domain knowledge directly in our deep learning architecture we fuse this knowledge with the RNN output as shown on the right side of Figure 2 in the green box. We refer to this as the *Fusion* layer and design two strategies to blend this information: (a) Early fusion and (b) Late fusion. In early fusion, we concatenate the domain features z and learned data representations o_T from Equation 1 and feed into a multi-layered feed forward network. By contrast, in late fusion we learn a representation of the domain knowledge by feeding it into a multi-layered feed-forward network and concatenate the data and knowledge representations. Finally, the output from each respective strategy is fed into an output prediction layers (Figure 2, pink box) as described in equation 2. Formally this can be defined as follows:

$$\textbf{Early Fusion:} \quad \hat{o}_T = \text{MLP}([o_T; z]) \qquad \textbf{Late Fusion:} \quad \hat{o}_T = [o_T; \text{MLP}(z)] \qquad (3)$$

Table 2: Features used: List of knowledge driven features (left) list of hospital acquired conditions (right)

(a) Knowledge derived Features

Category	Definition
Comorbidity	DX & PROC Categorization
	Charlson Comorbidity Index
Clinical	Length of Stay (index admission)
	Number of Inpatient Admissions during Previous 12 Months
	Number of Outpatient Visits during Previous 12 Months
	Number of ED Visits during Previous 12 Months
	Type of Index Admission
	Admission Source
	Discharge Disposition
	Discharge Diagnosis
	List of Hospital Acquired Complications During Index Admission
	Diagnosis Related Group
	Number of DX Codes on a Claim
Demographic	Age Group
	Gender
	Race Codes
	Dual Eligibility
	Reason for Medicare Eligibility
Others	Facility ID

(b) Hospital acquired conditions

Hospital Acquired Complications
Foreign Object Retained After Surgery
Air Embolism
Blood Incompatibility
Pressure Ulcer Stages III & IV
Falls and Trauma
Catheter-Associated Urinary Tract Infection (UTI)
Vascular Catheter-Associated Infection
Manifestations of Poor Glycemic Control
Surgical Site Infection, Mediastinitis, Following Coronary Artery Bypass Graft (CABG)
Surgical Site Infection Following Certain Orthopedic Procedures
Surgical Site Infection Following Bariatric Surgery for Obesity
Deep Vein Thrombosis and Pulmonary Embolism Following Certain Orthopedic Procedure

Table 3: Comparison of model performance (w/ linear embedding vs. med2vec) for prediction of two adverse events at hospital discharge (a) unplanned readmission and (b) unexpected mortality. Uncertainty based on top 10 models.

(a) Readmission

Algorithm	With Linear Emb		With Med2Vec	
	AUC	**Recall**	**AUC**	**Recall**
LR	0.628 ($\pm$0.004)	0.647	0.629 ($\pm$0.003)	0.554
RF	0.600 ($\pm$0.008)	0.712	0.478 ($\pm$0.001)	0.864
HistGBT	0.558 ($\pm$0.003)	0.218	0.442 ($\pm$0.002)	0.826
Early Fusion	0.678 ($\pm$0.004)	0.549	0.654 ($\pm$0.005)	0.54
Late Fusion	0.676 ($\pm$0.005)	0.51	0.652 ($\pm$0.006)	0.496

(b) Unexpected mortality

Algorithm	With Linear Emb		With Med2Vec	
	AUC	**Recall**	**AUC**	**Recall**
LR	0.800 ($\pm$0.006)	0.7839	0.774 ($\pm$0.004)	0.784
RF	0.785 ($\pm$0.004)	0.119	0.492 ($\pm$0.003)	0.697
HistGBT	0.766 ($\pm$0.003)	0.095	0.471 ($\pm$0.001)	0.557
Early Fusion	0.840 ($\pm$0.008)	0.811	0.800 ($\pm$0.004)	0.687
Late Fusion	0.844 ($\pm$0.005)	0.729	0.800 ($\pm$0.001)	0.711

Baselines: We evaluated our models against 30-day readmission prediction and, another closely related task, 30-day unexpected mortality. We split the dataset, stratified by number of events (e.g., readmission episodes) for a patient, for each of the problems randomly into training (70%), validation (15%), calibration (5%), and test folds (10%). The stratification ensured that the complete history of a patient belonged to only one data subset. We evaluated both classical models (Logistic Regression, Random Forest, and Histogram-based Gradient Boosting Tree) as well as our proposed methods, henceforth referred to as 'Early' and 'Late' fusion. As the dataset is highly sparse, we compared models under both linear and med2vec embedding strategies. The tasks are also highly unbalanced (See Table 1) such that failing to predict a true positive event may have a higher burden than over-predicting. To account for this aspect, we used state-of-the-art techniques such as SMOTE[17]for classical models. For deep models, we used a weighted loss between positives and negatives allowing us to penalize false negatives higher than false positives. While this is a standard setting for unbalanced data problems, for clinical tasks such as readmission predictions the final score from a predictive model is usually desirable to correlate with the probability of the event (so that these can be used as risk scores). Thus, once the models were trained we applied *Platt Scaling* for classical models and *temperature scaling* for deep learning models on the calibration set to calibrate the score estimates from the models - this ensures that these scores have a probabilistic interpretation and 0.5 can be chosen as the threshold to declare presence/absence. Our choices were motivated by existing literature[18] that found these choices to work well in practice.

Results

The deep learning models were implemented using 'pytorch', the classical models using 'scikit-learn', and were run on a cluster containing eight NVIDIA V100 GPUs. For each model, hyperparameter optimization was conducted using a grid search over a predefined grid and the hyperparameters were selected based on the model with the best performance on validation sets. All models were able to chose from the same set of features and the grid for the

hyperparameters were individually tuned for each model following best recommendations. For example, for our models we mainly chose between hidden size $(8 - 128)$, number of layers $(1 - 3)$, and batch size $(8 - 128)$. For classical models, we also generated standard summary of temporal features such as mean and counts of features. We report the comparison of model performance on the 10% hold-out test fold for readmission prediction in Table 3a. We report both area under the receiver operating characteristic curve (AUC) as a general metric and recall to evaluate the false-negative rate. Uncertainty around AUC based on the top 10 best fitted models is also reported. To ensure that model outputs can be mapped to probabilities, we always chose 0.5 as the threshold to calculate recall. We conducted initial experiments on readmission prediction problem where we compared a RNN only model (without any fusion) to the baselines. Our results were similar to [8] where RNN models failed to outperform the baselines. Overall, for 30-day readmission prediction, the proposed deep models performed at acceptable (or slightly better) AUC values, ranging from 0.652 to 0.682, compared to reported numbers from literature[8,19] on general readmission (not necessarily unplanned). We also notice that there was a performance gain with the deep learning models over classical machine learning models. For example, for the 30-day readmission prediction problem, the best performing deep learning model *Early Fusion* achieved an AUC of 0.682 and recall of 0.549 while the best performing conventional machine learning model is logistic regression with an AUC of 0.632 and recall of 0.647. *It should be noted, that our proposed models were successfully tuned to achieve very high Recall-at-top-k (close to 1) at the expense of lower precision*. We also observed that the deep learning models can be tuned more-readily to balance the trade-off between these metrics dependent on the performance indicators of interest such as AUC and Recall. For this study, we opted for models above an acceptable AUC threshold (e.g., 0.6 for 30-day readmission prediction) and higher recall to account for the rarity of the tasks.

As a secondary study, Table 3b reports on the same performance measures for 30-day unexpected mortality. The proposed models perform relatively well for this task, with the best AUC varying between 0.806 to 0.849 and recall ranging from 0.697 to 0.811 for the best models. It can be seen the deep learning models perform the best for this task as well with *Early Fusion* and *Late Fusion* achieving the top results.

Discussion

What did we learn from model selection - Does med2vec help Early or Late Fusion? We trained a suite of models and optimized for the hyperparameters to find the best validated version of each model. We performed a *quasi-ablation* study by selectively including and excluding different features such as outpatient claims and domain knowledge. Based on our experiments, it was found that the outpatient historical claims did not improve model performance when incorporating both inpatient claims and domain features. Also, we found that in general, the use of med2vec as a preprocessing technique did not improve the deep learning models but helped the classical models to attain better recall sometimes. We also found that 'Early Fusion' in general led to better performance over multiple problem setup than *Late Fusion*. This indicates the importance of learning non-trivial interaction components between domain knowledge and learned representations from longitudinal medical history of the patients. It should be noted that our modeling scenario is more restrictive in the sense that there is no overlap of patients between the different folds (such as in the training and in the test data sets).

What are the key drivers - is knowledge important? We evaluated the importance of knowledge by ascertaining their importance with respect to raw data features at a model level. However, due to the black-box nature of deep models, obtaining global importance is a non-trivial task. We obtained this in a post-hoc manner by conducting a study to identify globally important features using a logistic regression model (self-interpretable) as an explainer model. Tables 4a and 4b show the most important features identified at a global level for 30-day unplanned readmission prediction and 30 day unexpected mortality, respectively. It can be seen that many data-driven features, spanning both diagnosis and procedure codes, are identified as important. This highlights how critical it is to model the longitudinal medical history for a patient. However, we also found that domain knowledge features, such as *Charlson comorbidity* (for both readmission and mortality) and *Length of stay* (for mortality), are important measures. This highlights the importance of blending domain knowledge into our deep models.

Where does the model perform well - Personalized Model Performance Performance of clinical risk scores is typically reported at an aggregate level for an entire study population. However, patients in the population may vary significantly from one another (e.g., patients with vs. without cardiovascular disease; males vs. female; patients in

Table 4: Global Feature importance

(a) Readmission Prediction

Cateogry	CCS CATEGORY	Importance
PROC	Hemodialysis	2.80549
ICD9	Deficiency and other anemia	1.98106
PROC	Blood transfusion	1.66835
PROC	Arthroplasty knee	-1.54134
ICD9	Acute cerebrovascular disease	-1.52745
ICD9	Secondary malignancies	-1.48642
ICD9	Transient cerebral ischemia	-1.48194
ICD9	Aortic; peripheral; and visceral artery aneurysms	-1.34200
ICD9	Other and ill-defined cerebrovascular disease	-1.21932
PROC	Insertion; revision; replacement; removal of cardiac pacemaker or cardioverter/defibrillator	-1.18553
PROC	Endarterectomy; vessel of head and neck	-1.13404
PROC	Computerized axial tomography (CT) scan head	1.11995
ICD9	Osteoarthritis	-1.11774
ICD9	Other non-epithelial cancer of skin	-1.06222
Domain	*Charlson index*	1.05767
PROC	Laminectomy; excision intervertebral disc	1.05620
ICD9	Other ear and sense organ disorders	-1.04940
PROC	Diagnostic cardiac catheterization; coronary arteriography	-1.03238
PROC	Electrographic cardiac monitoring	1.02820
ICD9	Malaise and fatigue	-1.01513

(b) Mortality Prediction

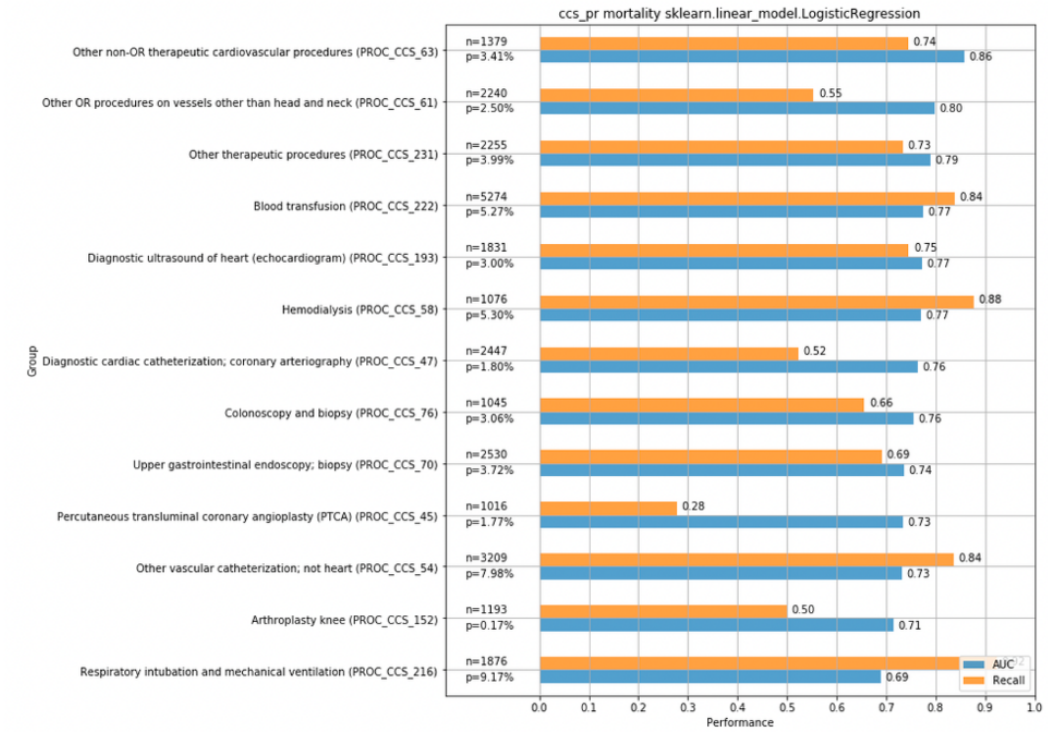
Category	CCS Category	Importance
ICD9	Delirium, dementia, and amnestic and other cognitive	0.046376
ICD9	Congestive heart failure; nonhypertensive	0.040182
ICD9	Respiratory failure; insufficiency; arrest	0.034240
ICD9	Residual codes; unclassified	0.033168
Domain	Charlson Index	0.031612
ICD9	Acute and unspecified renal failure	0.030669
ICD9	Deficiency and other anemia	0.027499
ICD9	Cardiac dysrhythmias	0.027440
ICD9	Disorders of lipid metabolism	0.026994
ICD9	Pneumonia (except that caused by tuberculosis)	0.026305
ICD9	Essential hypertension	0.025222
Domain	*Length of stay*	0.022936
ICD9	Nutritional deficiencies	0.022277
ICD9	Urinary tract infections	0.021677
ICD9	Chronic obstructive pulmonary disease and bronchiectasis	0.021063
ICD9	Other gastrointestinal disorders	0.018653
ICD9	Other aftercare	0.017234
ICD9	Coronary atherosclerosis and other heart disease	0.015617
ICD9	Septicemia (except in labor)	0.015360

different age ranges. In this respect, it is critical to characterize the performance of the risk score across patient subpopulations, so that clinicians can understand when a risk score is expected to be the most applicable to an individual patient. More specifically, knowing the operating ranges of a model can lead to more confident use of AI models. To support this need, we characterized the performance across predefined sub-cohorts of patients. In a post hoc analysis on a hold-out evaluation test set that was scored with the trained risk prediction model, we grouped patients based on a variety of baseline characteristics and recomputed prediction performance measures on the patient cohort.

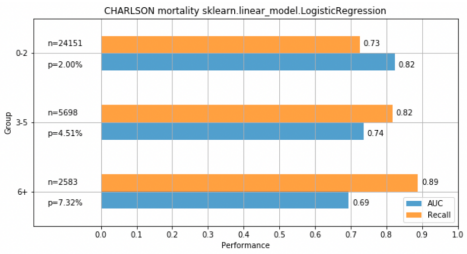
The investigated subgroups consisted of age, gender, race, CCS procedure categories, and Charlson index. We report 3 of the most interesting analysis in Figure 3. Figure 3a shows model performance for population subgroups based on CCS procedure categories, for 30 day mortality prediction. There is variability in performance across groups. The performance is significantly higher than the population average (AUC = 0.80) for patients with *Other non-OR therapeutic cardiovascular procedures (CCS 63)* (AUC = 0.86). It is much worse for patients that had *Percutaneous transluminal coronary angioplasty (PTCA) (CCS 45)* (AUC = 0.73), *Other vascular catheterization; not heart (CCS 54)* (AUC = 0.73), *Arthroplasty knee (CCS 152)* (AUC = 0.71), and *Respiratory intubation and mechanical ventilation (CCS 216)* (AUC = 0.69). This indicates that, while the model performs well at a population level, for patients with history of procedures (e.g, *Arithroplasty of Knee*), the model should be used with caution. Figure 3b illustrates a non-trivial variability in performance for patient subgroups based on their Charlson index. Patient groups with an index between 0-2 exhibit a slightly better performance (AUC = 0.82) than the population average (AUC = 0.80), while patient groups with higher indices, 3-5 and 6+ exhibit worse performance (AUC = 0.74 and 0.69, respectively). This indicates that at these Charlson Index ranges, the model is less certain. Figure 3c depicts the performance for patient subgroups based on their *CWF BENE MDCR* status. Almost all patients are in the "Aged without ESRD" group so the mortality prevalence (2.9%) and model performance (AUC = 0.81) are roughly the same as the population average (p = 2.9%, AUC = 0.80). Only 980 patients are in the *Aged with ESRD* group, which exhibits a higher mortality prevalence (3.2%) but substantially worse performance (AUC = 0.70).

Conclusion

Deep learning models can suffer in performance when data is highly sparse and irregular, as is commonly the case for observational patient data. When sufficient domain knowledge is available, classical models using hand-crafted features can supplement deep models. We proposed a novel modeling framework that enjoy the best of both worlds, i.e.,



(a) CCS procedure categories



(b) Charlson index

(c) CWF BENE MDCR status

Figure 3: Performance (AUROC and Recall) for patient subgroups based on (a) CCS procedures, (b) Charlson Index, and (c) medicare enrollment reasons. We see variations of model performance across sub-categories in (a-c).

the benefits of non-parametric and flexible sequential modeling of patient history, via deep learning, while receiving guidance from domain knowledge. Our empirical investigation, using large insurance claims data illustrates that such a method works efficiently across two tasks at discharge time: unplanned readmission and unexpected mortality. We also analyzed our model performance to determine conditions when (or when not) such models are confident in their predictions. Possible extensions of this work may include applying this approach on other data modalities (e.g., EHR data) to enable the assessment of readmission and risks prior to discharge.

References

1. Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 06 2018.

2. Cao Xiao, Tengfei Ma, Adji B Dieng, David M Blei, and Fei Wang. Readmission prediction via deep contextual embedding of clinical concepts. *PloS one*, 13(4):e0195024, 2018.

3. Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3504–3512. Curran Associates, Inc., 2016.

4. Awais Ashfaq, Anita Sant'Anna, Markus Lingman, and Sławomir Nowaczyk. Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, 97:103256, 2019.

5. Fei Wang, Lawrence Peter Casalino, and Dhruv Khullar. Deep learning in medicine—promise, progress, and challenges. *JAMA internal medicine*, 179(3):293–294, 2019.

6. Varun Gulshan, Lily Peng, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, 12 2016.

7. Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

8. Xu Min, Bin Yu, and Fei Wang. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on copd. *Scientific reports*, 9(1):1–10, 2019.

9. Tadahiro Goto, Taisuke Jo, Hiroki Matsui, Kiyohide Fushimi, Hiroyuki Hayashi, and Hideo Yasunaga. Machine learning-based prediction models for 30-day readmission after hospitalization for chronic obstructive pulmonary disease. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 16(5-6):338–343, 2019. PMID: 31709851.

10. Ahmed Allam, Mate Levente Nagy, George R. Thoma, and Michael Krauthammer. Neural networks versus logistic regression for 30 days all-cause readmission prediction. *Scientific Reports*, 9, 2019.

11. Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

12. LI Horwitz, C Partovian, Z Lin, JN Grady, J Herrin, M Conover, J Montague, C Dillaway, K Bartczak, LG Suter, JS Ross, SM Bernheim, HM Krumholz, and EE. Drye. Development and use of an administrative claims measure for profiling hospital-wide performance on 30-day unplanned readmission. *Ann Intern Med.*, 2014.

13. Leora Horwitz and et. all. Hospital-wide readmission measure methodology report, 2012.

14. Carl van Walraven, Irfan A Dhalla, Chaim Bell, Edward Etchells, Ian G Stiell, Kelly Zarnke, Peter C Austin, and Alan J Forster. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj*, 182(6):551–557, 2010.

15. Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1495–1504, New York, NY, USA, 2016. Association for Computing Machinery.

16. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

17. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

18. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

19. Daniel J Morgan, Bill Bame, Paul Zimand, Patrick Dooley, Kerri A Thom, Anthony D Harris, Soren Bentzen, Walt Ettinger, Stacy D Garrett-Ray, J Kathleen Tracy, et al. Assessment of machine learning vs standard prediction rules for predicting hospital readmissions. *JAMA network open*, 2(3):e190348–e190348, 2019.

Phenoflow: A Microservice Architecture for Portable Workflow-based Phenotype Definitions

Martin Chapman[1], Luke V. Rasmussen[2], Jennifer A. Pacheco[2], Vasa Curcin[1]
[1]King's College London, London, United Kingdom; [2]Northwestern University, Chicago, Illinois, USA

Abstract

Phenotyping is an effective way to identify cohorts of patients with particular characteristics within a population. In order to enhance the portability of a phenotype definition across institutions, it is often defined abstractly, with implementers expected to realise the phenotype computationally before executing it against a dataset. However, unclear definitions, with little information about how best to implement the definition in practice, hinder this process. To address this issue, we propose a new multi-layer, workflow-based model for defining phenotypes, and a novel authoring architecture, Phenoflow, that supports the development of these structured definitions and their realisation as computable phenotypes. To evaluate our model, we determine its impact on the portability of both code-based (COVID-19) and logic-based (diabetes) definitions, in the context of key datasets, including 26,406 patients at Northwestern University. Our approach is shown to ensure the portability of phenotype definitions and thus contributes to the transparency of resulting studies.

Introduction

Learning Health Systems require high-quality, routinely collected electronic health record (EHR) data to drive analytics and research, and translate the outputs of novel techniques such as machine learning into patient care and service improvement. To achieve this, the data used for research need not only be of high-quality, but methods associated with its use need be transparent and reproducible to ensure that any findings can be validated by the research community and generalised to other populations. At the core of this challenge is the ability to reliably identify clinically equivalent research-grade patient cohorts. These cohorts are precise enough to conduct meaningful research by identifying individuals with a particular disease, sets of comorbidities, medical histories, a demographic profile or any other relevant patient-specific information – a process known as EHR-based phenotyping[1].

The popularity of EHR data for research has increased the documentation and sharing of phenotypes derived from research datasets in order to stimulate reuse, reduce variation in phenotype definitions across data sources, and ultimately simplify and support the identification of clinically equivalent populations for research and healthcare applications. The reuse of existing phenotype definitions necessitates the ability to discover and access curated and validated phenotype definitions. Pioneering efforts in building standardised phenotype repositories, such as the Phenotype Knowledge Base (PheKB), CALIBER, Million Veterans Program (MVP) and *All of Us* consortium have achieved notable success within their research programmes, with thousands of registered usages[2,3].

In an attempt to ensure the portability of a phenotype across multiple research use cases, the logic that comprises a phenotype definition is often represented abstractly within these repositories, where it is structured as, for example, a list of codes (e.g. Figure 1), or as a data flow diagram (e.g. Figure 2). This abstract representation is designed to guide the development of a computable form of the phenotype, such as an executable script or a data pipeline, for a particular use case. However, in practice, the portability of these definitions is often low: a lack of clarity in the abstract definition, either in terms of terminology or structure, make them hard to interpret in order to produce a computable form, and the technical skill burden on the computable phenotype author is high, as the abstract nature of each definition means that little is communicated about the realisation of each phenotype in practice.

Methods

In order to address the difficulty of deriving computable forms from phenotype definitions, we propose a novel phenotype definition model, which aims to increase portability by improving clarity, and more explicitly defining the structure of computable forms. The formulation of the proposed model was based on the experiences of initiatives such as the UK eScience and US Cyberinfrastructure programmes[4], which developed *scientific workflow* models for orchestrating and coordinating their computational tasks. In addition, the functional (re-)modelling of different phenomena in a number of different domains was used as a basis for the proposed model – in particular, work in *hierarchical modelling*; the representation of a phenomena at different levels of abstraction[5], e.g. in bioinformatics software

architectures[6]. Finally, the authors themselves have developed a number of different models as part of prior studies, including work on complex systems[7], semantics for hierarchical composition[8] and phenotyping from large scale EHR repositories[9].

Vocabulary	Code	Term
ICD-10	U07.1	Diagnosis of COVID-19 confirmed by laboratory testing
...	...	...
ICD-11	RA01.0	Diagnosis of COVID-19 confirmed by laboratory testing
...	...	...
SNOMED-CT	840539006	COVID-19
...	...	...

Figure 1: Extract of code lists used for defining COVID-19 patients in an EHR system (Source: http://covid19-phenomics.org/).

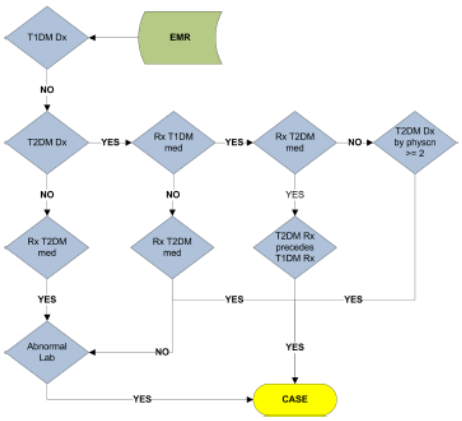


Figure 2: Data flow for defining T2DM patients in an EHR system (Source: https://phekb.org/phenotype/type-2-diabetes-mellitus).

In order to evaluate how our new *structured definition* model impacts portability, we first collected a set of 278 existing phenotype definitions from a number of different phenotype repositories, including PheKB (https://phekb.org) and CALIBER (e.g. https://portal.caliberresearch.org). We then re-authored these definitions according to our model, and used them to produce corresponding computable forms. In examining these definitions, we identified that they fall broadly into two categories: *code-based* definitions that identify patient cohorts using a list of clinical codes, and *logic-based* definitions that identify patients using a series of logical statements. To evaluate the impact of our model on the portability of the phenotype definitions in each of these categories, we selected a representative phenotype from each category, including a code-based Coronavirus disease 2019 (COVID-19) definition (Figure 1) and a logic-based Type 2 Diabetes Mellitus (T2DM) definition (Figure 2), and compared the portability of each re-authored definition with the original, using the *Knowledge conversion, clause Interpretation, and Programming* (KIP) phenotype portability scoring system[10].

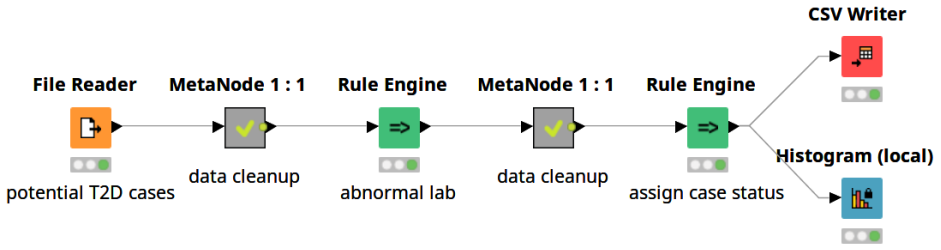


Figure 3: Original T2DM phenotype, implemented as the nodes of a KNIME pipeline.

Prior to applying the KIP scoring system, it was important to verify that our re-authoring approach resulted in structured definitions that still captured the required phenotype logic. To do this we executed the computable forms derived from the original definitions of our representative phenotypes and the computable forms derived from the structured definitions of these phenotypes against a given patient dataset, and verified that the same patients were identified by both implementations. For example, as a part of our evaluation, the phenotype definition for COVID-19 was obtained from CALIBER (http://covid19-phenomics.org), re-authored by one of the authors (MC), and a computable form produced from both the original definition (as one did not exist) and the structured definition. Similarly, the definition

for T2DM, and its corresponding Konstanz Information Miner (KNIME) pipeline implementation (Figure 3), were obtained from PheKB (https://phekb.org/phenotype/type-2-diabetes-mellitus). This definition, like the definition for COVID-19, was then re-authored by one of the authors (MC), and a new computable form was produced. The original COVID-19 implementation and the new computable form were then executed against a cohort of 1468 individuals who tested positive for COVID-19 at Guy's and St. Thomas' NHS Foundation Trust (GSTT), London, while the T2DM implementations were executed against a cohort of 26,406 possible T2DM patients, taken from the Northwestern Medicine Enterprise Data Warehouse (NMEDW), as well as against publicly available data from PheKB. In the case of T2DM, the execution of the computable form derived from the structured definition against these datasets is possible because it uses the same data input format as the original KNIME pipeline implementation, and similarly creates an output file in the same structure. The GSTT dataset included a subset ($n = 1153$ cases) of hospitalised COVID-19 patients, while the NMEDW dataset included a subset ($n = 23$ cases) of patients with T2DM that had previously undergone manual chart review, both of which acted as the gold standard to validate our algorithm against. In both cases, the results of executing the structured implementation were compared with the results of executing the original implementations to confirm the same exact cases and controls were found across their respective datasets.

Structured Phenotype Definition

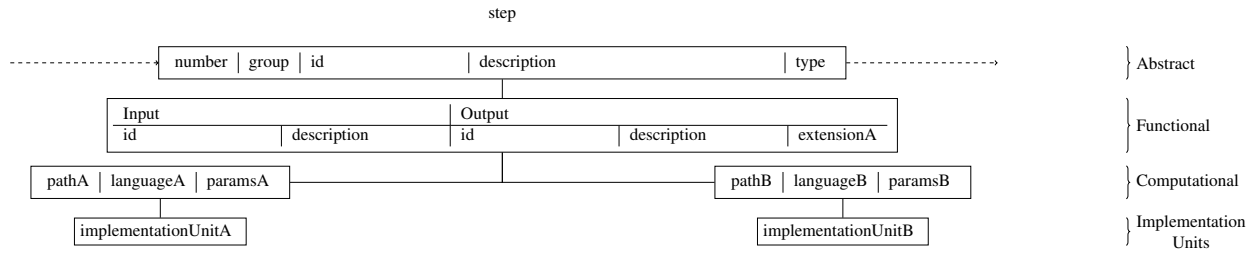


Figure 4: Structured phenotype definition model (step) and implementation units.

The structured phenotype definition model developed consists of a set of *layers*: abstract, functional and computational. A graphical overview of our model is given in Figure 4. Like traditional definitions, the **abstract layer** of a structured phenotype definition holds the logic of the phenotype. However, the abstract layer in our model is defined by two distinct features. Firstly, like the workflow models upon which our model is based, this layer consists of a number of sequential steps, each of which defines a single operation against a target dataset. However, steps may also be grouped, allowing for their functionality to be summarised by a single parent step. The second feature of this layer is a multi-dimensional description of each step, which consists of an ID, designed to summarise the purpose of the step using relevant clinical terminology; a longer description of the step, designed to offer a non-technical description of the logic of the step; and a categorisation of the logic of the step as an entity within a given concept ontology (broadly based on the axioms of the Phenotype Execution and Modelling Architecture (PhEMA) authoring tool (PhAT)[11]): *load* (loading data from a datasource), *logic* (generic logic to identify patients), *boolean* (boolean logic to identify patients) and *output* (the output of patients that exhibit a given phenotype to a specified location, e.g. disk). As a part of our model, we also constrain the type of logic that a step may have, depending upon its position. For example, the first step may only be of a *load* type, while the last may only be of an *output* type.

The **functional layer** of a structured phenotype definition is used to augment the information held in the abstract layer by adding metadata information about the inputs and outputs of each step. This explicit specification is an application of a functional programming paradigm[12], and includes identifiers for both the inputs and outputs, summarising their purpose using relevant clinical terminology; a longer description of both the inputs and output to each step, designed to offer a non-technical insight; and syntactical commitments for step output (e.g. file type).

Finally, the modular **computational layer** of a structured phenotype definition is used to describe the presence of one or more implementation units (e.g. a script, data pipeline module, etc.) for each (nested) step in the abstract (and functional) layers. This description includes information about the execution environment used to run the implementation unit, and how the unit is linked to that environment.

Generating computable phenotypes

In order to assist in the development of a computable phenotype from a structured phenotype definition, we develop a microservice architecture, *Phenoflow* (Figure 5). Software designed as a microservice architecture provides functionality based on the interactions between individual services. As a specialised type of service-oriented architecture, the microservice approach dictates that each service should only deliver one specific piece of system functionality, making it easier to achieve quality attributes such as scalability and resilience in practice[13]. This paradigm has been successfully used to structure software in several health domains, including the representation of clinical guidelines[14].

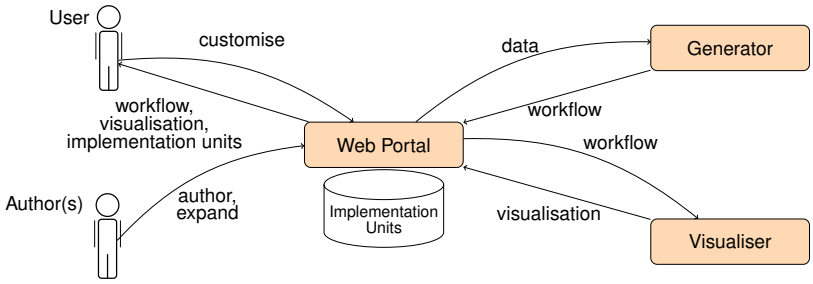


Figure 5: The microservices (web portal, generator and visualiser) that constitute the *Phenoflow* architecture.

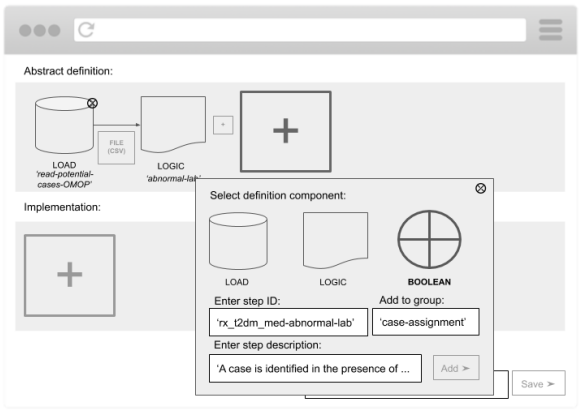


Figure 6: Visually defining the abstract and functional layers of a definition.

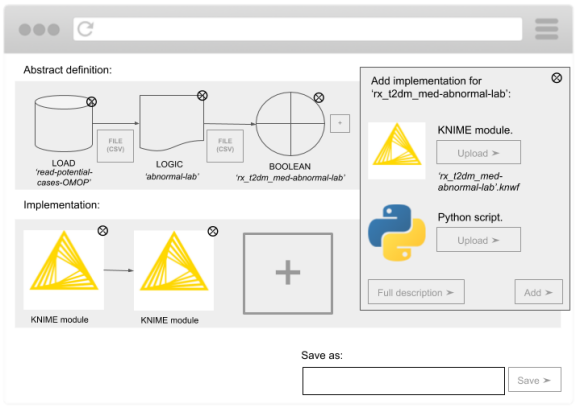


Figure 7: Providing an implementation unit for the third step in the abstract and functional layers, in order to generate a step in the computational layer and store this unit.

Phenoflow first allows a researcher to graphically author a definition through a web portal service, where they express each step of their new definition at the abstract and functional layers by selecting the type of each step, and then by labelling, describing and, if they wish to, grouping those steps, as well as describing their inputs and outputs. This process is represented in Figure 6, where a user is in the process of defining a new boolean expression, within their abstract layer, having already defined an initial data read from an Observational Medical Outcomes Partnership (OMOP) common data model (CDM) database, another piece of logic, and the fact that a CSV file is passed between these two steps. Authors can also (indirectly) reuse the definitions produced by others in the definition of their own phenotypes. This is common in the development of new definitions that contain the same logic, but target a different data source. For example, two definitions that differ only in *load* components that separately target the OMOP CDM and an Informatics for Integrating Biology and the Bedside (i2b2) dataset.

Following the specification of one or more steps in the abstract and functional layers, the researcher graphically connects each step to an implementation unit (e.g., a Python script, or a KNIME module; their choice across the steps does not have to be homogeneous), which they supply to the portal, in order to generate the computational layer. This process is represented in Figure 7, where the author is uploading the module of KNIME pipeline as the implementation counterpart of their priorly defined boolean expression step.

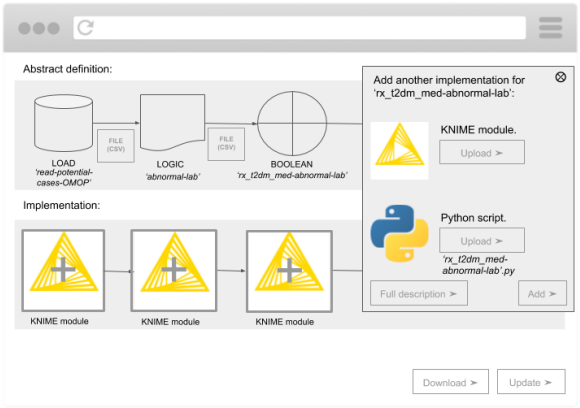


Figure 8: Adding additional implementation units for an existing step.

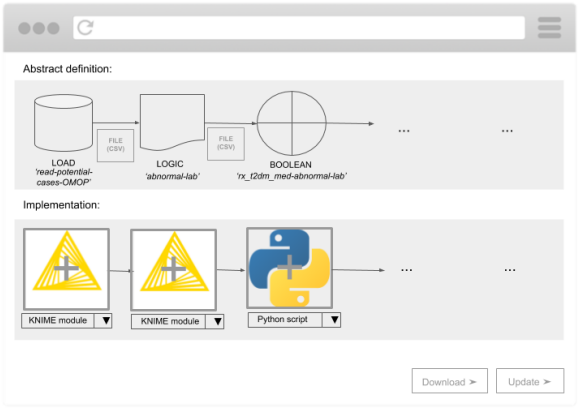


Figure 9: Customising a computable phenotype for local use.

If another researcher wishes to later supply an alternative implementation unit for any of the existing units, thus introducing an additional module in the computational layer, they can do so, and this process is represented in Figure 8. Here, another author has accessed a previously authored definition, and is in the process of adding an alternative implementation for the third step in the computational layer; previously implemented as a KNIME module, the second author is now uploading a Python realisation of the same abstract boolean expression.

Given the potential for multiple permutations of the computational layer, and associated implementation units, when accessing the definitions authored by others, a user is able to pick the permutation they wish to use in order to generate a computable phenotype for local use. This process is represented in Figure 9, where a user is selecting, from the stored implementation units, the exact structure of the computable phenotype; they have chosen a permutation that mixes KNIME and Python implementation units.

Once a phenotype has been defined, and a user has customised the implementation units connected to this definition, the information elicited by the web portal is sent to the generation service in the Phenoflow architecture, which, backed by *python-cwlgen* (https://github.com/common-workflow-language/python-cwlgen), instantiates the definition as a text-based Common Workflow Language (CWL) document, and sends it back. This document is then combined with the stored implementation units, and packaged as a download for the user to execute locally as a computable phenotype, using one of CWL's execution engines (e.g., *cwl-tool*, https://github.com/common-workflow-language/cwltool). As these engines typically integrate with container technology, we have developed several custom images to support the execution environments specified in the computational layer, including a custom KNIME Docker image.

Once it has received the CWL document back from the generator service, the web portal also sends this document to the visualisation service in the Phenoflow architecture, which, backed by *cwlviewer* (https://github.com/common-workflow-language/cwlviewer), sends back a visualisation of the abstract and functional layers expressed in the supplied workflow. This ensures that the text-based CWL instantiation of the definition is complemented by a visualisation that presents the definition in a format more commonly seen (e.g. as seen in Figure 2).

Results

	Knowledge	Clause	Programming	Total
COVID-19: Traditional Code	0	2	2	4
COVID-19: Structured Code	0	0	0	0
T2DM: Traditional Logic	1	1	2	4
T2DM: Structured Logic	0	1	0	1

Table 1: KIP scores indicating the portability of traditional code-based (COVID-19; GSTT) and logic-based (Type 2 Diabetes (T2DM); NMEDW) phenotype definitions and their structured counterparts.

Table 1 presents the KIP portability scores for traditional code-based and logic-based phenotype definitions, and

their structured counterparts. These scores were discussed and agreed upon by all of the authors, a subset of whom have extensive experience both applying and validating KIP scores. The KIP assigns a score between 0 and 2 to phenotype definitions under a number of different portability aspects, with higher scores indicating that a definition is less portable. To understand these scores better, the following sections present the impact of our model on our representative phenotype definitions for COVID-19 and T2DM, under each aspect of the KIP. Recall that we are able to directly compare the portability of the traditional and structured forms of a definition, as we have already verified that converting from one form to the other does not affect the logic of a phenotype.

Knowledge conversion

The first aspect of the KIP scoring system relates to the clinical knowledge required to develop a computable phenotype from its definition. For example, the original COVID-19 phenotype definition is based on common vocabularies, and is thus awarded a knowledge conversion score of 0. However, in the original T2DM phenotype definition, we note the use of some more complex medical concepts (e.g., *T2DM Rx precedes T1DM Rx*, Figure 2). Supplementary information about the meaning of this terminology is provided, but not within the definition itself, and in a different form (as an additional written document). As a result, we assign a score of 1 for knowledge conversion for the original T2DM definition.

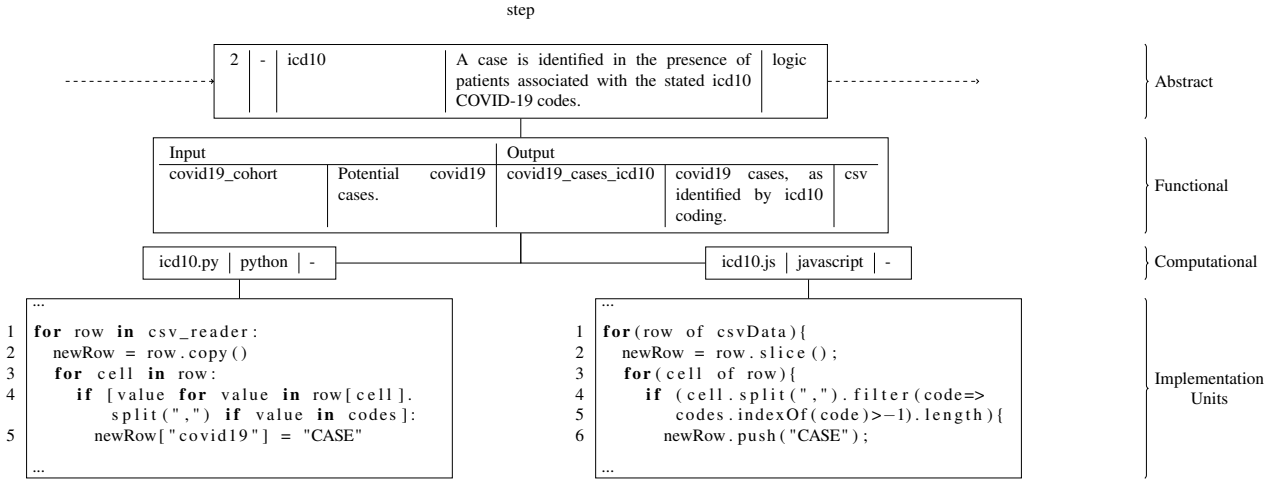


Figure 10: Individual step of COVID-19 structured phenotype definition and new implementation units.

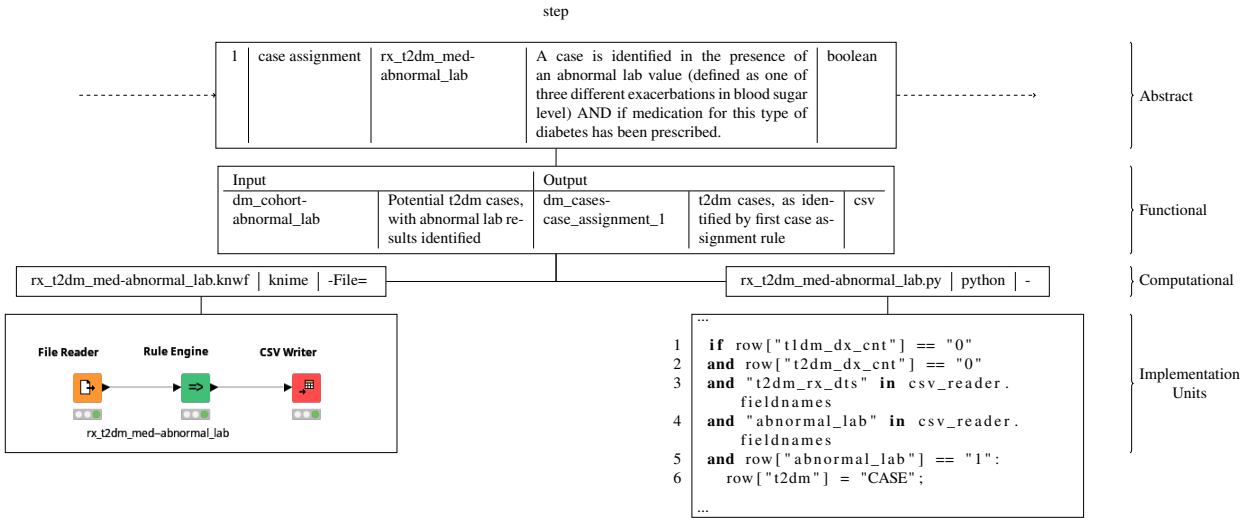


Figure 11: Individual step of T2DM structured phenotype definition and new implementation units.

In its structured form (Figure 10), the COVID-19 definition retains its portability level for knowledge transfer, and while the T2DM phenotype (Figure 11) retains the terminology found in the original definition (e.g. *rx_t2dm_med-abnormal_lab*), the impact this has on portability is lessened in two key ways. Firstly, additional information about the meaning of the terminology is provided within the abstract layer itself, in the description field of each step, ensuring that any medical terminology is supplemented by a longer, more accessible, description (e.g. an explanation of abnormal lab values). Secondly, the classification of each step as a type of operation from a pre-defined ontology ensures that even in the presence of medical terms, basic understanding about the logic of a step can still be extracted. For example, the classification of a step containing a *case assignment* rule as a boolean expression ensures that the use of medical terminology does not obscure its logic. Based upon these factors, the KIP system assigns a value of 0 for this aspect.

Clause interpretation

The second aspect of the KIP scoring system aims to identify any ambiguity in the logical clauses found in a phenotype definition, which may result in inconsistencies when realising this logic computationally. The existing T2DM phenotype definition (Figure 2) uses long conditional clauses (represented graphically), however the logic still has a clear interpretation. This leads to the attribution of a further KIP score of 1. In contrast, the COVID-19 definition, existing as a set of code lists, has a much less clear interpretation, omitting key information, such as the order in which the codes are to be applied, and how the lists are logically connected (e.g. conjunctive vs. disjunctive). This results in the awarding of a KIP score of 2 for this aspect.

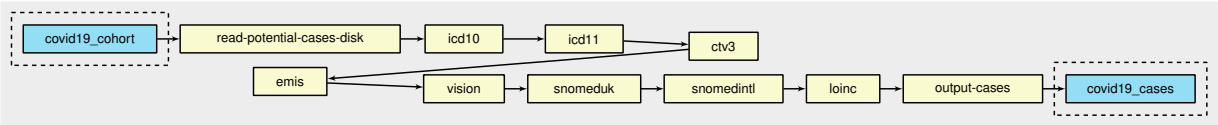


Figure 12: Visualisation of COVID-19 structured phenotype definition

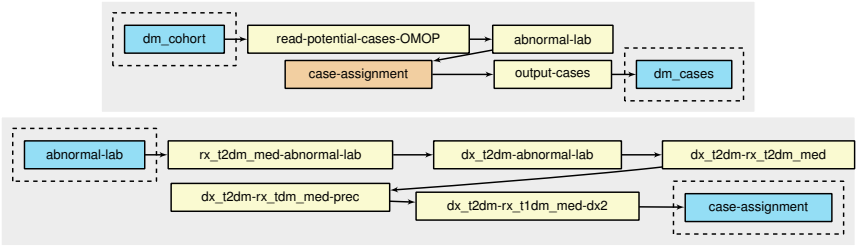


Figure 13: Visualisation of T2DM structured phenotype definition.

The translation of the code-based COVID-19 definition to the structured form enables much of this key information to be expressed explicitly. For example, as can be seen from Figure 12 where a visualisation of the re-authored COVID-19 definition is shown, the order in which each set of codes is to be applied is now clear, and their incremental application confirms a disjunctive connection. For this reason, a new portability score of 0 is assigned. The impact of the structured form on the T2DM definition is less marked. An additional visualisation containing the abstract layer of the re-authored T2DM definition is shown in Figure 13, where the second box shows a grouping of case assignment rules as a set of nested steps, referenced by the parent step shown in orange in the first. Each of the steps in this group, which are evaluated in sequence, contains an individual boolean expression, such as the one defined in Figure 11. This aims to increase the clarity of the interpretation further, by breaking down the long clauses seen in the original abstract layer (Figure 2). Moreover, the use of a group (nested steps) itself, furthers this clarity by allowing for the overall role of these steps to be more easily identified within the abstract layer. However, while there is a simpler overall structure, at the same time the longer descriptions within each step introduce different complexity. For these reasons, the same KIP score of 1 is assigned to this aspect under the structured definition.

Programming

The final aspect of the KIP scoring system relates to the programmatic complexity of implementation. The structure of the original COVID-19 definition suggests a low level of programming expertise required to produce a computable form (e.g. the requirement to produce a Python script to identify the stated codes within a dataset), while the T2DM definition suggests a moderate level of programming expertise (e.g. the requirement for a data pipeline to be produced to realise the stated logical conditions). However, the fact that little instruction can be extracted from each definition on how to develop a computable form in practice increases the complexity of implementation, and results in a disconnect between the two, which reduces the intelligibility of the implementation. This is particularly marked in T2DM, where the case assignment logic seen in Figure 2, while defined as separate operations in the definition, is obscured in a single node in the computable form (*assign case status*, Figure 3), making the correspondence between the two unclear. This reduced intelligibility makes it harder to reuse, or modify, the provided implementation in a new use case. As a result of this complexity, and the implications, a KIP score of 2 is awarded for both original definitions accordingly.

In contrast, the requirement for a distinct (set of) implementation unit(s) for each step in the abstract layer, introduced by the computational layer of a structured definition, each of which responds to the inputs, and produces the outputs, specified in the functional layer, provides a clear template for development. This lessens the implementation burden, in the case of both the COVID-19 and T2DM definitions, by either structuring new development, or allowing existing implementation units, which may have been developed locally, to be reused in order to produce the computable form of a definition.

In addition, a computable phenotype produced on the basis of a structured definition is inherently more intelligible, as the implementation holds a greater correspondence with the abstract layers. For example, in the case of T2DM, because each step in the abstract layers must be connected to an individual implementation unit, the case assignment logic is no longer obscured (as seen in Figure 11), as it was in the original computable form. As a result of this increased intelligibility, these computable phenotypes are more transparent, and thus reusable and more easy to modify, lessening the implementation burden on future developers. Moreover, assuming multiple implementation units exist for the same abstract step (previously written by other authors), which can be easily swapped in and out owing to the modularity brought by the computational layer, a user is more likely to find a unit written in a technology they are comfortable with, and can thus edit, again reducing the implementation burden. For example, our structured COVID-19 and T2DM definitions reference a mix of Python, Javascript and KNIME implementation units.

The ability to modify existing computable phenotypes structured according to our model is only increased by their delivery as executable CWL documents by the Phenoflow architecture. As CWL documents, modifications to these phenotypes can be rapidly tested against execution engines that leverage container technology to avoid having to manually install execution environments. All of these factors result in the attribution of a score of 0 for the KIP programming aspect, for both the COVID-19 and T2DM structured definitions.

Discussion and Conclusion

In this paper, we introduce a workflow-based, multi-layer model for the definition of a phenotype, and an associated microservice architecture, *Phenoflow*, which is used to define phenotypes under this model, and export them as workflows, which can later be executed against a dataset along with associated implementation units.

Overall, we note a number of improvements to portability when a phenotype definition is structured using our representation model, under the KIP scoring system. For code-based definitions, benefits are best seen in terms of a clarity of structure, brought by the requirements of our model, while fewer improvements in portability are seen in terms of terminology. For logic-based definitions, benefits are best seen in terms of clarity of terminology, brought by supplementary information in the abstract layer, while fewer improvements are seen in terms of clause interpretation, where the fact that long clauses are (necessarily) replaced with individual steps, introduces different complexity with equivalent effect. For both code and logic definitions, significant portability improvements can be seen in terms of programmatic complexity of implementation, where a structured definition both closely guides the implementation via the additional (functional and computational) layer information, and promotes the development of intelligible computable forms that can later be reused and modified by other authors. Portability is improved even further by the presence of the Phenoflow architecture, which facilitates the collation of implementation units, and facilitates the generation of computable phenotypes from structured definitions. Ultimately, the improved phenotype portability brought about by

our approach not only helps researchers reuse existing definitions in new studies, but also assists in determining the reproducibility of the methods found in published studies.

While the issue of translating an abstract phenotype definition into a computable form is well recognised within the research community, this work offers several key advancements to complement previously developed methods.

The electronic Medical Records and Genomics (eMERGE) Network has a significant record of representing phenotypes for dissemination and publication. This process was originally done by each institution within the eMERGE Network taking a narrative description of the phenotype pseudocode and an accompanying data flow diagram, and translating this into executable code that would run against their local data warehouse. This approach has now progressed to the use of pipeline-based executable representations, such as those using the KNIME analytics platform, which allows the definition of the computable form in a graphical manner. In addition, the eMERGE Network has adopted a common data model – the OMOP CDM – to facilitate the representation and dissemination of phenotype algorithms[15]. This has allowed the graphical authoring of phenotypes using the ATLAS authoring tool. This provides a human-readable representation of the logic, with the benefit of being stored in a format that may be automatically converted to an executable format across multiple database systems at different organisations. While this approach addresses the issues associated with translating an abstract definition into a compatible form and has facilitated the rapid sharing and execution of phenotypes, the OMOP CDM is not globally adopted (although it has seen wide growth and adoption in recent years), the representations are therefore not fully portable to other CDMs or local data models[16]. In contrast, phenotypes developed under our model are not tightly coupled to a single CDM; OMOP CDM is just one data source that can be referenced in a definition (as are standards such as i2b2 and Fast Healthcare Interoperability Resources (FHIR)).

The PhEMA project has also attempted to address the issue of translating an abstract definition to a computable form by proposing the use of a graphical authoring environment that can be used to generate a higher-level, standardised representation of the phenotype logic[17]. Initial work utilised the Quality Data Model (QDM), with more recent development adopting the Clinical Quality Language (CQL). A key aspect of PhEMA's approach is the use of *translators* to take the higher-level representation (QDM or CQL), and convert it into an executable format that may run against a particular CDM. For example, the approach of converting QDM into an executable KNIME pipeline allowed that KNIME representation to still be customised for local execution[18,19]. However, while this also aims to solve the issue of developing a computable phenotype based on an abstract representation, the translators available are often specific to an implementation format, such as KNIME. In contrast, phenotypes developed under our model are not coupled to a single implementation format.

Future work will explore the impact that Phenoflow has on the portability of additional types of phenotype definitions, including probabilistic definitions, the development of which is likely to leverage data processing tools such as the Flexible Data-Driven Pipeline *(FIDDLE)* framework[20]. In addition, future work will investigate how the multi-dimension annotations of the structured definition model can be leveraged in order to introduce new search and discovery capabilities into phenotype repositories. For example, the ability to use a wider range of search criteria, or to understand when existing definitions intersect with those currently in a repository, to assist in finding partial phenotype matches for a user's requirements, which can then be adapted to suit their needs. Future work will also focus on developing libraries of both abstract steps, and implementation units, to be made available to researchers wanting to customise an existing computable phenotype within Phenoflow for their research tasks. Such efforts are already underway in the UK, under the umbrella of the Health Data Research UK (HDR UK, https://www.hdruk.ac.uk) network which is developing a National Human Phenome portal, of which Phenoflow is a part. Moreover, a broader range of implementation languages will be supported, by developing implementation unit plugins, such as those that perform file type conversion, e.g., from a CSV file to a lightweight SQL table, so that an SQL script can be executed against the data within an individual step. Finally, we will investigate how our experiences of developing our structured definition model and the Phenoflow architecture can be extrapolated to a set of heuristics to be followed when designing and sharing novel phenotype definitions.

References

[1] Richesson R, Smerek M. Electronic health records-based phenotyping. In: Rethinking clinical trials: A living textbook of pragmatic clinical trials. Duke Clinical Research Institute; 2014. p. 1–19.

[2] Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability. Journal of the American Medical Informatics Association. 2016;23(6):1046–1052.

[3] Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. Journal of the American Medical Informatics Association. 2019;26(12):1545–1559.

[4] Hey T, Trefethen AE. Cyberinfrastructure for e-Science. Science. 2005;308(5723):817–821.

[5] Bernardi F, Santucci JF. Model design using hierarchical web-based libraries. In: Design Automation Conference. New York, New York, USA: ACM Press; 2002. p. 14–17.

[6] Hull D, Wolstencroft K, Stevens R, Goble CA, Pocock MR, Li P, et al. Taverna: a tool for building and running workflows of services. Nucleic Acids Research. 2006;34(Web-Server-Issue):729–732.

[7] Chapman M, Tyson G, McBurney P, Luck M, Parsons S. Playing hide-and-seek: an abstract game for cyber security. In: 1st International Workshop on Agents and CyberSecurity (ACySE); 2014. p. 1–8.

[8] Curcin V, Ghanem M, Guo Y. The design and implementation of a workflow analysis tool. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2010;368(1926).

[9] Curcin V, Bottle A, Molokhia M, Millett C, Majeed A. Towards a scientific workflow methodology for primary care database studies. Statistical Methods in Medical Research. 2010;19(4):378–393.

[10] Shang N, Liu C, Rasmussen LV, Ta CN, Caroll RJ, Benoit B, et al. Making work visible for electronic phenotype implementation: Lessons learned from the eMERGE network. Journal of Biomedical Informatics. 2019 11;99.

[11] Rasmussen LV, Kiefer RC, Mo H, Speltz P, Thompson WK, Jiang G, et al. A Modular Architecture for Electronic Health Record-Driven Phenotyping. AMIA Joint Summits on Translational Science. 2015;2015:147–51.

[12] Bird R, Wadler P. An Introduction to Functional Programming. Prentice Hall International (UK) Ltd.; 1988.

[13] Sam Newman. Monolith to Microservices: Evolutionary Patterns to Transform Your Monolith. O'Reilly; 2019.

[14] Chapman M, Curcin V. A Microservice Architecture for the Design of Computer-Interpretable Guideline Processing Tools. In: 18th International Conference on Smart Technologies (EUROCON); 2019. p. 1–6.

[15] Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. Journal of Biomedical Informatics. 2019 8;96.

[16] Rasmussen L, Brandt P, Jiang G, Kiefer R, Pacheco J, Adekkanattu P, et al. Considerations for Improving the Portability of Electronic Health Record-Based Phenotype Algorithms. In: AMIA Symposium; 2019. p. 755–764.

[17] Rasmussen LV, Kiefer RC, Mo H, Thompson WK, Jiang G, Pacheco JA, et al. The Phenotype Execution and Modeling Architecture (PhEMA) - A Standards-Based Composition of Software for Phenotype Algorithm Development. Northwestern; 2015.

[18] Mo H, Jiang G, Pacheco JA, Kiefer R, Rasmussen LV, Pathak J, et al. A Decompositional Approach to Executing Quality Data Model Algorithms on the i2b2 Platform. AMIA Joint Summits on Translational Science. 2016;2016:167–75.

[19] Pacheco JA, Rasmussen LV, Kiefer RC, Campion TR, Speltz P, Carroll RJ, et al. A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments. Journal of the American Medical Informatics Association. 2018 8;25(11):1540–1546.

[20] Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. Journal of the American Medical Informatics Association. 2020 oct;27(12):1921–1934.

Early Detection of Post-Surgical Complications using Time-series Electronic Health Records

David Chen, Ph.D.[1], Jun Jiang, Ph.D[1], Sunyang Fu, M.H.I[1], Gabriel Demuth, Ph.D.[2], Sijia Liu, Ph.D.[1], Gavin M. Schaeferle B.S.[2], Patrick M. Wilson M.P.H & M.S.[2], Elizabeth Habermann, Ph.D.[2], David W. Larson, M.D., M.B.A.[3], Curtis Storlie, Ph.D.[2], Hongfang Liu, Ph.D.[1]

1 Division of Digital Health Sciences, 2 Department of Health Science Research,

3 Department of Colorectal Surgery, Mayo Clinic, Rochester, MN, USA.

Abstract

Models predicting health complications are increasingly attempting to reflect the temporally changing nature of patient status. However, both the practice of medicine and electronic health records (EHR) have yet to provide a true longitudinal representation of a patient's medical history as relevant data is often asynchronous and highly missing. To match the stringent requirements of many static time models, time-series data has to be truncated, and missing values in samples have to be filled heuristically. However, these data preprocessing procedures may unconsciously misinterpret real-world data, and eventually lead into failure in practice. In this work, we proposed an augmented gated recurrent unit (GRU), which formulate both missingness and timeline signals into GRU cells. Real patient data of post-operative bleeding (POB) after Colon and Rectal Surgery (CRS) was collected from Mayo Clinic EHR system to evaluate the effectiveness of proposed model. Conventional models were also trained with imputed dataset, in which event missingness or asynchronicity were approximated. The performance of proposed model surpassed current state-of-the-art methods in this POB detection task, indicating our model could be more eligible to handle EHR datasets.

Introduction

Predicting post-operative complications (PSC) following medical procedures is a major topic of interest for clinical decision support and medical discovery research. The rate of post-operative complications are heavily dependent on the type of procedure, ranging from less than 1% in low risk procedures such as cataract surgery to >15% for craniectomies (1, 2). However, when they do occur, they can be devastating to the medical status of the patient and the cost associated with the events. This is particularly true if events are caught late and risky emergency surgery is required. Earlier identification and or prediction of patients at high risk may allow for less invasive inventions or postponement of surgery all-together if appropriate.

Models to identify patients at high risk of complications have thus-far been largely static-time models (3-5). Although these models can incrementally improve patient safety, they often ignore highly informative changes in longitudinal patient health status due to their implicit limitations of the algorithms. Some have extracted specific features of the temporal dynamics such as max rate of change, Fourier domain analysis, or pre-engineered patterns to identify patterns of interest in time-series (3). However, these features are sometimes obscured by noise or temporally distorted in real-time. Other models more suited for time-series data such as hidden Markov models or conditional random fields have also been proposed but do not incorporate the same ability to learn nonlinear relationships (6).

Recently, recurrent neural networks (RNN) have provided a methodology to efficiently learn nonlinear relationships from large time-series datasets without extensive subject matter expertise (7). They have been applied to a wide range of sequential data including stock market prediction, language modeling, and text summarization (8-10). However, the application of RNN to healthcare data is not straightforward due to the unique characteristics of healthcare time-series. Unlike the previously mentioned applications, healthcare data is often rife with data variability and veracity issues such as data asynchronicity and missingness. Previous applications of RNN in healthcare such as predicting mortality in the ICU or progression of Alzheimer's disease have made simplistic

assumptions of the data which likely reduce their accuracy (11-14). Recently, Nguyen et al. adopted minimalRNN with the concept of "model filing" (predictions/missing values of the RNN are used as inputs for the next timepoint) to predict Alzheimer's disease progression. The proposed strategy outperformed traditional "pre-processing" approaches (15). Tan et al. leveraged dual-attention mechanism to handle missing values. The method was achieved by assigning smaller weights to missing data in order to reduce its signal through an attention mechanism (16).

In this work, we develop a variant of RNN to explicitly address data asynchronicity and missingness through the mechanism called Gated to Missingness. Such a model has the potential to learn informative missingness that would otherwise be loss if imputation methods are carried out, thereby potentially reducing uncertainly and improving prediction accuracy. Our objective is to provide an algorithmic tool which readily accepts real-world healthcare data without the need for complex imputation or inaccurate assumptions. We demonstrate our proposed model on a real-world surgical cohort to predict PSCs.

Materials

This study was approved by the Mayo Clinic institutional review board. All patients included in this study gave written approval for their data to be used for research purposes. A retrospective dataset was constructed including 13,399 colorectal surgical cases (Demographics summarized in Table 1) in 12,402 unique patients performed at Mayo Clinic Hospital. The adverse event of interest was post-operative bleeding (POB), which was defined by any type of gastrointestinal hemorrhage occurring within one week following CRS (Colon and Rectal Surgery). The definition for POB can vary depending on the source with some dependent on blood volume lost, requirement for transfusion of blood products, and/or diagnostic and procedural code combinations (9). In a previous study, we leveraged a combination of ICD-9 diagnosis codes with laboratory results (drop of hemoglobin concentration greater than 3 g/dL) and/or transfusion records to identify only the presence of bleeding (10). In order to identify the time of bleed for this study, we used the order time of the hemoglobin blood test corresponding to the pre-defined drop in hemoglobin concentration and/or the order time of blood products.

Structured patient data (demographics, laboratory results, vitals, flowsheet data, and ICD-9 codes) were pulled from Mayo Clinic's unified data platform. Comorbidities found using ICD-9 codes diagnosed within 1 year of procedure date was considered "static data". Where defined, we reduced the dimensionality of the ICD-9 code using Charleston comorbidity index definitions. Any remaining codes which occurred in less than 1% of cases were discarded. We also created "static" demographics data (age, bmi, race) using the most recent demographics data recorded prior to the procedure. The "longitudinal" data was created using laboratory, vitals, and flowsheet data. All data recorded between time of admission and time of discharge or death. As the data falls on a continuous spectrum, we needed to discretize the time steps to reduce the possible length of each sequence. We settled on 4 hour time steps to keep the majority of high frequency data (such as vitals) while not trivially increasing low frequency data (such as troponin) missingness. If multiple data points are found during a single time step, the mean of the data points was used. The result of this is a matrix of data with X variables by Y time long. Because patients had varying length stays, all shorter stays were zero padded past the end of the sequence.

Table 1:
Characteristics of patients with POB compared to those without. Percentages indicate incidence rate within a group.

		No POB n=11,719	**POB n=1,680**	**p-value**
Demographics				
mean age		57.6 ± 17.8	61.5 ± 16.7	<0.001
male sex		6,206 (53.0%)	865 (51.5%)	0.271
BMI		30.3 ± 11.0	27.8 ± 8.6	0.656
Race				
	white	10,612 (90.6%)	1,447 (86.1%)	<0.001
	black	49 (0.4%)	18 (1.1%)	<0.001
	asian	74 (0.6%)	14 (0.8%)	0.433
	other	164 (1.2%)	35 (2.1%)	0.041
alcohol use		723 (6.2%)	123 (7.3%)	0.078
smoker		2,097 (17.9%)	277 (16.5%)	0.168
drug use		180 (1.5%)	39 (2.3%)	0.023
Indications for surgery				
rectal cancer		2,060 (17.6%)	321 (19.1%)	0.134

colon cancer	5,098 (43.5%)	654 (38.9%)	<0.001
inflammatory bowel disease	3,325 (28.4%)	461 (27.4%)	0.444
other	2,916 (24.9%)	244 (14.5%)	<0.001
Surgical characteristics			
partial colectomy	4064 (34.7%)	728 (43.3%)	<0.001
total colectomy	4531 (38.7%)	626 (37.3%)	0.281
proctectomy	5553 (47.4%)	650 (38.7%)	<0.001
laparoscopic	3,919 (33.4%)	206 (12.3%)	<0.001
elective surgery	565 (4.8%)	211 (12.6%)	<0.001
median ASA score	2 ± 0.71	3 ± 0.79	<0.001
Outcomes			
mean hospital stay	6.5 ± 4.7	13.4 ± 13.8	<0.001
death	52 (0.4%)	57 (3.4%)	<0.001

Methods

PSC-related Event modeling

To understand the efficacy of RNNs for this task, we compare RNN variants to conventional static time models. Given that we expect there would be specific temporal patterns which would indicate POB, we engineered features to capture some of the temporal dynamics of the data to feed into the static model. These include minimum, mean, and maximum values, as well as the minimum, mean, and maximum rate of changes for all temporally changing data. We also formatted the longitudinal data into sequences for use with RNNs starting from the end of surgery and ending at either a POB event or discharge from hospital. Data examples were illustrated in Figure 1, in which asynchronous events were demonstrated. For patients who experienced an event or were discharged before the 7th day, we padded to sequence with 0s. To limit the length of data, we cropped the time analyzed to be 7 days after end of surgery. To deal with asynchronousness of healthcare process, we structured the sequences to an interval of 4 hours between data points. We chose 4 hours given that it gave enough granularity to capture acute changes in physiological measurements but not too many to result in overly lengthy and highly sparse sequences. If multiple measurements fell between these points, the average of the measurement was taken. If no measurement was found during an interval, we initially left it as missing.

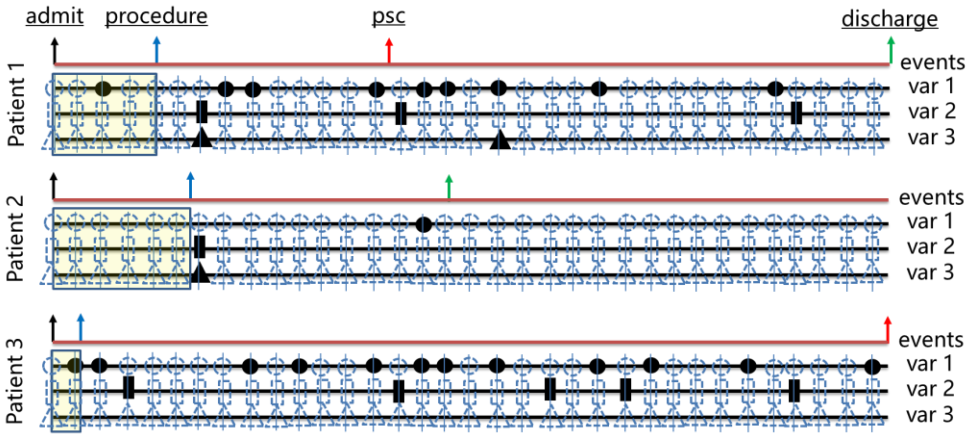


Figure 1. Examples of asynchronous patient data. Medical events were denoted with a black circle, rectangle or triangle. Black arrow denotes patient's admission, blue arrow denotes medical procedure (surgery), red arrow denotes post-surgical complications and green arrow denotes patient's discharge.

Recurrent Neural Networks

RNNs are powerful neural network architectures for learning from sequential data. RNNs comprise of "cells" which contain the weights and activation functions necessary to learn the non-linear relationships between different sequential elements. These cells are then recursively stacked, providing the architecture necessary to learn from

sequential data. The basic RNN cell (Figure 2a) consists of a single hyperbolic tangent activation which combines the input (x_t) from the previous hidden layer (h_{t-1}) and current element.

$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b)$$

Here, $W_{x,h}$ denotes the weights associated with the input and the hidden layers and b denotes the bias. The primary issue with this simplistic architecture is that as the length of the sequence increases, the back-propagated gradient which is necessary to learn the weights of each individual cell run into limitations of machine precision. This is called the vanishing or exploding gradient problem (17). Practically, this limits the length of sequences which can be accurately modeled by the basic RNNs to low double digits. RNN variants, such as long short-term memory (LSTM) (18) and the gated recurrent unit (GRU) (19) have been proposed to address this issue by separating long and short term dependencies into individual gates. This provides more pathways for the back-propagated gradient to continue through each cell. In this work, we focus on GRU and modifying it to adapt our task, since GRU has fewer parameters to learn compared to LSTM.

The GRU (Figure 2b) can be summarized by the following equations:

$$z_t = \sigma(W_{xz} x_t + W_{hz} h_{t-1} + b_z)$$
$$r_t = \sigma(W_{xr} x_t + W_{hr} h_{t-1} + b_r)$$
$$\tilde{h}_t = \tanh(W_{xh} x_t + W_{rh}(r_t \cdot h_{t-1}) + b_h)$$
$$h_t = (1 - z_t) \cdot h_{t-1} + r_t \cdot \tilde{h}_t$$

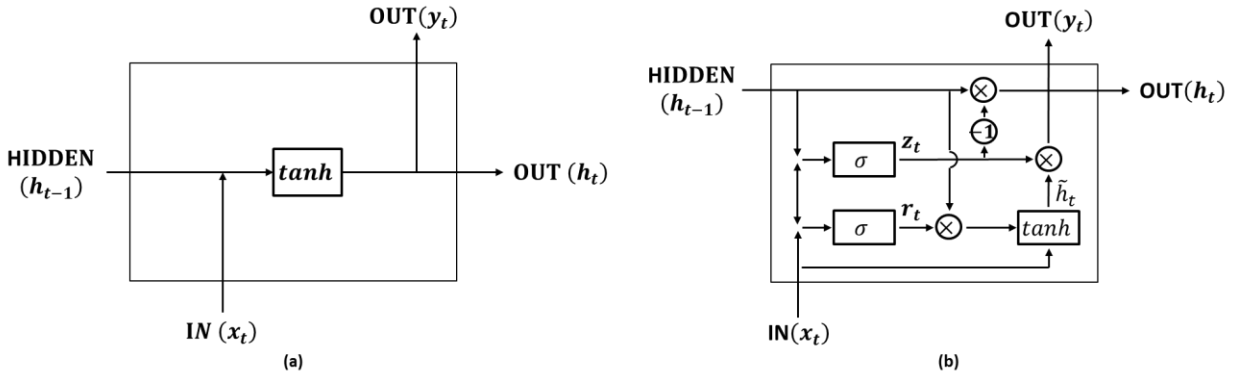


Figure 2. Cell of Standard RNN and GRU (gated recurrent unit)

Where z_t, r_t, and $\tilde{h}_t$ denote the update, reset, and weighted hidden layers respectively. W and b denote model parameters and σ denotes sigmoid activation function.

Although GRUs are built to deal with sequential data, they critically ignore two aspects of real-world data: missingness and time (Figure 1). The standard GRU assume that sequence data is complete and there are no missing elements. Most workflows do some sort of simple imputation such as mean or last observation carried forward. However, these simplistic strategies often destroy information associated with the missingness or time since real observed data. For example, laboratory tests measuring troponin values are rarely ordered in surgical setting, and as such is difficult for many models to learn from the data. However, the very lack of a troponin test has been shown to indicate positive patient health. This correlation can be intuitively understood that doctors would rarely, if ever order a troponin test for patients not suspected of experiencing myocardial ischemia. Likewise, uncertainty caused by long times between tests can also be informative. Furthermore, not only can simple imputation remove potential information implicit in the data, it can also introduce significant bias into the data due to gross errors different from the true. Therefore, the standard GRU is often not well suited for sequential healthcare data.

To address these problems of the standard GRU, we modify the standard GRU architecture (Figure 2b) to explicitly learn from both missingness and data veracity. Two separate sequential indicators were created for GRU cells to formulate missingness and time line in real-world data: a binary missingness indicator and a time from last measurement indicator. In order to evaluate model performance of including missingness and time line into training

process. GRU cells were augmented by adding missingness, time line or both signals, and were denoted as GRU-m, GRU-t and GRU-mt respectively. The architectures of the GRU cells were shown in Figure 3.

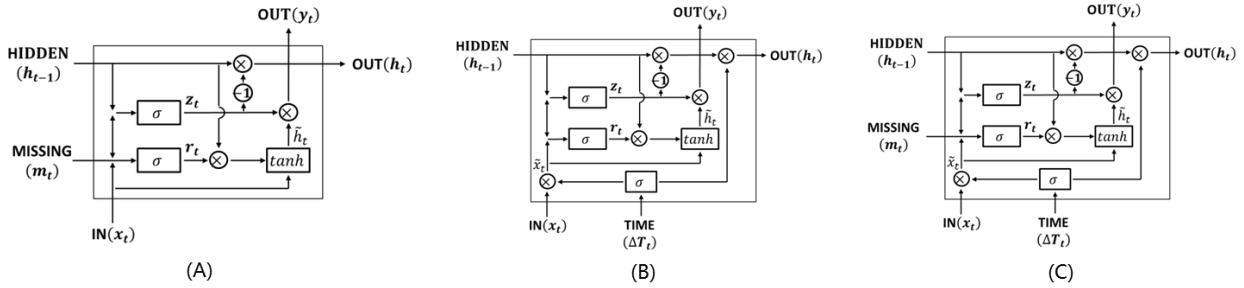


Figure 3. Augmented GRU cells with missingness and time sensitive inputs. A: Gated to missingness (GRU-m). B: Time sensitive GRU cell (GRU-t). C: Time sensitive GRU cell with gated to missingness (GRU-mt).

Gated to Missingness (GRU-m)

It is well known that missingness can be extremely informative (13). Therefore, we created a binary missingness indicator for a GRU cell to modulate missingness signals in training data. We imputed the missing elements simplistically into either the last observation carried forward (LOCF) or with the mean. With missingness m_t as input, the functionality of GRU-m cells can be summarized as below:

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + W_{mz}m_t + b_z)$$
$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + W_{mz}m_t + b_r)$$
$$\tilde{h}_t = \tanh(W_{xh}x_t + W_{rh}(r_t \cdot h_{t-1}) + W_{mz}m_t + b_h)$$
$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h_t}$$

Time Sensitive GRU cell (GRU-t)

Furthermore, the time between real results can also be highly informative. A time from last measurement indicator was created to capture time serials signals. With data to data certainty time (ΔT_t) as an extra input, GRU-t can be summarized by the following equations:

$$\gamma_t = \sigma\left(W_\gamma \Delta \boldsymbol{T_t} + b_\gamma\right)$$
$$\tilde{x}_t = x_t \times W_{\gamma \mathrm{x}}\gamma_t$$
$$z_t = \sigma(W_{xz}\tilde{x}_t + W_{hz}h_{t-1} + b_z)$$
$$r_t = \sigma(W_{xr}\tilde{x}_t + W_{hr}h_{t-1} + b_r)$$
$$\tilde{h}_t = \tanh(W_{xh}\tilde{x}_t + W_{rh}(h_{t-1} \times r_t) + b)$$
$$y_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t$$
$$h_t = y_t \times W_{\gamma \mathrm{h}}\gamma_t$$

Time Sensitive GRU cell with Gated to Missingness (GRU-mt)

Going forward, we use similar terminology to typical LSTM as the functions are similar, although more explicitly incorporating the characteristics of the input data sequence. Prior works have adopted similar approach. For instance, Che et al (13) embedded exponential decay terms following the input and hidden layers to explicitly reduce their contribution to the relationship. However, this assumes that either missingness or long times since last observations reduce data veracity when in fact, these can actually increase certainty. Additionally, that implementation is constrained to a single RNN layer as the missingness and veracity lags are cannot be abstractly passed onto a 2nd layer. Therefore, we make two modifications. First, we add two additional gates sensitive to missingness and data veracity interval. First of which is an "input gate" ($\tilde{x}_t$) which is sensitive to the data missingness (m_t). The second gate is a "forget gate" (f_t) which modulates the sensitivity of the data to data certainty time (ΔT_t). The second modification is outputting the vector of these gates, which provides an abstract representation of the missingness and data veracity time, allowing it to be used in an arbitrary sized, multi-layer deep hidden layer (Figure 4). The resulting cell which we term GRU-mt is represented by the functions as follows:

$$\gamma_t = \sigma\left(W_\gamma \Delta \boldsymbol{T_t} + b_\gamma\right)$$
$$\tilde{x}_t = x_t \times W_{\gamma \mathrm{x}}\gamma_t$$
$$z_t = \sigma(W_{xz}\tilde{x}_t + W_{hz}h_{t-1} + W_{mz}m_t + b_z)$$
$$r_t = \sigma(W_{xr}\tilde{x}_t + W_{hr}h_{t-1} + W_{mz}m_t + b_r)$$

$$\tilde{h}_t = \tanh(W_{xh}\tilde{x}_t + W_{rh}(h_{t-1} \times r_t) + W_{mh}m_t + b)$$
$$y_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t$$
$$h_t = y_t \times W_{\gamma h}\gamma_t$$

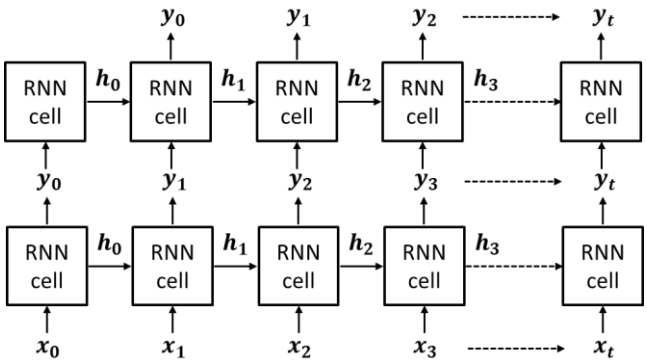


Figure 4. Multilayer RNN for arbitrary sized input.

Experiments

All models were trained to identify POB 24 hours prior to the identified time of bleeding or time of discharge. This provides a buffer for which possible targeted surveillance and corresponding intervention can be applied. We compared a standard static time gradient boosting machine (GBM) model with variants of RNNs. GBMs are widely used in predictive tasks due to their simplicity and robust performance out of the box. Because the GBM model is static time, the aggregated sequential data is inherently LOCF imputation. Furthermore, GBM inherently handles some missing data. Therefore, no mean imputation is done.

The dataset was randomly split into 80/20 training (n=10,729, cases=1,352) and testing (n=2,670, cases= 328) datasets by patient. Models were trained and optimized on the training dataset. The test was then used to evaluate the final performance of the model. Model hyperparameters such as learning rate, number of hidden layers, activation function, and number of training epochs were found using grid search. Area under the receiver operating curve (AUC) was used to assess the performance of different models. For both MLP and LSTM models, we further tuned learning rate, dropout, activation function, loss function, and number of training epochs. The optimization metric for the MLP and LSTM models was accuracy. Areas under the receiver operating curve (AUROC) were used to assess the performance of the different models.

Results

A total of 1,680 cases (12.5%) were found to have experienced POB. The median time to experience a POB event was 2.3±2.1. The overall 30-day mortality rate was 0.8% (109 cases). The median time spent in hospital was 1.7±4.6 days. The characteristics of the longitudinal dataset were summarized in Table 2, in which "Event #" denotes the event counts for each patient; "Missingness ratio" denotes the ratio of missingness counts to all feature dimensions; "Event span" denotes the time span for each patient (unit: days).

Table 2. Longitudinal Dataset Characteristics.

Unique PT	event # (mean±std)	Missingness ratio (mean±std)	Event span (mean±std)
12,402	120.41±514.70	0.90±0.07	5.96± 20.76

Table 3 shows the performance of various models. The optimal GBM models were trained using 300 trees, maximum depth of 3, and 40% of variables available for splitting. The optimal GRU-mt hyperparameters were found to be learning rate of 1e-5, dropout of 0.20, ReLU activation function, and binary cross entropy, and 75 training epochs. Static time GBM using both only pre-recovery data and all data taken 1 day prior to discharge or event yielded results comparable or higher to the deep learning models. This is particularly true of LOCF imputed models which did not yield better results compared to GBM models using only pre-recovery data. This is true of all GRU variants proposed. However, the deep learning models were able to better model post-surgical bleeding when null imputation and the proposed missingness gates were used. Specifically, the missingness gate yielded higher

results compared to using time since last observation gate. Furthermore, the combination of missingness gate and time since last observation gate yielded the highest results. 1D-CNN yielded the lowest results.

Table 3. Evaluation performance under multiple settings. LOCF refers to models trained using LOCF imputed data. Mean refers to models trained using mean imputed data.

Model	**AUC (LOCF)**	**AUC (Mean)**
Pre-recovery GBM	0.817	
1-day Prior GBM	**0.858**	
1D-CNN	0.811	0.721
GRU	0.821	0.793
GRU-m	0.811	0.842
GRU-t	0.818	0.829
GRU-mt	0.824	**0.866**

Discussion

POB following CRS is often the beginning of a cascade of serious complications, including ileus and anastomotic leak which are correlated with high risk of morbidity and mortality (20). Therefore, early detection and medical intervention is needed to avoid the worst of results. However, real-time prediction using electronic health records is difficult due to the high variability of completeness, timing, and order of data. The model developed in this work explicitly learns from these variabilities, thereby achieving higher predictive performance compared to standard static time classification methods and standard GRU-based methods.

As shown in the results, missingness of data can greatly impact the ability of deep neural networks to accurately model the data. More so, different methods of imputation can either improve or confound the results, as evident in the greatly reduced performance of the standard GRU model using null imputation. Without explicit knowledge of the data, it can be hard for neural networks to learn the necessary relationships in the data. Here, we ask the model to recognize certain values in a sequence to be more relevant than others. Yet, because we do not have an extreme amount of data, this task is difficult for the model to do. This observation has been seen in other similar tasks such as natural language processing tasks where part of speech flags or start or end sentence flags are useful for model performance (21).

It is not particularly surprising that the static time GBM models yield similar performance to the deep learning models. It is well known that complex models with large number of parameters are not well suited to small datasets. However, there are other aspects which may somewhat favor the static time model. First, the GBM model took data from a specific time point (1 day prior to event). It does not need to explicitly integrate longitudinal data (because it cannot) and therefore has less confounding information to use for prediction. Second, we imputed missing data of the static dataset using MissForest (22). This has shown to yield more accurate imputation compared to simplistic methods such as LOCF or mean imputation. Therefore, the comparison between the models are not entirely fair or straight forward.

In this work we tested both a missingness gate and a time since last observation gate. The missingness information produced a higher increase in performance compared to the time since last observation. We speculate that this is due to the design of the gate as it provides each cell with the ability to either “ignore” or “emphasize” missing values. If the cell ignores missingness, the gradient should carry forward to the next cell. However, if the cell emphasizing missingness, cell can stop the continued propagation of the gradient. It is interesting that such a setup seems particularly useful when null imputation is used rather than LOCF imputation. We speculate that the model still has a hard time identifying “wrong” information.

The results presented here demonstrate the utility of machine learning methods in predicting POB in CRS patients using data from EMR. Unlike many prior studies, variables were extracted from the EMR with few limits on the scope. Compared to well annotated and carefully planned clinical registry data (such as ACS-NSQIP), this methodology allows for a more unbiased, data driven approach to research. With this approach, a large number of

patients may be included with minimum overhead, whereas adding additional patients to clinical registries can be prohibitively expensive in both time and money.

We recognize that this study is limited in several ways. For one, we included only a single dataset. Although methods such as cross-validation would provide a more robust indication of generalizability, ideally our method still needs to be proven in other datasets with different characteristics such as different time intervals or different patterns of data missingness. We plan to validate this model using prospective data from all 3 Mayo sites in the future. Also, the dataset we used was moderate in size. We have shown in other studies that significantly larger datasets might produce better results even using standard data processing methods (23). Furthermore, an increase in the number of parameters is actually contraindicated if the dataset is small as regularly seen in healthcare. Therefore, the proposed method is certainly not an off-the-shelf tool for all longitudinal healthcare applications. Second, related to the limited size of training data, the proposed method does assume that missingness is not random. If missingness is random, it is entirely possible that the model would perform worse due to less regularization caused by the increased number of model parameters. However, we believe that because missingness and time-intervals are treated as a gate instead of an explicit delay, our model maybe more robust to such dataset characteristics compared to prior works. Finally, the proposed technique does not impute the data. Therefore, in real-world applications, it would not be appropriate to use such a model to predict laboratory results. Rather, it should be used to inform caregivers the need for additional surveillance.

Conclusion

Real-time prediction of post-surgical complications can be difficult due to the asynchronous and highly missing nature of clinical workflow. Conventional RNN architectures have difficulty dealing with such data, and often times are not superior to off-the-shelf static-time machine learning models which incorporate temporal features such as GBMs. We propose a novel RNN architecture which incorporates knowledge of missingness and time from last observation, thereby explicitly learning informative missingness. The proposed GRU-mt model improved upon existing architectures and imputation methods, potentially providing clinicians with a tool to better distribute limited human resources and accurately direct targeted surveillance. We hope this novel architecture provides a tool to better analyze and model sequential healthcare data.

References

1. Pingree MF, Crandall AS, Olson RJ. Cataract surgery complications in 1 year at an academic institution. Journal of Cataract & Refractive Surgery. 1999;25(5):705-8.
2. Chang V, Hartzfeld P, Langlois M, Mahmood A, Seyfried D. Outcomes of cranial repair after craniectomy. Journal of neurosurgery. 2010;112(5):1120-4.
3. Kunutsor S, Whitehouse M, Blom A, Beswick A. Systematic review of risk prediction scores for surgical site infection or periprosthetic joint infection following joint arthroplasty. Epidemiology & Infection. 2017;145(9):1738-49.
4. Sohn S, Larson DW, Habermann EB, Naessens JM, Alabbad JY, Liu H. Detection of clinically important colorectal surgical site infection using Bayesian network. Journal of Surgical Research. 2017;209:168-73.
5. Chen D, Afzal N, Sohn S, Habermann EB, Naessens JM, Larson DW, et al. Postoperative bleeding risk prediction for patients undergoing colorectal surgery. Surgery. 2018;164(6):1209-16.
6. Esling P, Agon C. Time-series data mining. ACM Computing Surveys (CSUR). 2012;45(1):1-34.
7. Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: a review. Data Mining and Knowledge Discovery. 2019;33(4):917-63.
8. Cao Z, Wei F, Dong L, Li S, Zhou M, editors. Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization. AAAI; 2015: Citeseer.
9. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
10. Selvin S, Vinayakumar R, Gopalakrishnan E, Menon VK, Soman K, editors. Stock price prediction using LSTM, RNN and CNN-sliding window model. 2017 international conference on advances in computing, communications and informatics (icacci); 2017: IEEE.

11. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association. 2017;24(2):361-70.
12. Wang T, Qiu RG, Yu M. Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks. Scientific reports. 2018;8(1):1-12.
13. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Scientific reports. 2018;8(1):1-12.
14. Che Z, Purushotham S, Khemani R, Liu Y, editors. Interpretable deep models for ICU outcome prediction. AMIA Annual Symposium Proceedings; 2016: American Medical Informatics Association.
15. Nguyen M, He T, An L, Alexander DC, Feng J, Initiative AsDN. Predicting Alzheimer's disease progression using deep recurrent neural networks. NeuroImage. 2020:117203.
16. Tan Q, Ye M, Yang B, Liu S, Ma AJ, Yip TC-F, et al., editors. DATA-GRU: Dual-Attention Time-Aware Gated Recurrent Unit for Irregular Multivariate Time Series. Proceedings of the AAAI Conference on Artificial Intelligence; 2020.
17. Hochreiter S, Bengio Y, Frasconi P, Schmidhuber J. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. A field guide to dynamical recurrent neural networks. IEEE Press; 2001.
18. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735-80.
19. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078. 2014.
20. Boccola MA, Buettner PG, Rozen WM, Siu SK, Stevenson AR, Stitz R, et al. Risk factors and outcomes for anastomotic leakage in colorectal surgery: a single-institution analysis of 1576 patients. World journal of surgery. 2011;35(1):186-95.
21. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical Concept Extraction: a Methodology Review. Journal of Biomedical Informatics. 2020:103526.
22. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.
23. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ digital medicine. 2019;2(1):1-5.

Refined, regionally-specific data standards reveal heterogeneity in Hispanic death records

Shamsi Daneshvari Berry, PhD, MS[1,2], Heather J.H. Edgar, PhD[2], Carmen Mosley, PhD[2], Keith Hunley, PhD[2]
1: Western Michigan University Homer Stryker MD School of Medicine, Kalamazoo, MI
2: University of New Mexico, Albuquerque, NM

Abstract

Hispanic ethnicity can be captured with differing levels of granularity using various data standards, including those from the Office of Management and Budget, Health and Human Services and National Academy of Medicine. Previous research identified seven subgroups of Hispanics in New Mexico using open-ended interviews and information about the culture/history of the state. We examined age and manner of death to determine whether differences among subgroups are hidden by less-refined categorization. Significant differences in the mean age at death were found between some groups, including Spanish and Mexican Americans. We found an association between specific manners of death codes and subgroups. However, significance disappeared when manners of death were grouped (e.g. accident, homicide, etc.). This indicates that while certain manners of death are associated with group membership, overall types of death are not. Data descriptors for Hispanics should reflect more refined, regionally relevant groups, in order to unmask heterogeneity.

Introduction

Race and ethnicity are common data elements used in the Census, healthcare, research, and education. However, how race and ethnicity are classified can differ depending on the standard implemented. The Office of Management and Budget (OMB), after extensive testing and public engagement, concluded that Hispanic ethnicity should be captured as "Hispanic or Latino" and "Not Hispanic or Latino." [1] A great deal of research on health disparities in Hispanics and Latinos/as continues to use these overarching categories. [2,3] There also has been considerable research on health and mortality in subgroups of Hispanics.[4–8] Hummer et al.[7] found differences in the age at death among Mexican Americans, Cuban Americans, Central and South American and other Hispanics. They also discovered that mortality differed between the subgroups using age, sex and cause of death. Weinick et al.[4] found differences in emergency room visits, inpatient visits and prescription usage among Mexicans, Cubans and Puerto Ricans. Fenelon et al.[5] found mortality differences between 12 subgroupings of Hispanics using region of origin and nativity. Additional studies have found differences in fertility, birth rates, life expectancies and morbidities. [6,8] These examples point to the importance of considering the heterogeneous nature of Hispanic subgroups for understanding health and mortality disparities.

Health and Human Services (HHS) has suggested the use of categories that can be "rolled up" into the OMB standard when needed. These groups are "Cuban," "Mexican and Mexican American," "Puerto Rican," and "Another Hispanic group."[9] For the 2010 and 2020 censuses the groupings were "Mexican, Mexican Am., Chicano," "Puerto Rican," "Cuban,", and "another Hispanic, Latino, or Spanish origin."[10] These groups have specific cultural and historical significance in some states, such as New Mexico, Florida and New York, but not in others. In many cases, researchers collect data on these relatively refined categories, but then use the "rolled up" groupings as the sample sizes for refined groups can be small, reducing the likelihood of achieving statistical significance in analyses.

The National Academy of Medicine (NAM) report of 2009 suggested using 36 subgroups of the group, "Hispanic," with a greater range of relevance in various regions of the country, in conjunction with the overarching OMB standard[11] (see Table 1). Here, we consider the Hispanic population in New Mexico, using locally relevant ethnic subgroup categories suggested by previous research.[12] We used a list of seven subgroups within the New Mexican Hispanic population: "Chicano/a," "Hispanic," "Latino/a," "Mexican," "Mexican American," "Nuevomexicano/a," and "Spanish."[12] Unlike the HHS or OMB standards, these categories are specific to the culture and history of New

Mexico. They also match the NAM categories with the addition of "Nuevomexicano/a" and "Hispanic." Hispanic is added to the list of subgroups because it is the most common term chosen by New Mexicans. Note that "Hispanic" describes both a subgroup and a catch-all term for all subgroups combined. Nuevomexicano/a is a regionally and temporally specific term used in New Mexico; individuals who identify using this term have similar migration histories to individuals who identify as Spanish. [12,13]

Understanding variation among subgroups such as these may be useful in unraveling the "Hispanic Paradox," in which, in the US at large, Hispanics have better health outcomes than non-Hispanic Whites (NHW,) despite lower socioeconomic standing. [14–19]

Table 1. National Academy of Medicine subgroups of Hispanic Ethnicity.

Andalusian	Catalonian	Ecuadorian	Mexican American	Salvadoran
Argentinean	Central American Indian	Gallego	Mexicano	South American
Asturian	Chicano	Guatemalan	Nicaraguan	South American Indian
Balearic Islander	Colombian	Honduran	Panamanian	Spaniard
Bolivian	Costa Rican	La Raza	Paraguayan	Spanish Basque
Canal Zone	Criollo	Latin American	Peruvian	Uruguayan
Castilian	Cuban	Mexican	Puerto Rican	Valencian
				Venezuelan

Background

The New Mexico Decedent Image Database (NMDID) was created in 2020, and houses 15,243 full-body computed tomography (CT) scans and associated health and lifestyle data. [20] Contained within this database are the fields of race, ethnicity, and Hispanic subgroup. Next of kin were contacted to collect data not available in the medical examiner's database, including Hispanic subgroup. "Hispanic" is a variable in the ethnicity, race, and subgroup fields, reflecting responses from next of kin.

NMDID also contains information regarding age at death and manner of death. Age at death is the actual age each decedent was when they died. It provides a proxy for life expectancy, which can be used to compare among groups. Manner of death is one of five broad categories: accident, homicide, natural, suicide, and unknown. Manner of death code provides more specific information within one of these broad categories, such as pneumonia (natural) or automobile accident (accident). The decedents represented in NMDID with manner of death information (n=15,236) include 38.6% accidents, 34.6% natural, 7.4% homicides, 15.4% suicides, and 4% undetermined. Overall, 11% of deaths in New Mexico from mid-2010 to mid-2017 are included in the database. Excluded from the sample are most expected deaths, such as elderly persons in nursing homes, and physician-attended deaths. Thus, the analyses and results presented here are biased such that they underrepresent the very oldest in the living population, and over represent younger individuals.

Using this information, we have three aims: 1) determine whether there is a significant difference in age at death among the seven subgroups of Hispanics, 2) determine whether there is an association between manner of death and each Hispanic subgroup, and 3) determine whether there are significant associations between manner of death codes and Hispanic subgroups within the NMDID sample.

Materials and Methods

Only de-identified data from decedents was used to create NMDID, so no Institutional Review Board approval was required for this research. Approval from the research committee of the New Mexico Office of the Medical Investigator, where the data originate, was obtained in 2015.

NMDID was queried to find all decedents with a Hispanic subgroup identified. Hispanic subgroup identification was provided by the next of kin when contacted by NMDID staff for additional information. If next of kin identified the

decedent as Hispanic in either race or ethnicity, they were asked how the decedent would have identified their own subgroup. Subgroup names were provided to next of kin, and included: Chicano/a, Hispanic, Latino/a, Mexican, Mexican American, Nuevomexicano/a, and Spanish. There are 639 decedents of Hispanic ethnicity or race that also have Hispanic subgroup information, comprising 13% of Hispanics in the database. The subgroups and counts are available in Table 2. In addition, data were queried for age at death in years, manner of death, and manner of death code.

Table 2. Distribution of subgroups in the New Mexico Decedent Image Database

Hispanic Subgroup	Count
Chicano/a	99
Hispanic	196
Latino/a	46
Mexican	71
Mexican American	76
Nuevomexicano/a	28
Spanish	123
Total	639

The mean ages at death in years were compared among subgroups using ANOVA. We then ran a post hoc Tukey analysis to determine which groups were significantly different from one another. In addition, mean age between NHW and the subgroups were compared using the same methods.

Each decedent was assigned both a category manner of death (Accident, Homicide, Natural, Suicide, or Unknown) and one of 60 specific manner of death codes (such as pneumonia or automobile accident) by a forensic pathologist. We next determined whether there are associations between manner of death, both category and specific code, and subgroup membership. To account for the large sample size and high number of specific manner of death codes, we performed Monte Carlo simulations of the Fishers exact test with N=100,000 iterations. All analyses were performed in SAS Studio.[21]

Results

Age at death

The mean age in years for all 639 decedents is 45.45 years (SD:18.91), with range 0-99 years (with 12 decedents less than 10 years of age). The mean and standard deviation of age at death in years for the seven subgroups are listed in Table 3. For comparison, among the entire New Mexican population, mean age of death in 2017 was 81.2 years for women and 75.3 for men, demonstrating a bias toward younger age of death for this sample.[22] Within NMDID, the NHW (n=8,767) average age at death is 52.69 years (with 232 decedents under the age of ten).

Table 3. Mean age at death in years, with standard deviations, for the seven Hispanic subgroups in New Mexico.

Hispanic subgroup	Mean age in years	Standard deviation
Chicano/a	48.52	16.55
Hispanic	43.29	19.20
Latino/a	42.91	17.46
Mexican	43.05	20.33
Mexican American	41.58	21.00
Nuevomexicano/a	48.90	15.69
Spanish	50.34	18.18

A significant difference was found in the mean age at death in years among the seven Hispanic subgroups (p=0.0029). See Figure 1A for the distribution of ages at death among the subgroups. There are significant differences in the Spanish-Hispanic (p=0.019) and Spanish-Mexican American comparisons (p= 0.0236), with

Spanish dying seven years later than Hispanics and eight years later than Mexican Americans. When the seven subgroups were compared to NHW in NMDID, a significant difference was found ($p<0.0001$) in NHW-Hispanic ($p<0.0001$), NHW-Latino/a ($p=0.0143$), NHW-Mexican ($p=0.0008$), and NHW-Mexican American ($p<0.0001$) comparisons. Only Chicano/a, Nuevomexicano/a, and Spanish did not differ significantly from NHW.

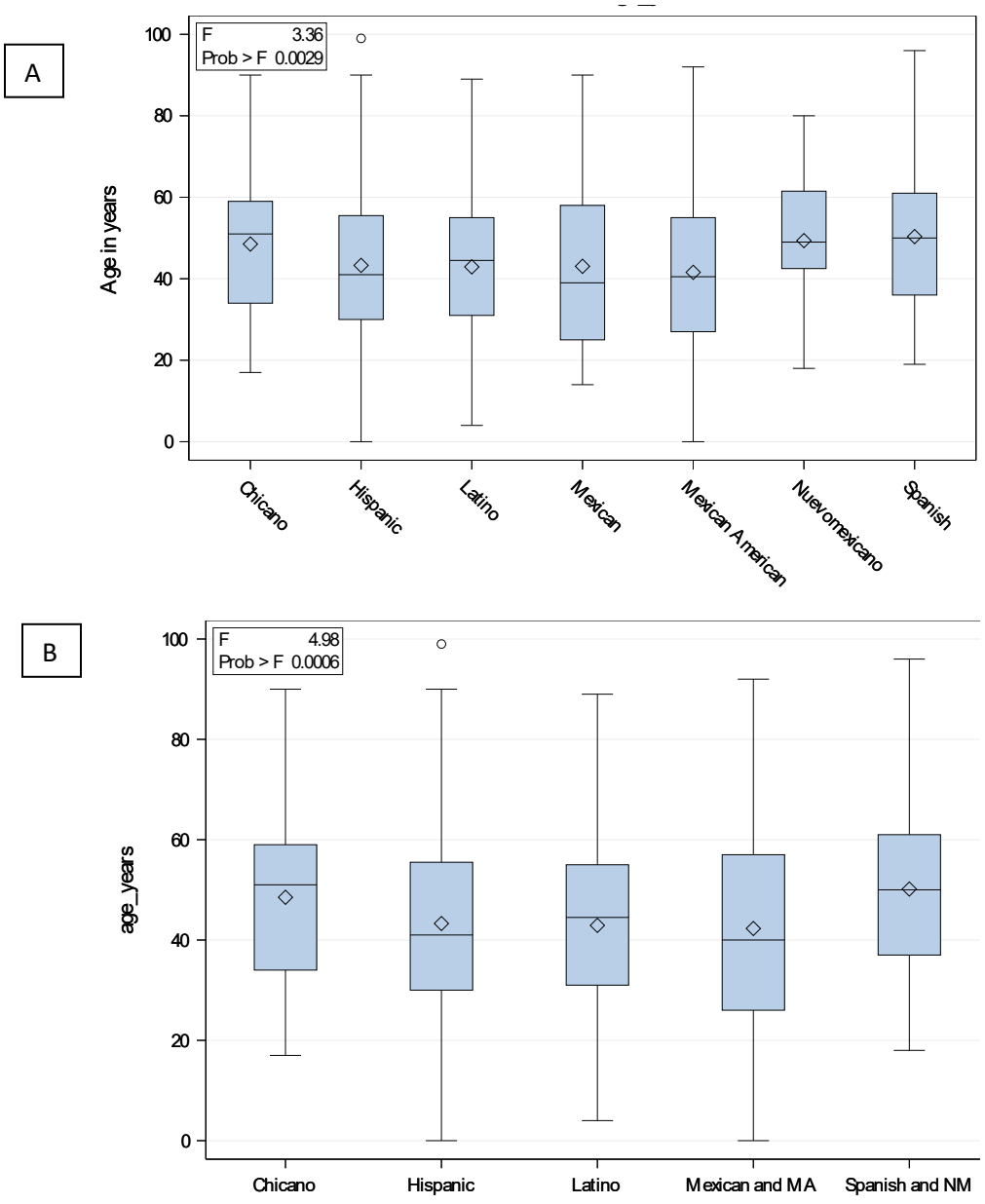


Figure 1. A: Boxplots of age at death between the 7 Hispanic subgroupings in New Mexico. B: Boxplot of age at death between Hispanic subgroups when similar groups are combined.

Hunley et al.,[12] describes similarities of history and migration among some of these subgroups. Following these descriptions and to further explore the importance of subgroup specificity, Mexican American and Mexican subgroups were combined, as were Nuevomexicano/a and Spanish. Differences in mean age at death are even more clear when these subgroups are considered together ($p=0.0006$). Significant differences are found between Spanish/Nuevomexicano/a and Mexican/Mexican American ($p= 0.0028$) and Spanish/Nuevomexicano/a and Hispanic ($p= 0.0065$). See Figure 1B for the distribution of age at death when similar subgroups are combined.

Manner of Death

There are fewer homicides among Spanish, and fewer suicides among Chicano/as than expected. However, there is no significant relationship between subgroup and manner of death, when all subgroups are considered independently (p=0.3465) or related subgroups are combined (p=0.2223). See Table 5 for the contingency table of manner of death in all groups.

Table 5. Manner of death groupings by Hispanic subgroupings in New Mexico. Frequency of each cell is above, with expected count below.

Manner of Death	**Hispanic subgroup**							
Frequency (n) Expected	**Chicano /a**	**Hispanic**	**Latino /a**	**Mexican**	**Mexican American**	**Nuevomexicano /a**	**Spanish**	**Total**
Accident	54 43.69	83 86.498	17 20.3	28 31.333	33 33.54	13 12.357	54 54.282	282
Homicide	13 10.225	21 20.244	5 4.7512	11 7.3333	9 7.8498	1 2.892	6 12.704	66
Natural	23 29.746	61 58.892	15 13.822	20 21.333	21 22.836	10 8.4131	42 36.958	192
Suicide	6 11.775	25 23.311	6 5.471	11 8.4444	8 9.0391	2 3.3302	18 14.629	76
Undetermined	3 3.5634	6 7.0548	3 1.6557	1 2.5556	5 2.7355	2 1.0078	3 4.4272	23
Total	99	196	46	71	76	28	123	639

Manner of death code

There are 60 manner of death codes used in this sample, ranging from natural, various automobile accidents, drug and alcohol use, to falls (see Table 4 for list of specific codes and total counts in sample). A significant association was found between cause of death code and subgroup (p=0.0364). In contingency table (not shown), more Chicano/as died of drug use, more Spanish and fewer Mexicans died of prescription pill overdose, more Spanish died of exposure, and fewer Spanish died of gun shots than expected. When similar subgroups are combined, the significance increases (p=0.0123).

Table 4. Specific Manner of Death Codes in Sample

Manner code	Count
ASCVD/COPD/Seizure disorder	1
Accident-specify	3
Asphyxia/airway obstruction/suffocation	5
Beaten by assailant(s)	9
Bitten/mauled/stung/kicked by (bee, dog, snake, horse, (name agent))	1
Blunt trauma/multiple injuries/subdural hematoma	1
Choked on (bolus of food, toy, etc.)	2
Contacted electrical current via (outlet, telephone lines, ungrounded chain saw)	1
Crushed/suffocated by (car falling from jack, plastic bag over head)	6
Cyclist accident	3

Driver of auto in collision with (auto, pickup, truck, minivan, ATV, motorcycle, (other motor vehicle type))	10
Driver of auto that left roadway (and overturned and/or became pinned underneath or in auto)	3
Driver of motorcycle (explain circumstances briefly, e.g. left roadway and struck tree, overturned in roadway, etc.)	5
Driver of motorcycle in collision with (auto, pickup, truck, minivan, ATV, motorcycle, (other motor vehicle type))	6
Driver of pickup that left roadway (and overturned and/or became pinned underneath or in auto)	5
Driver of truck that left roadway (and overturned and/or became pinned underneath or in truck)	1
Drowned in (tub, arroyo, pool, (this includes all non-recreational water accidents))	5
Drowned while swimming (this includes recreational swimming and rescue attempts)	2
Fall from (chair, table, mesa, cliff)	8
Fall from height/same height	1
Fall from standing height	8
Gunshot wound	2
Hanged self	28
Homicide-specify	3
Ingested alcohol (ethanol)	9
Ingested and/or injected illicit drug(s) - (in combination with ethanol)	94
Ingested and/or injected prescription medications	54
Ingested or injected medication	5
Ingested, injected or inhaled non-prescription medication (illicit, volatiles)	4
Inhaled toxic substance (toxic substances inhaled accidentally)	3
Inhaled toxic substance - (toxic substances abused to achieve intoxication)	2
Jumped from	1
Natural	192
Neglect/Starvation	1
Passenger .in (airplane, balloon, hang glider, (other aircraft type) that crashed) - (Also parachutist)	2
Passenger in auto in collision with (auto, pickup, truck, minivan, ATV, motorcycle, (other motor vehicle type))	7
Passenger in auto in collision with (tree, embankment, rock, wall, (other fixed object))	1
Passenger in auto that left roadway (and overturned and/or became pinned underneath or in auto)	6
Passenger in pickup in. collision with (auto, pickup, truck, minivan, ATV, motorcycle, (other motor vehicle type))	1
Passenger on motorcycle in collision with (auto, pickup, truck, minivan, ATV, motorcycle, (other motor vehicle type)	1
Pedestrian homicide (ie, Struck with auto by assailant(s)	1
Pedestrian struck by (auto, pickup, truck, train, (other motor vehicle type))	17
Pedestrian struck by (bicycle, (other non-motor vehicle))	2
Pedestrian struck by Motor vehicle	2
Pilot of (airplane, balloon, hang glider, (other aircraft type) that crashed)	1
Received blow/collided with	1
Remained outdoors exposed to (cold, heat) - ((while intoxicated))	6
Shot by assailant(s) with firearm	42
Shot self with firearm	34
Skeletal, mummified or decomposed remains	4
Slashed with	1
Stabbed by assailant(s)	7
Strangled by assailant(s)	3
Suffocated self with	1

Suicide-specify	1
Suicide as pedestrian	1
Undetermined after autopsy and/or toxicology	4
Undetermined-specify	3
Victim of (car bomb, letter bomb, (type of device)) explosion	1
Victim of (house, car, trailer, open range, (other site)) fire	5

Discussion

Hispanic is a panethnic term that is used to describe individuals with origins from the Caribbean, Spain, South and Central America, etc. As such, it includes individuals with a wide variety of geographical, economic and social backgrounds. [14,23,24] Hispanics are often grouped together for biomedical research, but do not represent a homogenous ethnicity. [14] In this study we used age at death, manner of death, and manner of death code to determine whether there are differences among culturally and historically relevant subgroups in New Mexican Hispanics. Previous studies have shown differences in health, mortality, fertility and birth rates using Hispanic subgroupings with vocabulary standards defined by the HHS or a select subset. [5–8]

Results show a significant difference in age at death between the Spanish and Mexican American and Spanish and Hispanic subgroups, with the Spanish dying seven to eight years later than these comparison groups, respectively. These differences are even more notable when culturally and historically similar groups are combined. These results indicate the value of understanding of the culture and history from which ethnic terminology standards develops. For example, individuals may use the term Nuevomexicano/a to indicate their long history in the state and their Spanish decent. [25] Given this background, it is not surprising that differences in age at death are made clearer when Spanish and Nuevomexicano/a groups are combined. This is because "Spanish" as a self-identifier is used by older individuals and "Nuevomexicano/a" by younger people with similar backgrounds.[13] It appears that, while ethnic terminologies change over time, the challenges affecting age at death remain consistent.

The age at death of NHW drawn from the same medicolegal sample is significantly higher than that of the seven Hispanic subgroups, with Hispanics, Latino/as, Mexicans, and Mexican Americans all dying younger than NHWs. Spanish and Nuevomexicano/a, groups that are traditionally considered privileged in the state, do not die significantly younger than do NHW.[25,26] These results do not lend support for the Hispanic Health Paradox, as Mexicans and Mexican Americans, who have been shown to be more likely born outside of New Mexico or descended from more recent immigrants, are dying at a younger average age than NHWs.[13]

While there are no significant associations between manners of death and subgroups there is an association between specific manner of death codes and subgroups. More Chicano/as die of drug use, more Spanish and fewer Mexicans die of prescription pill overdose, more Spanish die of exposure, and fewer Spanish die of gun shots than expected if manner of death codes were unrelated to subgroup membership. Recognition of this heterogeneity can be used for targeted interventions, related to licit and illicit drug use, alcohol use, and gun safety.

These results make clear that OMB and HHS terminology standards are not specific nor culturally relevant enough for research into health and mortality research, at least not among Hispanics in New Mexico. The NAM standard, although not used widely, identifies and provides terminology for more specific regional variation. If the OMB or HHS standards had been used in this analysis, we would have missed heterogeneity important for understanding public health, especially between the combined Spanish and Nuevomexicano/a versus Mexican and Mexican American groups. These results indicate that researchers should strive to capture more regionally relevant ethnicity data on subjects. This will likely require large samples and the introduction of ethnological methods into the development of ethnic terminologies and vocabulary standards.

Conclusion

There are significant differences in ages at death between Spanish and Mexican American subgroups of Hispanics in New Mexico. In addition, specific manner of death codes are associated with seven ethnic subgroups. Without

terminology to reflect the variation in ethnicity collected in this database, important mortality data is lost. The subgroup terms used in this study come from an understanding of the culture and history of place, and reflect different migration histories among the different subgroups. [13] The results indicate that it is of utmost importance to define the subgroupings of Hispanic beyond that suggested by HHS and OMB, when studying health and mortality. These results support the contention made by several other authors that important data regarding health patterns can be lost if Hispanics are treated as a homogeneous group.[5–8]

The results also argue against the Hispanic Paradox, as groups with more recent migration histories (Mexicans and Mexican Americans) die at younger ages than those with long histories in the United States (Spanish and Nuevomexicano/a). If subgroups are used when studying Hispanic health, it may elucidate different outcomes than has been captured with a panethic term, clarifying the "paradox."

How can culturally relevant, specific group differences be captured across the United States? Some vocabulary standards, like HHS, include the most common the subgroups within major regions of the United States. The NAM report suggested even further specificity within Hispanic ethnicity, naming 36 different groups, all with the common thread of having a Spanish-speaking background. The current study included additional subgroups specific to New Mexico and not included in the NAM report. The fact that results were strengthened when some groups were combined indicates the need for continued reflexive analysis, as terminologies are not natural features of populations, but rather ephemeral categories relevant only at particular times, in particular geographies.

Although we recommend using the NAM as a core terminology standard for ethnicity, it is missing a vital subset that many Hispanics in New Mexico use to describe themselves: Hispanic. In addition, other regional terminologies may become apparent when ethnological research is conducted in different states and regions. Therefore, we recommend using the NAM ethnicity standard with the addition of the subgroup of Hispanic, perhaps with additional, region-specific terms.

References

1. Office of Management and Budget. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. *Fed Regist* 1997; 62: 58782–58790.

2. Colen CG, Ramey DM, Cooksey EC, et al. Racial disparities in health among nonpoor African Americans and Hispanics: The role of acute and chronic discrimination. *Soc Sci Med* 2018; 199: 167–180.

3. McClelland S, Perez CA. The pervasive crisis of diminishing radiation therapy access for vulnerable populations in the United States—part 3: Hispanic-American patients. *Adv Radiat Oncol* 2018; 3: 93–99.

4. Weinick RM, Jacobs EA, Cacari Stone L, et al. Hispanic Healthcare Disparities: Challenging the Myth of a Monolithic Hispanic Population Hispanic Healthcare Disparities Challenging the Myth of a Monolithic Hispanic Population. 2004; 42: 313–320.

5. Fenelon A, Chinn JJ, Anderson RN. A comprehensive analysis of the mortality experience of hispanic subgroups in the United States: Variation by age, country of origin, and nativity. *SSM - Popul Heal* 2017; 3: 245–254.

6. Sutton PD, Mathews TJ. Birth and fertility rates for states by Hispanic origin subgroups: United States, 1990 and 2000. *Vital Health Stat 21* 2006; 1–4, 6–95.

7. Hummer RA, Rogers RG, Amir SH, et al. *Adult Mortality Differentials among Hispanic Subgroups and Non-Hispanic Whites*. 2000.

8. Garcia MA, Garcia C, Chiu C-T, et al. A Comprehensive Analysis of Morbidity Life Expectancies Among Older Hispanic Subgroups in the United States: Variation by Nativity and Country of Origin. *Innov Aging*; 2. Epub ahead of print 1 June 2018. DOI: 10.1093/geroni/igy014.

9. HHS Implementation Guidance on Data Collection Standards for Race, Ethnicity, Sex, Primary Language,

and Disability Status | ASPE, https://aspe.hhs.gov/basic-report/hhs-implementation-guidance-data-collection-standards-race-ethnicity-sex-primary-language-and-disability-status (accessed 11 August 2020).

10. Beck JD, Offenbacher S. Systemic Effects of Periodontitis: Epidemiology of Periodontal Disease and Cardiovascular Disease. *J Periodontol* 2005; 76: 2089–2100.

11. Ulmer C, McFadden B, Nerenz DR. *Race, ethnicity, and language data: Standardization for health care quality improvement*. National Academies Press, 2009. Epub ahead of print 30 December 2009. DOI: 10.17226/12696.

12. Hunley K, Edgar H, Healy M, et al. Social Identity in New Mexicans of Spanish-Speaking Descent Highlights Limitations of Using Standardized Ethnic Terminology in Research. *Hum Biol* 2017; 89: 28.

13. Healy ME, Hill D, Berwick M, et al. Social-group identity and population substructure in admixed populations in New Mexico and Latin America. *PLoS One* 2017; 12: e0185503.

14. Valles SA. The challenges of choosing and explaining a phenomenon in epidemiological research on the '"Hispanic Paradox"'. Epub ahead of print 2016. DOI: 10.1007/s11017-015-9349-1.

15. Markides KS, Eschbach K. Hispanic Paradox in Adult Mortality in the United States. 2011, pp. 227–240.

16. Smith DP, Bradshaw BS. Rethinking the Hispanic paradox: Death rates and life expectancy for US non-Hispanic White and Hispanic populations. *Am J Public Health* 2006; 96: 1686–1692.

17. Crimmins EM, Kim JK, Alley DE, et al. Hispanic Paradox in Biological Risk Profiles. *Am J Public Health* 2007; 97: 1305–1310.

18. Palloni A, Morenoff JD. *Interpreting the Paradoxical in the Hispanic Paradox Demographic and Epidemiologic Approaches*, https://deepblue.lib.umich.edu/handle/2027.42/73920 (2001, accessed 25 August 2020).

19. Tarraf W, Jensen GA, Dillaway HE, et al. Trajectories of Aging among U.S. Older Adults: Mixed Evidence for a Hispanic Paradox. *Journals Gerontol - Ser B Psychol Sci Soc Sci* 2020; 75: 601–612.

20. Edgar, HJH; Daneshvari Berry, S; Moes, E; Adolphi, NL; Bridges, P; Nolte K. New Mexico Decedent Image Database. *Office of the Medical Investigator, University of New Mexico*. Epub ahead of print 2020. DOI: doi.org/10.25827/5s8c-n515..

21. SAS. SAS Studio.

22. NM-IBIS - Complete Health Indicator Report - Life Expectancy From Birth, https://ibis.health.state.nm.us/indicator/complete_profile/LifeExpectBirth.html (accessed 26 August 2020).

23. Mora G. *Making Hispanics: How activists, bureaucrats, and media constructed a new American*. 2014.

24. Liz J. The categorization of Hispanics in biomedical research: US and Latin American perspectives. *Philos Compass* 2020; 15: e12655.

25. Casandra D. Salgado. Mexican American Identity: Regional Differentiation in New Mexico. *Sociol Race Ethn* 2020; 6: 179–194.

26. Gonzales PB. *The Political Construction of Latino Nomenclatures in Twentieth-Century New Mexico*. 1993.

Leveraging Spatial Information in Radiology Reports for Ischemic Stroke Phenotyping

Surabhi Datta, MS[1], Shekhar Khanpara, MD[2], Roy F. Riascos, MD[2], Kirk Roberts, PhD[1]
[1]School of Biomedical Informatics, [2]McGovern Medical School
The University of Texas Health Science Center at Houston, Houston, TX

Abstract

Classifying fine-grained ischemic stroke phenotypes relies on identifying important clinical information. Radiology reports provide relevant information with context to determine such phenotype information. We focus on stroke phenotypes with location-specific information: brain region affected, laterality, stroke stage, and lacunarity. We use an existing fine-grained spatial information extraction system–Rad-SpatialNet–to identify clinically important information and apply simple domain rules on the extracted information to classify phenotypes. The performance of our proposed approach is promising (recall of 89.62% *for classifying brain region and* 74.11% *for classifying brain region, side, and stroke stage together). Our work demonstrates that an information extraction system based on a fine-grained schema can be utilized to determine complex phenotypes with the inclusion of simple domain rules. These phenotypes have the potential to facilitate stroke research focusing on post-stroke outcome and treatment planning based on the stroke location.*

Introduction

Ischemic stroke (IS) accounts for around 87% of all strokes in the United States[1]. Clinical trials and epidemiological studies targeted toward investigating communication, cognitive, and emotional changes after stroke are interested in analyzing specific subsets of patient records pertaining to certain characteristics of IS for treatment and prognosis research. Radiological findings documented in head computed tomography (CT) and brain magnetic resonance imaging (MRI) reports provide important information to develop IS phenotypes. Understanding and identifying various clinically important information from the report text can facilitate in constructing fine-grained phenotypes. In this work, we propose to utilize spatial information in the reports to construct IS phenotypes. We develop and evaluate a natural language processing (NLP) pipeline for IS phenotyping by using spatial information extracted from the reports. More specifically, we use the spatially-related imaging features and their brain locations as well as the potential diagnoses information to classify the phenotypes.

Effects of stroke in a patient are dependent on the areas of the brain affected[2,3]. Based on the side and the particular location of the stroke, different body functions are impaired. For example, stroke in the right side of cerebral hemisphere results in *left-sided weakness or paralysis*, *visual*, and *spatial problems*, stroke in the cerebellum manifests in a different set of effects such as *ataxia*, *dizziness*, *nausea*, and *vomiting*, whereas stroke located in the brainstem results in problems associated with *breathing*, *balance*, and *coma*. Moreover, the effects can be further specified based on the particular lobe of the cerebral hemisphere that is affected. For example, *sensation* and *spatial awareness* are impacted with stroke in the parietal lobe whereas *language* and *memory* are impaired with stroke in the temporal lobe. A previous work[4] has demonstrated that location of stroke infarct influences the functional outcome following an ischemic stroke as measured by modified Rankin Scale, a commonly used scale for rating stroke outcome in clinical trials. Further, a few studies[5,6] have focused on the brain locations affected by stroke for improving treatment of post-stroke depression and predicting post-stroke language outcome. Therefore, categorizing imaging reports according to stroke location–or in other words, constructing phenotypes incorporating the stroke location–holds potential benefits for clinical research studies that focus on targeted treatment based on the specific brain region affected.

We construct the IS phenotypes by using the brain location information in the reports both directly and indirectly. Direct use refers to including the side and the specific brain region affected by stroke in the phenotypes. Indirect use of location includes deriving other crucial information such as stroke stage based on the particular brain region a certain imaging feature is detected. Besides these, we also use the IS-related potential diagnoses information directly in the phenotypes to extract the stroke stage in cases when it is included as part of the diagnosis phrase (e.g., *subacute* stage in the diagnosis phrase '*subacute infarction*').

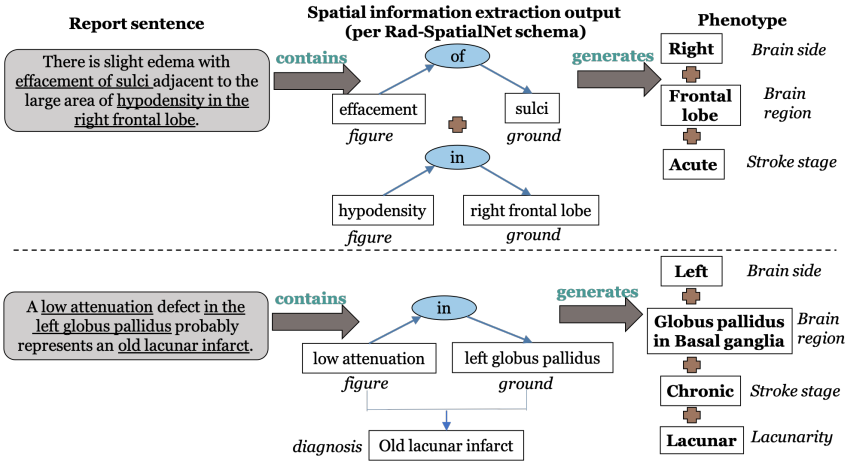


Figure 1: Examples of stroke phenotypes using spatial relations from reports. Blue ovals contain spatial triggers.

Sentences from a report

There is an **acute infarction** in the **lateral aspect** of the **right cerebellar hemisphere**. There are several small **acute infarctions** in the **right midbrain**. There is **encephalomalacia** and **gliosis** in the **inferior left cerebellar hemisphere**.

Brain side	Brain region	Stroke stage	Lacunarity
Right	Cerebellar hemisphere	Acute	No
Left	Cerebellar hemisphere	Chronic	No
Right	Midbrain	Acute	No

Phenotypes containing brain region-specific information

Figure 2: Granular phenotypes considered in this work (shown for a sample report).

Consider the two examples shown in Figure 1 from head CT reports. The first sentence captures information corresponding to mass effect like *sulcal effacement* along with imaging feature such as *cortical hypodensity* that helps to indicate that an *infarction* is *acute*. The second sentence detects an area of *low attenuation* in the left side of *globus pallidus*, part of *basal ganglia*. The sentence also describes that this finding indicates that the infarct is *lacunar* and thus *chronic*. Therefore, for the first example, we see that spatial relations between imaging features and brain locations (as indicated by phrases like '*effacement of sulci*' and '*hypodensity in the right frontal lobe*') encode important radiological information that facilitates in determining the diagnosis (i.e., *infarction*) and its stage (i.e., *acute*). Also, note that although *acute* is not mentioned explicitly in this sentence, identifying the spatial relations help in inferring that the stroke is *acute*. Thus, spatial relations present in imaging reports can directly be utilized for constructing stroke phenotypes containing fine-grained location information along with additional derived information like stroke stage. We, therefore, use our previously proposed spatial representation schema–Rad-SpatialNet[7] to extract spatial information from reports which can subsequently be used for extracting important IS phenotypes.

Prior studies have attempted to extract IS-related information from radiology reports. Wheater et al.[8] developed brain imaging phenotypes, however, these phenotypes lacked specificity in the brain location information and were classified as only cortical or deep. Other works identified reports with acute IS[9,10] and silent brain infarcts[11]. However, these studies focused on limited information like classifying reports based on presence/absence of IS, acuity, and middle cerebral artery (MCA) territory involvement. Alternatively, we aim to construct specific stroke phenotypes containing more granular information for each stroke affected brain area and this makes the task more complex compared to performing binary classification of the reports. We illustrate the granularity and complexity of our phenotypes in Figure 2. Note that the phenotypes consider information at the level of both side and region of the brain affected. Thus we see the stage is *acute* for right cerebellar hemisphere and *chronic* for the left side.

Therefore, using spatial information from the reports forms an intuitive way to extract such fine-grained information for constructing the phenotypes. In this paper, we define the fine-grained stroke phenotypes described above with input from radiology experts. For automatic labeling of the reports with the relevant phenotypes, we first identify the spatial relations using a transformer-based model (BERT[12]) for each report. We then apply rules based on domain knowledge on the extracted spatial information to classify the phenotypes. Finally, we evaluate our system by comparing the automatically generated phenotypes with the gold phenotypes for a set of head CT and brain MRI reports. Thus the main contributions of our work include:

- Classify fine-grained ischemic stroke phenotypes by applying simple domain rules on top of spatial information extracted from neuroradiology reports.
- Phenotypes contain information targeted at the level of a specific side and region of the brain affected.

Related Work

Numerous works have focused on identifying certain subgroups of stroke patients using NLP techniques with the aim to facilitate timely patient triaging to select appropriate group of patients highly likely to encounter severe consequences. We highlight the relevant studies by categorizing them in the following three subsections:

Identifying stroke/ischemic stroke Sedghi et al.[13] converted medical narratives to codified text based on expert provided sign and symptom phrases and they applied ML algorithms on the codified sentences to predict the presence of stroke in a patient. Majersik et al.[14] applied NLP-based approaches by adding context to n-grams that classified ischemic, hemorrhagic, and non-stroke cases with high precision by using different combination of clinical report types. A study by Kim et al.[10] utilized document-feature matrix vectorization techniques to classify brain MRI reports for identifying acute ischemic stroke. Govindarajan et al.[15] developed ML-based NLP approaches to identify whether the stroke is ischemic or hemorrhagic based on some pre-defined symptoms and patient factors.

Classifying stroke subtypes Two studies focused on automatically classifying stroke patients based on standard stroke subtype classification systems–the Trial of Org 10172 in Acute Stroke Treatment (TOAST) and the Oxfordshire Community Stroke Project (OCSP). Garg et al.[16] developed ML-based approaches to classify patients according to the TOAST ischemic stroke subtyping using neurology progress notes and neuroradiology reports for better patient management and outcome prediction. Sung et al.[17] constructed features based on the medical entities identified by MetaMap and then applied traditional ML techniques to classify stroke patients based on four clinical syndromes taken from OCSP classification system that considers the anatomical location of stroke.

Identifying stroke features A recent study by Ong et al.[9] classified radiology reports based on three outcomes - presence of stroke, involvement of MCA location, and stroke acuity by using text featurization methods such as bag of words, term frequency-inverse document frequency, and GloVe. These are considered as three separate classification tasks and they employed traditional ML models and recurrent neural networks to predict the outcomes. On the other hand, we aim to construct more specific phenotypes (e.g., 'acute right frontoparietal stroke') which can be fairly easily developed from more general information extracted from the reports (e.g., extracting 'hypoattenuation' in right frontoparietal distribution as well as identifying 'effacement of sulci' in the same report).

Most of the important information, especially those describing or relating to abnormal findings, are mentioned as part of the spatial descriptions between brain imaging observations and their corresponding anatomical structures. Often times, determining granular phenotypes is dependent on these specific information documented in the reports. Fu et al.[11] developed both rule-based and ML methods to identify incidental silent brain infarct and white matter disease patients from the EHRs. As reported in Fu et al.'s work, some of the false positive errors generated by the ML-based text classification system are usually contributed by certain disease locations (e.g., right occipital lobe) that often co-exist with expressions related to the disease/outcome of interest (e.g., silent brain infarct in their case). Thus, developing a set of constraints using domain knowledge on the spatial information in the reports has the potential to diminish such false positive cases. Moreover, developing constraints based on the spatial relationships between imaging observations and anatomical locations forms a natural way to predict a stroke-associated outcome of interest. This also enhances the interpretability of the automatic phenotype construction system as it closely replicates a clinician's workflow to select eligible group of patients for treatment plans and clinical recommendations.

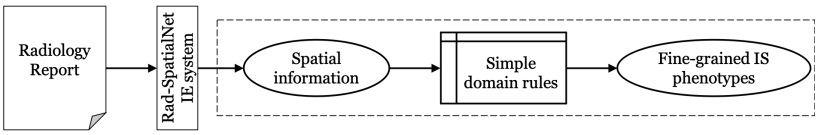


Figure 3: Pipeline for ischemic stroke (IS) phenotype classification. Dashed box indicates the main contribution of this work. IE - information extraction.

Wheater et al.[8] developed a rule-based NLP system to automatically label neuroimaging reports with a pre-defined set of 24 phenotypes. Their system incorporates manually crafted domain lexicons as well as a chunking step for extracting the radiological entities and relations from the text. Simple rules are then developed based on the presence of certain entities and relations to construct the final labels for each report. Inspired by this, we also develop the phenotype construction step similar to Wheater et al. However, our initial step of information extraction is more spatial information-oriented where we extract some common radiographic information by using advanced transformer-based language model and thus avoid the tedious process of developing manual rules for entity and relation extraction.

Thus the focus of our work is to demonstrate how automatically extracted important spatial information from neuroimaging reports can potentially be used to develop granular ischemic stroke phenotypes. To our knowledge, this is a first attempt toward using radiographic information connected through spatial trigger expressions in radiology reports for stroke phenotyping.

Materials and Methods

We use the output of a spatial information extraction (IE) system (information represented following the Rad-SpatialNet schema) to classify the granular ischemic stroke phenotypes. A set of simple domain rules are applied on the output of the IE system for classifying the phenotypes. The following sections describe the dataset along with a brief overview of the Rad-SpatialNet schema used in this study. This is followed by descriptions of the phenotype annotation process and our proposed pipeline for ischemic stroke phenotyping. An overview of our approach is shown in Figure 3.

1 Dataset

We select a set of 150 MIMIC reports (containing a mix of brain MRIs and head CTs) to classify the ischemic stroke phenotypes. These 150 reports contain at least one of the ICD-9 ischemic stroke-relaed diagnosis codes from 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11, 434.91, and 436. We refer to this phenotyping dataset as RAD-IS-P. To train our spatial information extraction (IE) model, we use 400 MIMIC-III[18] radiology reports (consisting of chest X-rays, brain MRIs, and babygrams) annotated following Rad-SpatialNet schema as part of our earlier work[7]. Since we extract stroke phenotypes from both types of neuroradiology reports, i.e. MRIs and CTs, we annotated a few (15) head CT reports following the same schema to add to the training data for our spatial IE system. Thus, we use the combined set of 415 reports for training the IE model. We refer to this dataset as RAD-SPATIAL-IE.

2 Rad-SpatialNet schema

This spatial representation schema has been proposed in our previous work[7]. We use this schema to extract spatial information from the RAD-IS-P data. According to this, a spatial frame is constructed for each spatial description mentioned in a radiology report sentence. The spatial trigger forms the lexical unit of a spatial frame and the other related important clinical contextual information constitutes the frame elements. So for the sentence–'*Hypodensity is noted in the pons which likely represents a lacunar infarct*', a spatial frame is instantiated by the spatial trigger '*in*' and the elements associated to this lexical unit are Figure ('*hypodensity*'), Ground ('*pons*'), Hedge ('*likely represents*'), and Diagnosis ('*lacunar infarct*'). The frame elements that are not present in this example sentence but are part of the Rad-SpatialNet schema include Relative Position, Distance, Position Status, Reason, and Associated Process. We apply domain rules on Figure, Ground, and Diagnosis elements extracted from the reports to classify stroke phenotypes.

3 Ischemic stroke phenotype annotation

Each MRI and CT report is annotated with important IS features as validated by a practicing radiologist. These features are identified based on both their clinical importance as well as taking into account the types of information covered in Rad-SpatialNet schema. The pre-defined features are described as follows:

Brain region affected	Frequency	Brain region affected	Frequency
Cerebral hemisphere	26	Basal ganglia	38
Cerebral hemisphere - Frontal lobe	61	Thalamus	6
Cerebral hemisphere - Occipital lobe	30	Cerebral peduncle	2
Cerebral hemisphere - Parietal lobe	46	Internal/External capsule	8
Cerebral hemisphere - Temporal lobe	29	Corona radiata	4
Cerebellum	35	Insula	15
Brainstem	9	Watershed	4

Table 1: Annotated phenotypes per brain region.

1. Brain side - the laterality of the brain that is affected
2. Brain region - refers to the specific brain area affected due to reduced blood and oxygen supply
3. Stroke stage - three main stages used to describe the CT manifestations of stroke: acute, subacute, and chronic (as described in Birenbaum et al.[19]). Additionally, some reports document the stage information as acute/subacute, so we also consider acute/subacute separately
4. Lacunarity - whether infarct is lacunar or not. Lacunar infarcts are usually small noncortical infarcts (diameter of 0.2 to 15 mm) and are caused by occlusion of a small perforating artery

Multiple combinations of these four features can be present in a report. In such cases, we label each report with a maximum of five combinations of brain side, region, stroke stage, and lacunarity. For the example in Figure 2, the resulting feature combinations used for annotating the report are – 1. right, cerebellum, acute, not lacunar, 2. left, cerebellum, chronic, not lacunar, and 3. right, brainstem (midbrain), acute, not lacunar. Another point to note is that if the stroke stage is directly available as part of the spatial information extracted from the report, we use that information to annotate the report, otherwise the stage annotation is determined based on certain additional conditions/domain constraints applied over the extracted spatial information. For example, in the sentence "*There are several small acute infarctions in the right midbrain*" in Figure 2, *acute* was directly available as part of the Figure frame element *acute infarctions* identified in context to the spatial trigger *in*. However, in the last sentence, the stage is annotated as *chronic* because of the presence of terms like *encephalomalacia* and *gliosis*. Using this annotation scheme, the RAD-IS-P dataset was annotated with the stroke phenotypes by a radiologist (SK). A brief statistics of the brain region-wise phenotype annotations are shown in Table 1.

4 Proposed Pipeline

We describe the sequential stages of our phenotype extraction system in the following sections.

4.1 Spatial information extraction

We use an existing BERT-based sequence labeling system for extracting the spatial information from the reports[7]. This includes identifying the spatial triggers in a sentence followed by identifying the associated frame elements for each extracted trigger. Both spatial trigger and frame element extraction are framed as sequence labeling task. The frame elements identified by the BERT system for each of the spatial triggers in a sample head CT report sentence are illustrated in Figure 4. Specifically, in this work, we re-train the BERT-based frame element extractor using the RAD-SPATIAL-IE data with updated annotation spans for a few frame elements as described below.

Updates to Rad-SpatialNet for Ground and Diagnosis frame elements Note that for each anatomical location phrase labeled as Ground element in our previous work[7], the associated laterality terms such as 'left', 'right', and 'bilateral' were annotated as elements in context to that anatomical radiological entity. Similarly, for some of the potential diagnoses labeled as Diagnosis element, the associated temporal descriptors such as 'acute', 'evolving', and 'chronic' were also annotated as elements in context to the diagnosis radiological entity. Thus, the laterality and the temporal descriptor terms were not part of the Ground and Diagnosis frame elements respectively (in turn not directly connected to the spatial triggers) and thus were not identified by the spatial frame element extraction system. However, considering the need to capture laterality and diagnosis temporality information for our phenotyping task, we updated the mention spans of the Ground and Diagnosis elements in the sentences to support this work. Consider the following examples:

1. Include the laterality of the anatomical location
   Rad-SpatialNet[7] – *There is hypodensity in the left basal ganglia.*
   This paper – *There is hypodensity in the left basal ganglia.*

Spatial trigger (lexical unit)	Frame elements
Spatial Frame - 1	
in	Figure (*areas of restricted diffusion*) Ground (*vascular territory*) Hedge (*suggesting*) Diagnosis (*thromboembolic ischemic changes*)
Spatial Frame - 2	
of	Figure (*vascular territory*) Ground (*right MCA*)
Spatial Frame - 3	
on	Figure (*hyperintense foci*) Ground (*right occipital lobe, right basal ganglia*) Hedge (*suggesting*) Diagnosis (*thromboembolic ischemic changes*)
Spatial Frame - 4	
on	Figure (*hyperintense foci*) Ground (*right temporal lobe*) Relative Position (*distally*) Hedge (*suggesting*) Diagnosis (*thromboembolic ischemic changes*)

Figure 4: Spatial frames extracted for a sample sentence–*There are areas of restricted diffusion **in** the vascular territory **of** the right MCA, also some scattered hyperintense foci noted **on** the right occipital lobe, right basal ganglia and distally **on** the right temporal lobe suggesting thromboembolic ischemic changes.*

2. Include laterality and location descriptor whose span falls in between a laterality phrase and the anatomy phrase
   Rad-SpatialNet[7] – *A small area of white matter hyperintensity in the right frontal subcortical region.*
   This paper – *A small area of white matter hyperintensity in the right frontal subcortical region.*
3. Include the temporality of the potential diagnosis
   Rad-SpatialNet[7] – *Hypoattenuation in the right frontoparietal distribution consistent with acute infarction.*
   This paper – *Hypoattenuation in the right frontoparietal distribution consistent with acute infarction.*

In the first example we see that '*left*' has been included in the Ground element, and in the second example both '*right*' and '*frontal*' are included in the Ground element span. In the third example,'*acute*' is included in the Diagnosis element span. The spatial trigger (lexical unit for a spatial frame) is '*in*' for all the examples.

4.2 Automatic IS phenotype extraction

For each report, we use rules on top of the output of the BERT-based element extractor to automatically classify the phenotypes. We combine the spatial frames identified by the element extractor at the report level. We also keep a track of all the spatial frames predicted by the BERT extractor for each sentence in a sequential order (the order in which the spatial triggers appear in a sentence). This helps to combine the frames when the Ground element associated to a trigger is same as the Figure element of the next trigger. For example, in "*acute infarction in the lateral aspect of right cerebellum*", IS-related finding (*infarction*) is connected to the corresponding location (*right cerebellum*) through the common frame element *aspect* of the two spatial frames with triggers *in* and *of* appearing sequentially in the sentence. For each spatial trigger identified in a sentence, the following steps are performed:

1. First, the spatial triggers and the frame elements relevant to ischemic stroke are filtered. For this, we check if any of the Figure/Diagnosis element spans detected in relation to a trigger is IS-related. If one of the pre-defined IS-related imaging finding keywords (as shown in the first two rows of Table 2) is present in any of the element spans, the following steps are performed.
2. For extracting the brain side, we check for the presence of any laterality-related term in the predicted Ground element span (e.g., *left* for left, and *both*, *bilateral* for bilateral). Additionally, if the Ground elements are *thalami* and *capsules*, we assign the side as *bilateral*. In other cases, *unspecified* is assigned. Moreover, in cases (e.g., *infarction involving left frontal and parietal lobes*) when the same laterality is linked to multiple regions, each region is assigned the laterality separately. Here, *left* is assigned to both *frontal* and *parietal* lobes although *left* does not appear in the Ground span *parietal lobes*.
3. For identifying the brain region, the presence of any of the keywords developed for each of the pre-defined brain areas are checked in the detected Ground element span (e.g., keywords for mapping the brain region as 'Basal ganglia' are *basal ganglia*, *caudate*, *caudate nucleus*, *caudate head*, *caudate nucleus head*, *putamen*, *globus pallidus*, and *lentiform nucleus*). These keywords are built with domain expert input. Additionally, for Ground element spans involving two lobes, we assign both the cerebral lobes (e.g., *frontal* and *parietal* lobes are assigned for Ground element span – *frontoparietal*).

Table 2: Keywords for identifying IS finding, IS stage, and lacunarity from the frame element spans to classify the phenotypes.

Item	Keywords
IS-related imaging finding (CT)	hypodensity, hypodensities, hyperdensity, hyperdensities, hypodense, hypoattenuation, hypo-attenuation, low attenuation, low-attenuation, hypoattenuating, hypo-attenuating, low attenuating, low-attenuating, decreased attenuation, lacune, infarct, lesion
IS-related imaging finding (MRI)	restricted diffusion, slow diffusion, susceptibility artifact, signal, infarct
IS stage - Subacute	sub-acute, subacute, sub acute, evolving
IS stage - Acute	acute
IS stage - Chronic	encephalomalacia, gliosis, known, old, previous, prior
Lacunarity	lacune, lacunar

Table 3: Domain constraints applied on BERT predicted spatial frame elements to determine ischemic stroke stage.

Modality	Acute	Chronic
CT	(hypodensity/hypoattenuation in cortical/subcortical region) AND (hyperdense MCA OR hyperdensity in basilar artery OR loss of gray-white matter differentiation OR sulcal effacement)	(hypodensity/hypoattenuation in cortical/subcortical region AND (prominence of ventricles/sulci OR atrophy)) OR gliosis/encephalomalacia
MRI	(slow diffusion/restricted diffusion in cortical/subcortical region) OR (loss of flow void in MCA/basilar artery)	facilitated diffusion in cortical/subcortical region OR gliosis/encephalomalacia OR dilation of ventricles

4. For identifying the stroke stage for each pair of brain region and side, two sequential steps are involved. First, we check for the presence of any stage-related term directly in the predicted Figure/Diagnosis element span. Since the term *acute* is also contained in *subacute*, we priortize the search for subacute over acute. If not found, domain constraints are applied over the predicted spatial frame elements (this step also takes into account the other spatial relationships predicted in the same report in connection to the same brain region). If the stage is not determined by these two steps, we assign the label – *Can't determine*.
5. Similarly, for identifying the lacunarity for each pair of brain region and side, we check for the presence of lacunar-specific terms in the Figure/Diagnosis element span. We assign a binary lacunarity label – *Yes* if lacunar and *No* otherwise.

The keywords developed for IS-related imaging findings as well as for identifying the stroke stage and lacunarity from the frame element spans are shown in Table 2. These keywords as well as the domain constraints for inferring the stage are developed in collaboration with the radiologist who created the gold phenotypes. A few predominant constraints are demonstrated in Table 3.

Experimental Settings and Evaluation

We use the $\text{BERT}_{\text{LARGE}}$ model for fine-tuning the spatial information extraction task by initializing the model parameters obtained after pre-training BERT on MIMIC-III clinical notes for $300,000$ steps[20]. For extracting the spatial triggers from the RAD-IS-P data, we use the trained model from our previous work[7]. However, for extracting the frame elements, we re-train the BERT-based element extractor on the RAD-SPATIAL-IE dataset using the updated gold spans of Ground and Diagnosis frame elements for capturing the laterality and temporality information, respectively. We perform 10-fold cross-validation for evaluating the performance of the element extractor model. For each of the 10 iterations, we split the reports in RAD-SPATIAL-IE such that reports in 8 folds are used for training and 1 fold each are used for validation and testing. The model is fine-tuned by setting the maximum sequence length at 128, learning rate at $2e-5$, and number of training epochs at 4. We use cased version of the models. Among the 10 versions of the trained model checkpoints (generated for 10 folds of the dataset), we select the version based on the highest F1 measure on the validation set to predict the spatial frame elements from the RAD-IS-P data used for phenotype classification. Additionally, to provide a sense of the performance of the spatial information extraction system on stroke-related reports (that are more representative of the ones used for phenotyping), we annotated a random set of 20 reports from the RAD-IS-P dataset according to the Rad-SpatialNet schema and evaluated the system's performance on these 20 reports. For our phenotyping task, we report the precision, recall, and F1 measures of the phenotype extraction system based on various meaningful subsets or combinations of stroke features described in Section 3.

Table 4: 10 fold CV results on RAD-SPATIAL-IE for BERT-based spatial frame element extraction model using gold and predicted spatial triggers. P - Precision, R - Recall.

Main Frame Elements	Gold spatial triggers			Predicted spatial triggers		
	P(%)	R(%)	F1	P(%)	R(%)	F1
FIGURE	81.39	84.26	82.77	67.53	71.08	69.14
GROUND	92.01	93.41	92.69	70.87	80.13	75.09
HEDGE	75.51	83.08	78.91	68.94	74.05	71.19
DIAGNOSIS	54.73	78.41	64.06	48.49	67.67	55.95
RELATIVE POSITION	87.47	81.01	83.54	60.13	66.35	62.17
DISTANCE	75.83	80.83	75.53	73.63	80.00	74.25
POSITION STATUS	68.59	66.20	66.97	61.42	64.45	61.55
OVERALL	82.60	85.31	83.92	66.95	73.17	69.81

Table 5: BERT-based spatial frame element extractor's performance on 20 stroke reports (taken from RAD-IS-P). P - Precision, R - Recall.

Spatial triggers used	Overall P (%)	Overall R (%)	Overall F1
Gold annotated triggers	72.80	80.87	76.62
Predicted triggers	65.71	73.48	69.38

Results

The average precision, recall, and F1 scores of extracting spatial triggers from the RAD-SPATIAL-IE data are 86.14%, 79.55%, and 82.66, respectively. For the 20 stroke reports (selected from the RAD-IS-P data), the precision, recall, and F1 values for spatial trigger extraction are 93.70%, 76.28%, and 84.10, respectively. These predicted triggers are used further by the element extractor model in the end-to-end evaluation (shown under the 'Predicted spatial triggers' column in Table 4). Table 4 also highlights the average 10-fold CV performance measures of the BERT-based element extractor using the gold spatial triggers. The frame elements Associated Process and Reason have very low performance scores as they occur very rarely in the whole dataset and also not used for phenotyping. We additionally illustrate the overall precision, recall, and F1 measures (considering all the spatial frame elements) of the frame element extractor on the 20 stroke report subset in Table 5.

The results of our phenotype extraction system are shown in Table 6. We calculate the performance metrics of the system based on different combinations of the features (i.e., brain region, side, stroke stage, and lacunarity) that are potentially useful for clinical research studies. The precision, recall, and F1 values are calculated by comparing the distinct combinations of the features per report identified by the system to those of the gold annotated ones. This gives an idea about how well the system performs in classifying various subsets of meaningful features. Since stroke stage and lacunarity are associated with a specific brain region and side pair, we report the performance of the system including the stage and lacunarity features along with brain region and side in the last four rows of the table. Note that for stroke stage, we show the results both by considering various stage types and also by grouping the three stage types–acute, subacute, and acute/subacute together.

Discussion

This work focuses on identifying complex ischemic stroke phenotypes mainly from the perspective of the stroke location (brain region and side). We utilize the output of a spatial information extraction (IE) system (developed in our previous work) and apply simple neuroradiology-specific rules to classify these phenotypes. Note that the phenotypes we tackle in this work consider information at the level of specific brain area that is affected by stroke. Thus, this involves identification of information related to a stroke affected region in the brain from the report text. Our Rad-SpatialNet schema allows for easy identification of such related information as this captures the spatial relations between imaging findings and brain locations as well as the associated potential diagnoses. This becomes even more useful when the same report contains infarcts of different stages in different brain locations. Figure 2 illustrates an example where three different brain regions are affected and the stroke stage varies according to the region and its laterality.

Table 6: Phenotype extraction results. BR - brain region, CS - corresponding side, SS - stroke stage, SS_CO - SS with coarse types (*acute/chronic*), LC - lacunarity.

Phenotype variant	Example	Precision(%)	Recall(%)	F1
BR	*cerebellum*	73.58	89.62	80.81
BR + CS	*cerebellum, left*	68.34	85.47	75.95
BR + SS_CO	*cerebellum, chronic*	55.53	82.0	66.22
BR + CS + SS_CO	*cerebellum, left, chronic*	49.67	74.11	59.48
BR + CS + SS	*cerebral hemisphere - frontal lobe, bilateral, subacute*	46.32	56.96	51.09
BR + CS + LC	*basal ganglia, bilateral, yes*	62.53	77.2	69.09
BR + CS + SS_CO + LC	*basal ganglia, bilateral, chronic, yes*	48.59	72.49	58.18

We observe that applying simple domain rules that are mainly based on keyword search and a small set of constraints over the output of the spatial IE system results in satisfactory performance in classifying complex stroke phenotypes. This highlights both the information coverage of the Rad-SpatialNet schema and the sufficiently promising performance of the spatial IE system. Another point to note is that the information covered through Rad-SpatialNet are generic enough to extend our phenotype classification approach to other types of diseases/conditions beyond neuroradiology domains.

We briefly discuss the errors of the phenotype extraction system here. Most of the errors related to missing the brain region (referring to the recall of 89.62% in Table 6) is because of the Ground elements that are not predicted by the spatial IE system. There are also a very few cases where spatial triggers are not present explicitly (e.g., *left cerebellar infarct*). The existing Rad-SpatialNet schema doesnot capture such implicit relations and thus such regions are missed. Some of the errors related to stroke stage classification (when all the stage types are considered) is due to the ambiguity involved in distinguishing the acute and the subacute stages. Oftentimes, it becomes difficult to assess the stroke timing based on the report content (one of the major reasons for low recall for BR + CS + SS shown in Table 6). A small number of errors also occur when only acute and chronic stage information is considered because the output of the spatial IE system sometimes missed the specific stage-related term (e.g., *evolving*, *chronic*) in the predicted Diagnosis/Figure element span. Moreover, the report does not contain other spatial relations to satisfy the domain constraints for stage inference. Another reason of stage-related errors is when the stage information is mentioned in a following sentence in the report that does not contain any spatial relations (e.g., '*These lesions suggest old infarction'*). Lacunar-related errors happen mainly because their inferences sometimes depend on the specific sizes mentioned in the sentence (e.g., *lesion of 7 mm in diameter*) that are currently not captured in the Rad-SpatialNet schema. Taking into account a few limitations as described here in the Rad-SpatialNet schema, we aim to emphasize that there are rare instances of such scenarios overall across reports and we intend to further incorporate these information in the Rad-SpatialNet in our future work. We also see that the precision values are low, and one of the main reasons is that many of the stroke locations are referenced multiple times in a report and are expressed differently or with varying levels of specificity. For example, *left frontal lobe* is mentioned in the report's Findings section, whereas *left MCA* is mentioned in the Impressions section. This results in generating some false positive brain regions (e.g., *parietal* and *insula* here) as MCA (middle cerebral artery) maps to parts of *frontal* and *parietal lobes* as well as *insula* (the brain regions where MCA supplies blood to). The performance of our phenotype extraction system reflects the challenging nature of this complex phenotyping task and we aim to improve its performance and evaluate on an augmented dataset in a later work.

However, the phenotyping results suggest that the Rad-SpatialNet schema that we used in this work is robust enough considering the complexity of the phenotypes. We want to highlight that the current Rad-SpatialNet schema can be leveraged further to classify more granular aspects of the stroke location. Specifically, the RelativePosition frame element (e.g., *superior*, *inferior*) can be used to classify the subregions of a brain region like *cerebellum*. For instance, in the sentences of the same report–"*New acute infarction involving the superior left cerebellar hemisphere*" and "*Encephalomalacia and gliosis are again seen in the inferior left cerebellar hemisphere*", the stroke stage is *acute* in case of left cerebellum (superior) and *chronic* for left cerebellum (inferior). Thus, spatial information documented in the reports when extracted with detailed contextual information facilitates the classification of fine-grained phenotypes.

Conclusion

We used the output of an existing spatial information extraction system based on the Rad-SpatialNet schema to classify complex IS phenotypes. We demonstrated that a generalizable and fine-grained representation schema like Rad-SpatialNet could be utilized for determining detailed phenotypes that often requires information about various related radiological entities (such as findings, brain locations, and diagnoses). Our phenotypes are mainly based on specific brain regions affected by stroke. We have shown that satisfactorily good results can be achieved by applying simple domain rules on top of the IE system's output to classify the phenotypes.

Acknowledgments This work was supported in part by the National Institute of Biomedical Imaging and Bioengineering (NIBIB: R21EB029575) and the Patient-Centered Outcomes Research Institute (PCORI: ME-2018C1-10963).

References

[1] Stroke Facts — Cdc.Gov; 2020. Available from: https://www.cdc.gov/stroke/facts.htm.
[2] Effects of Stroke; 2020. Available from: https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/effects-of-stroke.
[3] Hui C, Tadi P, Patti L. Ischemic Stroke. In: StatPearls. StatPearls Publishing; 2020. Available from: http://www.ncbi.nlm.nih.gov/books/NBK499997/.
[4] Cheng Bastian, Forkert Nils Daniel, Zavaglia Melissa, et al. Influence of Stroke Infarct Location on Functional Outcome Measured by the Modified Rankin Scale. Stroke. 2014;45(6):1695–1702.
[5] Shi Y, Zeng Y, Wu L, et al. A Study of the Brain Functional Network of Post-Stroke Depression in Three Different Lesion Locations. Scientific Reports. 2017;7(1):14795.
[6] Price C, Seghier M, Leff A. Predicting Language Outcome and Recovery After Stroke (PLORAS). Nature reviews Neurology. 2010;6(4):202–210.
[7] Datta S, Ulinski M, Godfrey-Stovall J, et al. Rad-SpatialNet: A Frame-Based Resource for Fine-Grained Spatial Relations in Radiology Reports. In: Language Resources and Evaluation Conference; 2020. p. 2251–2260.
[8] Wheater E, Mair G, Sudlow C, et al. A Validated Natural Language Processing Algorithm for Brain Imaging Phenotypes from Radiology Reports in UK Electronic Health Records. BMC Med Inform Dec Mak. 2019;19(1):184.
[9] Ong CJ, Orfanoudaki A, Zhang R, et al. Machine Learning and Natural Language Processing Methods to Identify Ischemic Stroke, Acuity and Location from Radiology Reports. PLOS ONE. 2020;15(6):e0234908.
[10] Kim C, Zhu V, Obeid J, et al. Natural Language Processing and Machine Learning Algorithm to Identify Brain MRI Reports with Acute Ischemic Stroke. PloS One. 2019;14(2):e0212778.
[11] Fu S, Leung LY, Wang Y, et al. Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports. JMIR Medical Informatics. 2019;7(2).
[12] Devlin J, Chang MW, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT; 2019. p. 4171–4186.
[13] Sedghi E, Weber JH, Thomo A, et al. Mining Clinical Text for Stroke Prediction. Network Modeling Analysis in Health Informatics and Bioinformatics. 2015;4(1):16.
[14] Majersik Jennifer J, Mowery Danielle, Zhang Mingyuan, et al. Towards High-Precision Stroke Classification Using Natural Language Processing. Stroke;49(Suppl_1):92.
[15] Govindarajan P, Soundarapandian RK, Gandomi AH, et al. Classification of Stroke Disease Using Machine Learning Algorithms. Neural Computing and Applications. 2020;32(3):817–828.
[16] Garg R, Oh E, Naidech A, et al. Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing. J Stroke Cerebrovasc. 2019;28(7):2045–2051.
[17] Sung SF, Lin CY, Hu YH. EMR-Based Phenotyping of Ischemic Stroke Using Supervised Machine Learning and Text Mining Techniques. IEEE Journal of Biomedical and Health Informatics. 2020:1–1.
[18] Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data. 2016;3:160035.
[19] Birenbaum D, Bancroft LW, Felsberg GJ. Imaging in Acute Stroke. Western Journal of Emergency Medicine. 2011;12(1):67–76.
[20] Si Y, Wang J, Xu H, et al. Enhancing Clinical Concept Extraction with Contextual Embeddings. Journal of the American Medical Informatics Association. 2019;26(11):1297–1304.

Exploring the Hazards of Scaling Up Clinical Data Analyses: A Drug Side Effect Discovery Case Report

Franck Diaz-Garelli PhD[1], Todd R. Johnson PhD[2], Mohammad H. Rahbar PhD[2], Elmer V. Bernstam MD, MSE[2]
[1]University of North Carolina at Charlotte, Charlotte, NC
[2]The University of Texas Health Science Center at Houston, TX

Abstract
We assessed the scalability of pharmacological signal detection use case from a single-site CDW to a large aggregated clinical data warehouse (single-site database with 754,214 distinct patient IDs vs. multisite database with 49.8M). We aimed to explore whether a larger clinical dataset would provide clearer signals for secondary analyses such as detecting the known relationship between prednisone and weight. We found significant weight gain rate using the single-site data but not from using aggregated data (0.0104 kg/day, p<0.0001 vs. -0.050 kg/day, p<.0001). This rate was also found more consistently across 30 age and gender subgroups using the single-site data than in the aggregated data (26 vs. 18 significant weight gain findings). Contrary to our expectations, analyses of much larger aggregated clinical datasets did not yield stronger signals. Researchers must check the underlying model assumptions and account for greater heterogeneity when analyzing aggregated multisite data to ensure reliable findings.

Introduction
Implementation of electronic health records (EHRs) has enabled secondary analysis of clinical data for a variety of purposes[1,2]. Traditional time-consuming and costly research methods such as randomized controlled trials (RCTs)[2–5] are being complemented with secondary analyses of clinical data generated via clinical practice as a core activity of burgeoning learning healthcare systems[6,7]. As one important use case, 32% of novel therapeutics approved by the FDA were associated with a post-marketing safety event[8]. Thus, post-marketing discovery and surveillance for drug side effects (pharmacoepidemiology) using retrospective analyses of large clinical data sets has gained interest.

Though the current abundance of machine-readable clinical data sets enables such analyses, clinical data are prone to quality issues[9–13]. These issues are sometimes considered part of the noise that may mask signals in EHR data. It is often argued that larger data sets will amplify the signals, allows a more reliable estimate for the noise and improves the statistical power for detecting associations[14–18]. In the healthcare setting, this can be achieved in two ways. Healthcare systems can wait until their databases have enough patients to study a specific phenomenon, in which case, a more traditional research approach such as single site experimental study such as an RCT is developed and implemented. An alternative strategy, is for multiple healthcare systems to aggregate their data in clinical data warehouses (CDWs)[19,20] to increase statistical power[21], making clinical signals more detectable by secondary analysis methods such as regression analysis, as long as the underlying assumptions of these models are met. Research efforts are currently geared toward building large CDWs,[22–27] yet we found few studies in the literature reporting the impact of EHR data aggregation[28,29] or providing solutions to its consequences[30] and we found none illustrating its impact on traditional analytical methods such as regression analysis.

In this article, we attempt to "rediscover" the association between prednisone, a commonly prescribed corticosteroid, and weight gain[31,32] with longitudinal linear regression methods using two databases: a single-site CDW[33] and a much larger aggregated CDW. We chose this association because it is well-accepted by clinicians[34] and is defined by relatively objective numerical data (i.e., drug administration and weight change over time). Further, weight measurement does not require complex equipment and is often recorded during routine clinical care for a wide variety of patients across care settings. Thus, investigating the relationship between prednisone and weight gain is, in a sense, the "best case" scenario for assessment of factors associated with drug-side effect. Our goal was to "rediscover" a known side effect without leveraging knowledge about the side effect. Thus, we approximated the process of monitoring for previously unknown side effects. The contribution of this work is to illustrate an existing challenge of scaling up analyses using large, aggregated CDWs.

Methods
Data Sources - We used longitudinal linear regression methods to analyze the known relationship between prednisone and weight gain using real EHR data extracted from a single-site CDW and an aggregated CDW database. The single-

site data set was extracted from an outpatient clinic's EHR production database and contained 754,214 distinct patient IDs with data from April 2004 to January 2014. The aggregated data set, Cerner Health Facts (Cerner Corporation, Kansas City, MO) a HIPAA de-identified database, contained over 49.8 million distinct patient IDs with data from January 2000 to October 2014. We selected patients with at least one prednisone prescription, all their recorded weights and covariates such as age and gender. Descriptive statistics are shown in Table 1. No missing values were found for age, and gender variables in either data set. The weight variable was normally distributed. Because the distribution of drug exposure was not normal, we categorized exposure into a binary variable encoded as high (above mean) vs. low (below mean) exposure. We filtered out patients under 21 years of age and extreme outliers for weight (i.e., weight>400 kg). A second round of filtering was performed on the weight variable by excluding measurements that were more than three standard deviations away from the mean on both sides. This study has been approved by the Committee for the Protection of Human Subjects (the UTHSC-H IRB) under protocol HSC-SBMI-13-0549.[35]

Table 1 - Descriptive statistics for single and aggregated datasets. Note that the original aggregated database contains over sixty times more patients than the single-site database, whereas the extracted data has slightly over 18 times more patients.

	Single-Site	Aggregated
Number of Distinct Patient IDs in Database	754,214	49,826,219
Timespan Covered	Apr. 2004 - Jan. 2014	Jan. 2000 - Oct. 2014
Number of Patients Included (% of Database)	9,767 (1.29%)	169,944 (0.34%)
Number of Weight Measurements	93,617	2,278,953
Missing Drug Exposure	15.40%	20.40%
Average Age (in years) for 1st Prescription	54±15	63±17
Median Age (in years) for 1st Prescription	55	64
Mean±SD Weight	84.2±22kg	82.5±24kg
Mean±SD Prednisone Exposure	312±697mg	781±1,740mg

Statistical Analysis - We used summary statistics such as mean, median and extreme values (e.g., values more than three standard deviations away from the mean) to screen the data for outliers, missing values and erroneous data. We assessed normality of continuous variables based on histograms. To detect weight gain over time, we built a longitudinal linear regression model using generalized estimating equations (GEE) method. Our longitudinal regression model predicted the main outcome (weight) based on drug exposure (high/low), the number of days from the first prescription (time), the patient's age at the time of first prescription, the patient's gender. An autoregressive correlation structure was used to account for potential correlation between multiple measurements from the same individuals. Statistical significance was set at α=0.05. A model was built on weight as the dependent variable with independent variables including, time and a binary exposure group (cumulative prednisone dose below or above the mean) along with known covariates: gender and age. The observational time span (or time windowing) was varied around the time of prescription between 90 days before prescription and 360 days after prescription to improve model fit. We used the Quasi-likelihood under Independence Model Criterion (QIC) [36] as a quantitative measure of goodness-of-fit. QIC is a generalization of Akaike's information criterion which represents the goodness of fit of statistical models typically employed with GEE methods and informs model selection. We used SAS (version 9.4, SAS Institute Inc., Cary, NC) for statistical analysis.

Sensitivity Analysis - To assess the robustness of our results, we built regression models for multiple sub-populations contained in our data set using the same methods described above. For example, we built separate regression models for males only, females only, initial weight below or above average and for patients with ages below or above average at the time of prescription. These subgroup analyses aimed to identify which subgroups of the cohort contributed to the overall effect. If the effect varies by multiple subgroups, it indicates potential effect modifications that should be explored.

Results
We built longitudinal linear regression models to optimize model fit (i.e., reduce QIC) exploring time windows between 10 days prior and 120 days after prescription. The best fit was found for 7-90 days and 7-32 days for the single-site and aggregated data respectively (Table 2). This matches what is clinically expected: patients gain weight within the first three months of starting prednisone. These conditions returned the largest effect sizes for the relationship between prednisone and weight in both data sets: 0.0114 kg/day ($p<0.0001$) for the single-site data and -0.050 kg/day ($p<0.001$) but no association with exposure to prednisone ($p=0.847$) based on the aggregated data.

Overall, our longitudinal linear regression (i.e., all data after prescription) yielded no statistical significance for time ($p=0.900$), exposure ($p=0.389$) or the interaction between these two variables ($p=0.237$). The aggregated dataset showed significance for all variables, most probably due to the large size of the dataset. However, the model estimated that the weight gain was negative over time, though the amount of weight loss was not statistically significant. Model fit was measured using QIC, with lower values indicating better fit, but no specific reference ranges. Both models showed a very high QIC=70,020 (single site) and QIC=835,847 (aggregated) indicating very poor fit of the model to the data.

Setting an upper temporal boundary at 90 and 32 days from first prescription for each data set respectively brought all variables beyond the significance threshold (time: $p<0.0001$, exposure: $p=0.0266$, interaction between these two variables: $p<0.0001$), improving the fit (QIC=15,669) for the single site data set. The final coefficient estimate value for time was 0.0104 kg/day (~3.8 kg/year), which revealed a positive correlation between time and weight increase in patients prescribed prednisone. On the other hand, the aggregated data showed significance for all variables. The QIC value improved to QIC=393,087.

We found larger QIC values indicating a poorer fit for larger upper time limits (60 and 90 days). Setting a delayed lower time limit where the data considered for the regression started at 7 days after the first prescription improved fit (QIC=7335), preserving statistical significance for all variables except binary (high/low) exposure ($p=0.0645$) in the single-site dataset. Adding this lower time limit to the aggregated data set yielded significance for all variables except exposure ($p=0.847$) and the interaction between exposure and time ($p=0.068$). However, the estimate for the time variable was negative (-0.050 kg/day) indicating a slight decrease in weight over time. It is important to note that the minimum QIC found for the aggregated data set was much higher than the single-site data set (QIC=120,133 vs. 7335), suggesting that a larger, more complex data set may include much more variability and may be much harder to account for heterogeneity in variance by this model. In summary, we found weight gain with prednisone in the single-site data matching the expected outcome but slight weight loss with prednisone in the aggregated data.

Sensitivity Analysis - To assess the robustness of our findings, we performed sensitivity analysis by doing subgroup analyses (Figure 1). The single-site CDW showed a positive relationship between time and weight for all categories. Only age categories above 70 and patients with an initial weight below average showed no significance. These findings suggest homogeneity within the subgroups of this single-site data set with an upward trend detectable in multiple subgroups. On the other hand, the findings from the aggregated data were inconsistent. Regression analysis provided 11 negative estimated effects (only 5 statistically significant) and 4 positive estimated effects (none statistically significant). The largest effects were found for patients in their 20s (estimated effect=-0.514 kg/day, $p=0.338$), 50s (estimated effect=-0.426 kg/day, $p=0.261$), age below average (estimated effect=-0.292 kg/day, $p=0.046$) and male patients (estimated effect=-0.320 kg/day, $p=0.006$). These results suggest that this data set contains heterogeneous sub-populations that have very different weight changes over time.

We compared size of data set and sampling rate (i.e., measures per patient) for the two data sets (Table 3). Overall, the aggregated data had only a slightly lower sampling rate than the single-site data set (8.65 vs. 9.56 measurements/patient). In contrast, for each optimized time window presenting the best QIC fit (i.e., 7-90 days and 7-32 days for the single-site and aggregated datasets respectively and found that the aggregated data had roughly twice the sampling rate compared to the single site data. However, the aggregated data set contained multiple visit types such as outpatient, inpatient and emergency, whereas the single-site database contained outpatients only. Comparing outpatient data only, we found 2.96 vs. 9.65 measurements/patient for the single and aggregated data sets respectively and 1.26 vs. 1.84 measurements/patient respectively within the selected time window. In this case, the lower sampling rate may partly explain the lower p-values due to lower statistical power for the single-site data in both analyses ($p=0.014$ vs. $p<0.0001$). The lower measurement count per patient could be explained by the difference in time windows 7-90 days vs. 7-32 days. However, using the 7-90 day time window and outpatients only the CDW data set returned a smaller sample than the single-site data set (1.69 vs. 1.84 measurements/patient). Analysis of these data showed weight gain (estimate=0.0114 kg/day), yet not significant ($p=0.487$).

Table 2 – Results from Longitudinal Linear Regression models with weight as the dependent variable under varying parameter assumptions based on the single-site dataset (n_1=93,617) and the aggregated dataset (n_2=2,278,953).

		Overall Regression (After Prescription)				Upper Time Limit				Upper and Lower Time Limit			
		Single-site		Aggregated		Single-site		Aggregated		Single-site		Aggregated	
Parameter		**Estimate**	**p-value**	**Estimate**	**p-value**	**Estimate**	**p-value**	**Estimate**	**p-value**	**Estimate**	**p-value**	**Estimate**	**p-value**
Intercept		83.45	<.0001	93.8	<.0001	82.05	<.0001	92.6	<.0001	84.2715	<.0001	92.2	<.0001
Time (Days from Prescription)		0	0.9004	-0.0032	<.0001	-0.0048	0.0038	-0.061	<.0001	0.0104	<.0001	-0.050	<.0001
Age (in years)		-0.087	<.0001	-0.2399	<.0001	-0.0729	<.0001	-0.228	<.0001	-0.101	<.0001	-0.213	<.0001
Gender	M	14.43	<.0001	8.3245	<.0001	14.78	<.0001	8.60	<.0001	14.10	<.0001	7.81	<.0001
Gender (Ref.)	F												
Exposure (Classified)	Hi	-0.56	0.389	0.149	.0002	-1.40	0.0266	0.146	<.0001	-1.64	0.0616	-2.51	0.847
Exposure (Ref.)	Lo												
Days*Exposure (Class)	Hi	-0.0003	0.237	-0.00008	<.0001	0.0158	<.0001	-0.0194	<.0001	0.0156	0.0055	1.47	0.068
Days* Exposure (Ref.)	Lo												
QIC Value		79,020		835,487		15,669		393,087		7,335		120,133	

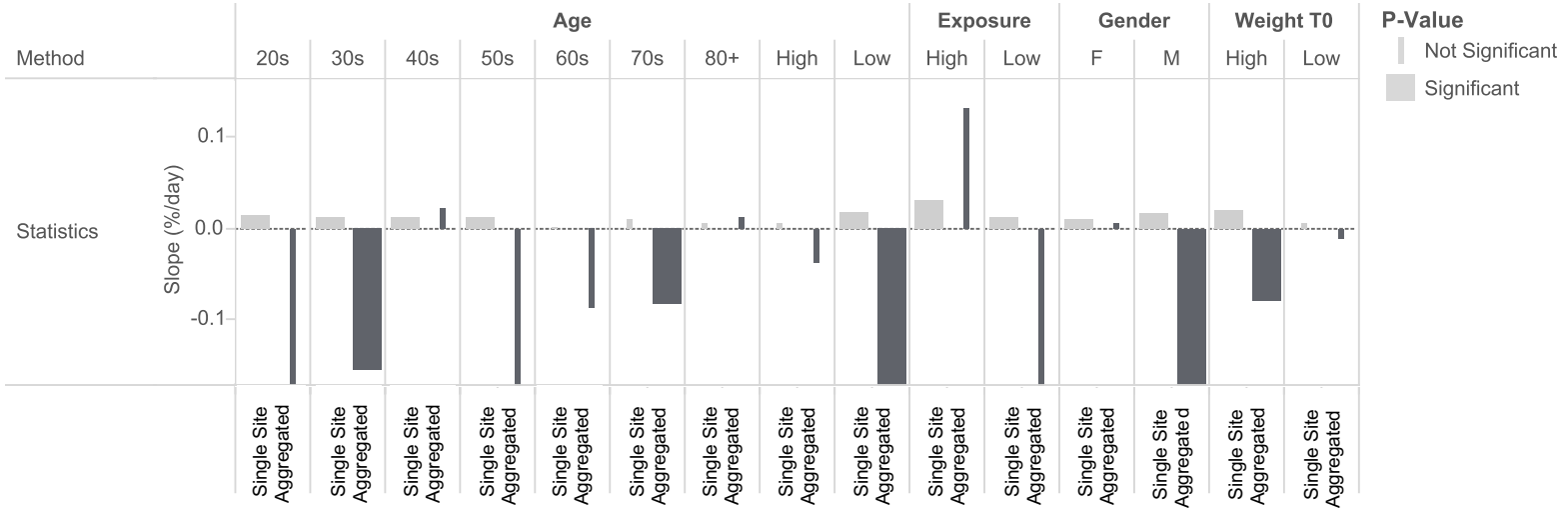


Figure 1 - Relationship between weight and follow up time for sub-populations within the single-site and aggregated data sets. The single-site data set presents similar results for most sub populations, whereas the aggregated data set returned much more variable results.

Table 3 – Regression analysis results summary for each data set and analytical method with patient counts and total number of measurements. Black arrows indicate statistically significant findings, whereas gray arrows represent p-values greater than 0.05.

	Single-Site		***Aggregated***				
Data set	*Full Data Set*	*Days 7-90 (Strongest Signal)*	*Full Data Set*	*Days 7-32 (Strongest Signal)*	*Outpatients Only*	*Outpatients Days 7-32*	*Outpatients Days 7-90*
Patient Count	9,854	4,812	169,056	34,194	37,890	7,008	12,055
Total Number of Weight Measurements	94,233	8,872	1,355,508	127,550	112,098	8,874	20,370
Average Number of Weight Measurements Per Patient	9.56	1.84	8.02	3.73	2.96	1.27	1.69
Statistical Estimate (kg/day)	0 (p=0.900)	0.0104 (p<0.0001)	-0.0032 (p<.0001)	-0.050 (p<.0001)	0.0005 (p=0.781)	-0.0280 (p=0.0014)	0.0114 (p=0.487)
Finding	→	↗	↘	↘	↗	↘	↗

Finally, we built a model by controlling for the site where the weight measure was recorded on the aggregated dataset; this is often done for multisite aggregated data. Controlling for site did not change our conclusions regarding rate of change in weight over time, finding very similar estimates to that in the final model from Table 2 (e.g., -0.0503, p<.0001 vs. -0.050, p<.0001 for the time variable). The dataset contained 146 different sites out of which 35 returned statistically significant estimates for the effect of time on weight; This revealed significant differences across sites. We then evaluated potential interaction between exposure and site terms (i.e., exposure*site), finding 63 statistically significant interaction terms out of 273, revealing that prednisone exposure depended on the clinical site in 23% of all sites. We also found very similar estimates (e.g., -0.0498, p<.0001 vs. -0.050, p<.0001 for time) in this model that accounted for interactions, which led to the same weight loss conclusion. These additional results indicate a robust statistical difference across sites and a significant interaction between the site and exposure in this particular aggregated database. However, controlling only for the study site did not improve the model, change the relationship between weight and time or the conclusions regarding the effect of prednisone on weight in the larger aggregated dataset.

Discussion

We found that prednisone was associated with weight loss, weight gain or no effect depending on the analysis methods, data set and assumptions. The much larger aggregated data set seemed more heterogeneous, including multiple visit types (e.g., outpatient, inpatient, emergency department, etc.), highly variable sampling rate (i.e., number of weight measurements per patient), multiple sources of data that seemed to present statistically distinct results and an interaction between the exposure and the site. This counters the idea that a larger aggregated data set would present a stronger overall signal and yield more robust findings in all circumstances, as we did not consistently find the expected association between prednisone and weight gain using traditional statistical regression methods. Subgroups analyses also yielded more inconsistent results in the aggregated dataset compared to the single-site dataset. In summary, the aggregated dataset appeared more heterogeneous and presented more analytical challenges during analyses.

Many studies have explored issues related to EHR data quality and bias[2,9,37–42]. However, the heterogeneous nature of CDWs has only been mentioned as a potential hazard[43–48] and few publications have shown or quantitated the impact of heterogeneous aggregated clinical data sets on analysis.[28–30] Data quality is often defined as "fitness for purpose" [49–51]. However, "fitness for purpose" is difficult to define in the absence of a clear understanding of the data sources constituting a clinical dataset.[52,53] Without this knowledge, it is difficult to anticipate threats to "fitness for purpose", because the processes that produced the analytical data are unknown and potentially unknowable due to data merging and deidentification. In other words, it is much more difficult to assess whether an aggregated

dataset is “fit-for-a-particular-purpose” than for a single-site dataset (e.g., a local CDW) where data production processes are known by the analytical team. Simply put, it is more challenging to assess whether aggregated data are “fit-for-a-particular-purpose” than for a single-site/local CDW where data production processes are known. This case report illustrates this limitation and our findings s*uggest that CDW heterogeneity and data quality may affect analytical outcomes that may not be mitigated by using a larger database*.

Our results suggest that these data quality and potential biases can change analysis outcome and are not necessarily mitigated by using a larger database. This may be due to the fact that clinical data reflect workflow, convention and multiple other factors in addition to “biology”.[38,54–57] The number of such confounders grows with the aggregation of data from multiple care contexts including institutions, settings (e.g., outpatient vs. inpatient, primary care vs. subspecialty, medical vs. surgical, etc.) and other contexts. We found that visit types (inpatient vs. outpatient vs. emergency department) was one of the strongest confounders, which is more intimately related to clinical workflow than the biology of the studied population. This is consistent with existing work.[56,57] We also found differences across sites and interactions that must be taken into account for reliable analysis [58]. This hints at the potential existence of analytical challenges particular to aggregated CDW (i.e., heterogeneous clinical workflow-derived data collected into a single database) that go beyond pitfalls expected in analyses making use of data produced for research purposes (e.g., multi-site data). These challenges must be further defined, and their effects evaluated in future research.

Though more data generally means more statistical power and narrower confidence intervals[32,33], it is important to consider potential heterogeneities[34] in large multisite aggregated clinical data sets. Because patients from diverse institutions (i.e., diverse clinical workflows), receiving diverse modes of treatment are included, the data may not be comparable[14,27,35–37]. This has been often cited in the statistical literature[64,65] but could be easily overlooked in applied clinical informatics data reuse. The complexity of data production processes and aggregated databases are a constant threat to appropriate data use[66]. It may be possible to tease out this complexity by employing reasonable assumptions that rely on clinical and healthcare workflow knowledge[33]. For example, Hripcsak et al.[67] were able to reproduce previous Pneumonia Outcomes Research Team (PORT) studies using EHR data, but only after eliminating the vast majority (up to 90%) of the patients. In addition, their analysis (like ours) also benefited from knowing the expected results.

Understanding the workflow that produced the data being analyzed is often challenging; particularly for multisite data aggregated across institutions and care settings. Incomplete understanding of how a data set was produced is, generally a hazard to analysis and interpretation of analytical results from large clinical data sets[66]. Clinical workflow may be difficult to define for a particular clinic or patient type even at a single institution. For example, are patients weighed routinely on every visit or only when there is reason to suspect a change or a specific clinical question? Are prescriptions from all providers, some of whom may be from other institutions, reliably entered into the source EHR? Our findings suggest that this threat grows with clinical data aggregation. They also suggest caution when reusing large aggregated CDWs for secondary analyses and the use of automated “big data” approaches to find previously unrecognized side effects. Also, just like in RCT data analysis, there are other statistical threats such as Simpson’s paradox[68,69] and sub-group heterogeneity[65,70] that must be considered under even if the research team has perfect knowledge of how the data were produced.

A strength of our study is that we used two relatively large, robust data sets. The single-institution data included over 750 thousand patients and the multi-site database included over 49.8 million patients; among the largest clinical data sets that exist in the US.

Our study has limitations that will be addressed in future work. First, the analytical methods selected were not comprehensive. However, we ran similar regression on both datasets for the sake of comparability and explored additional regressions on the aggregated dataset that uncovered its complexities. To better understand the reasons for our findings, we chose traditional statistical models rather than neural network-based machine learning approaches (e.g., deep learning methods) that produce “opaque” models[71]. Though these methods could potentially enable accurate analyses despite poor data quality and hetorogeneity,[72] understanding their role in analysis scalability was beyond the scope of this initial work. Most analysis decisions were driven by the need to replicate of a previous analysis[33] where the effect was found for the single-site dataset. This was done to allow for comparable analytical results. For example, we used a longitudinal regression models that are based on GEE methods with a categorization of the prednisone exposure variable per the initial analysis setup. Other advanced methods and their adaptation to

these aggregated datasets will be explored in future work. Second, we chose a particular drug-side effect association. Thus, our results may not apply to other associations, particularly side effects that do not evolve over time. For example, it may be easier to discover events such as myocardial infarctions[55] than trends over time. Third, we did not investigate all possible sub-populations and potential covariates (e.g., race and ethnicity), concurrent clinical events such as other clinical conditions, drugs taken or surgical procedures. Many additional variables could potentially improve accuracy, some of which may be difficult to accurately assess (e.g., compliance with medications). An alternative approach would be a case-control retrospective study design with propensity scoring to control for potential confounding effects. Although these issues are clearly important, our primary goal was to assess the impact of size of data set on the detection of drug side effect associations in large aggregated clinical data sets rather than to build an optimal model for this particular case or to discover new biology.

Conclusion

Analysis of larger clinical databases does not necessarily generate clearer overall signals, in particular if the large data set is aggregated from multiple heterogeneous data sources. Analyses that successfully detected a known association in a single-site dataset were unable to detect this same association in a much larger aggregated data set despite of leveraging significant knowledge about the side effect (e.g., expected timing relative to exposure). Analyses of large, aggregated, anonymized data sets require attention to additional details addressing their heterogeneity beyond basic analysis design point typically considered for smaller, single-site clinical datasets.

Acknowledgements

This work was partially supported in part by UTHealth Innovation for Cancer Prevention Research Pre-doctoral Fellowship (Cancer Prevention and Research Institute of Texas grant #160015), NIH NCATS grants UL1 TR001420, UL1 TR000371 and UL1 TR001105, NIH NCI grant U01 CA180964, NSF grant III 0964613, the Brown Foundation, Inc., NIGMS Institutional Research and Academic Career Development Award (IRACDA) program (K12-GM102773) and the Bridges Family Partnership, Ltd. - Sally and Joe Bridges, Jennifer and Todd Darwin, and Beth and Drew Cozby. This work was conducted with data provided by and support from the Cerner Corporation and UTHealth School of Biomedical Informatics. The content is solely the responsibility of the author(s) and does not necessarily represent the official views of the Cerner Corporation, the UTHealth School of Biomedical Informatics, The University of North Carolina at Charlotte or the National Institutes of Health.

References

1. Safran C. Reuse Of Clinical Data. IMIA Yearb. 2014;9(1):52–4.
2. Weiner MG, Embi PJ. Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? Ann Intern Med. 2009 Sep 1;151(5):359–60.
3. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. Medical data mining: knowledge discovery in a clinical data warehouse. Proc AMIA Annu Fall Symp. 1997;101–5.
4. Blum RL. Discovery, confirmation, and incorporation of causal relationships from a large time-oriented clinical data base: The RX project. Comput Biomed Res. 1982 Apr;15(2):164–87.
5. Frawley WJ, Piatetsky-Shapiro G, Matheus CJ. Knowledge discovery in databases: An overview. AI Mag. 1992;13(3):57.
6. Safran C. Reuse Of Clinical Data. IMIA Yearb. 2014;9(1):52–54.
7. Safran C. Using routinely collected data for clinical research. Stat Med. 1991 Apr;10(4):559–564.
8. Downing NS, Shah ND, Aminawung JA, Pease AM, Zeitoun J-D, Krumholz HM, et al. Postmarket Safety Events Among Novel Therapeutics Approved by the US Food and Drug Administration Between 2001 and 2010. JAMA. 2017 May 9;317(18):1854–63.
9. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. Med Care. 2013 Aug;51(8 0 3):S30–7.
10. Liaw S-T, Taggart J, Dennis S, Yeo A. Data quality and fitness for purpose of routinely collected data - a general practice case study from an electronic Practice-Based Research Network (ePBRN). AMIA Annu Symp Proc. 2011;2011:785–94.
11. Nobles AL, Vilankar K, Wu H, Barnes LE. Evaluation of data quality of multisite electronic health record data for secondary analysis. In: 2015 IEEE International Conference on Big Data (Big Data). 2015. p. 2612–20.
12. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring Diagnoses: ICD Code Accuracy. Health Serv Res. 2005 Oct 1;40(5p2):1620–39.

13. Farzandipour M, Sheikhtaheri A, Sadoughi F. Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10). Int J Inf Manag. 2010 Feb 1;30(1):78–84.
14. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete [Internet]. WIRED. [cited 2014 Sep 24]. Available from: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory
15. John Walker S. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Int J Advert. 2014 Jan;33(1):181–3.
16. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;2:3.
17. Belle A, Thiagarajan R, Soroushmehr SMR, Navidi F, Beard DA, Najarian K. Big Data Analytics in Healthcare [Internet]. BioMed Research International. 2015 [cited 2019 Feb 7]. Available from: https://www.hindawi.com/journals/bmri/2015/370194/abs/
18. Sun J, Reddy CK. Big Data Analytics for Healthcare. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2013 [cited 2019 Feb 7]. p. 1525–1525. (KDD '13). Available from: http://doi.acm.org/10.1145/2487575.2506178
19. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed Health Data Networks: A Practical and Preferred Approach to Multi-Institutional Evaluations of Comparative Effectiveness, Safety, and Quality of Care. Med Care. 2010 Jun;48:S45–51.
20. Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. J Am Med Inform Assoc. 2014 Jul 1;21(4):730–6.
21. Tabachnick BG, Fidell LS. Using multivariate statistics, 5th ed. Boston, MA: Allyn & Bacon/Pearson Education; 2007. xxvii, 980 p. (Using multivariate statistics, 5th ed).
22. Hripcsak G, Duke J, Shah N, Reich C, Huser V, Schuemie M, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. MEDINFO. 2015;15.
23. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci. 2016 Jul 5;113(27):7329–36.
24. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An Evaluation of the THIN Database in the OMOP Common Data Model for Active Drug Safety Surveillance. Drug Saf. 2013 Feb 1;36(2):119–34.
25. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar 1;17(2):124–30.
26. Natter MD, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, et al. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J Am Med Inform Assoc. 2013 Jan 1;20(1):172–9.
27. Segagni D, Tibollo V, Dagliati A, Perinati L, Zambelli A, Priori S, et al. The ONCO-I2b2 project: integrating biobank information and clinical data to support translational research in oncology. Stud Health Technol Inform. 2011;169:887–91.
28. Seneviratne MG, Kahn MG, Hernandez-Boussard T. Merging heterogeneous clinical data to enable knowledge discovery. In: Biocomputing 2019 [Internet]. WORLD SCIENTIFIC; 2018 [cited 2020 Aug 13]. p. 439–43. Available from: https://www.worldscientific.com/doi/abs/10.1142/9789813279827_0040
29. Glynn EF, Hoffman MA. Heterogeneity introduced by EHR system implementation in a de-identified data resource from 100 non-affiliated organizations. JAMIA Open. 2019 Dec 1;2(4):554–61.
30. Fu S, Leung LY, Raulli A-O, Kallmes DF, Kinsman KA, Nelson KB, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. BMC Med Inform Decis Mak. 2020 Mar 30;20(1):60.
31. Baker JF, Sauer BC, Cannon GW, Teng C-C, Michaud K, Ibrahim S, et al. Changes in Body Mass Related to the Initiation of Disease-Modifying Therapies in Rheumatoid Arthritis. Arthritis Rheumatol Hoboken NJ. 2016 Aug;68(8):1818–27.
32. WUNG PK, ANDERSON T, FONTAINE KR, HOFFMAN GS, SPECKS U, MERKEL PA, et al. Effects of Glucocorticoids on Weight Change During the Treatment of Wegener's Granulomatosis. Arthritis Rheum. 2008 May 15;59(5):746–53.
33. Diaz-Garelli J-F, Bernstam EV, MSE, Rahbar MH, Johnson T. Rediscovering drug side effects: the impact of analytical assumptions on the detection of associations in EHR data. AMIA Summits Transl Sci Proc. 2015 Mar 25;2015:51–5.
34. PredniSONE Tablets [Package Insert]. Ridgefield, CT : Boehringer-Ingelheim Inc. 2012.

35. Guerrero SC, Sridhar S, Edmonds C, Solis CF, Zhang J, McPherson DD, et al. Access to Routinely Collected Clinical Data for Research: A Process Implemented at an Academic Medical Center. Clin Transl Sci. 2019;12(3):231–5.
36. Pan W. Akaike's Information Criterion in Generalized Estimating Equations. Biometrics. 2001 Mar 1;57(1):120–5.
37. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Summits Transl Sci Proc. 2010;2010:1.
38. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Bias Associated with Mining Electronic Health Records. J Biomed Discov Collab. 2011 Jun 6;6:48–52.
39. Rea S, Bailey KR, Pathak J, Haug PJ. Bias in Recording of Body Mass Index Data in the Electronic Health Record. AMIA Summits Transl Sci Proc. 2013 18;2013:214–8.
40. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak. 2014 Jun 11;14(1):51.
41. Diaz-Garelli J-F, Wells BJ, Yelton C, Strowd R, Topaloglu U. Biopsy Records Do Not Reduce Diagnosis Variability in Cancer Patient EHRs: Are We More Uncertain After Knowing? AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2018;2017:72–80.
42. Diaz-Garelli J-F, Strowd R, Wells BJ, Ahmed T, Merrill R, Topaloglu U. Lost in Translation: Diagnosis Records Show More Inaccuracies After Biopsy in Oncology Care EHRs. AMIA Summits Transl Sci Proc. 2019 May 6;2019:325–34.
43. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data. Med Care. 2012 Jul;50(Suppl):S49–59.
44. Conway M, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, et al. Analyzing the Heterogeneity and Complexity of Electronic Health Record Oriented Phenotyping Algorithms. AMIA Annu Symp Proc. 2011;2011:274–83.
45. Chute CG, Pathak J, Savova GK, Bailey KR, Schor MI, Hart LA, et al. The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2011;2011:248–56.
46. Sun J, Wang F, Hu J, Edabollahi S. Supervised Patient Similarity Measure of Heterogeneous Patient Records.
47. Grissom RJ. Heterogeneity of variance in clinical data.
48. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013 Jan 1;20(1):117–21.
49. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan;20(1):144–151.
50. Holve E, Kahn M, Nahm M, Ryan P, Weiskopf N. A comprehensive framework for data quality assessment in CER. AMIA Summits Transl Sci Proc. 2013 18;2013:86–8.
51. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs [Internet]. 2016 Sep 11 [cited 2017 Apr 6];4(1). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051581/
52. Diaz-Garelli J-F, Bernstam EV, Lee M, Hwang KO, Rahbar MH, Johnson TR. DataGauge: A Practical Process for Systematically Designing and Implementing Quality Assessments of Repurposed Clinical Data. EGEMs Gener Evid Methods Improve Patient Outcomes. 2019 Jul 25;7(1):32.
53. DiazVasquez J. DataGauge: A Model-Driven Framework for Systematically Assessing the Quality of Clinical Data for Secondary Use. UT SBMI Diss Open Access [Internet]. 2016 Aug 16; Available from: http://digitalcommons.library.tmc.edu/uthshis_dissertations/33
54. Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Adjusting for Unmeasured Confounders in Pharmacoepidemiologic Claims Data Using External Information: The Example of COX2 Inhibitors and Myocardial Infarction. Epidemiology. 2005 Jan;16(1):17–24.
55. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The Tell-Tale Heart: Population-Based Surveillance Reveals an Association of Rofecoxib and Celecoxib with Myocardial Infarction. PLoS ONE. 2007 Sep 5;2(9):e840.
56. Diaz-Garelli J-F, Strowd R, Ahmed T, Wells BJ, Merrill R, Laurini J, et al. A tale of three subspecialties: Diagnosis recording patterns are internally consistent but Specialty-Dependent. JAMIA Open [Internet]. 2019 Aug 5 [cited 2019 Sep 6]; Available from: https://academic.oup.com/jamiaopen/advance-article/doi/10.1093/jamiaopen/ooz020/5543799

57. Diaz-Garelli F, Strowd R, Lawson VL, Mayorga ME, Wells BJ, Lycan TW, et al. Workflow Differences Affect Data Accuracy in Oncologic EHRs: A First Step Toward Detangling the Diagnosis Data Babel. JCO Clin Cancer Inform. 2020 Jun 1;(4):529–38.
58. Kp V, M L. The Effect of Ignoring Statistical Interactions in Regression Analyses Conducted in Epidemiologic Studies: An Example with Survival Analysis Using Cox Proportional Hazards Regression Model. Epidemiol Open Access [Internet]. 2016 [cited 2016 Oct 24];06(01). Available from: http://www.omicsonline.org/open-access/the-effect-of-ignoring-statistical-interactions-in-regression-analysesconducted-in-epidemiologic-studies-an-example-with-survival-2161-1165-1000216.php?aid=69316
59. Aliferis CF, Statnikov A, Tsamardinos I, Schildcrout JS, Shepherd BE, Jr FEH. Factors Influencing the Statistical Power of Complex Data Analysis Protocols for Molecular Signature Development from Microarray Data. PLOS ONE. 2009 Mar 17;4(3):e4922.
60. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting Sample Size Required for Classification Performance. 2012 [cited 2016 Sep 7]; Available from: http://dl.umsu.ac.ir//handle/Hannan/26109
61. Fletcher J. What is heterogeneity and is it important? BMJ. 2007 Jan 11;334(7584):94–6.
62. Hoffman S, Podgurski A. Big Bad Data: Law, Public Health, and Biomedical Databases. J Law Med Ethics. 2013 Mar 1;41:56–60.
63. Tatonetti N, Ye P, Daneshjou R, Altman R. Data-Driven Prediction of Drug Effects and Interactions. Sci Transl Med. 2012 Mar 14;4(125):125ra31-125ra31.
64. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. The Lancet. 2000 Mar 25;355(9209):1064–9.
65. Gelman A, Auerbach J. Age-aggregation bias in mortality trends. Proc Natl Acad Sci. 2016 Feb 16;113(7):E816–7.
66. Van Der Lei J. Use and abuse of computer-stored medical records. Methods Inf Med. 1991;30(2):79–80.
67. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Using discordance to improve classification in narrative clinical databases: An application to community-acquired pneumonia. Comput Biol Med. 2007 Mar;37(3):296–304.
68. Julious SA, Mullee MA. Confounding and Simpson's paradox. BMJ. 1994 Dec 3;309(6967):1480–1.
69. Wagner CH. Simpson's Paradox in Real Life. Am Stat. 1982 Feb 1;36(1):46–8.
70. Gelman A. Commentary: P Values and Statistical Practice. Epidemiology. 2013;24(1):69–72.
71. Tollenaar N, Heijden PGM van der. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. J R Stat Soc Ser A Stat Soc. 2013;176(2):565–84.
72. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. Npj Digit Med. 2018 May 8;1(1):1–10.

Integrating Biomedical Informatics Training into Existing High School Curricula

Avantika R. Diwadkar, MS[1], Susan Yoon, PhD[2], Jooeun Shim, MSEd[2], Michael Gonzalez, PhD[1], Ryan Urbanowicz, PhD[1], Blanca E. Himes, PhD[1]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, US

[2]Graduate School of Education, University of Pennsylvania, Philadelphia, PA, US

Abstract

Growing demand for biomedical informaticists and expertise in areas related to this discipline has accentuated the need to integrate biomedical informatics training into high school curricula. The K-12 Bioinformatics professional development project educates high school teachers about data analysis, biomedical informatics and mobile learning, and partners with them to expose high school students to health and environment-related issues using biomedical informatics knowledge and current technologies. We designed low-cost pollution sensors and created interactive web applications that teachers from six Philadelphia public high schools used during the 2019-2020 school year to successfully implement a problem-based mobile learning unit that included collecting and interpreting air pollution data, as well as relating this data to asthma. Through this project, we sought to improve data and health literacy among the students and teachers, while inspiring student engagement by demonstrating how biomedical informatics can help address problems relevant to communities where students live.

Introduction

Biomedical informatics—an interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health[1]—is a specialty that arose in response to the need for computing in biology and medicine. As data-intensive computing has become essential to research and practice in various health domains, the demand for professionals trained in data science and biomedical informatics has steadily risen[2–4]. Further, as data-driven health decisions are increasingly necessary in daily life, health literacy (i.e., the capacity to obtain, communicate, process, and understand basic health information and services to make appropriate health decisions) and data literacy (i.e., the ability to read, work with, analyze, and argue with data) are critical life skills for all citizens[5,6]. Long-standing training programs in biomedical informatics exist at the post-graduate and undergraduate levels, and elements of data science and biomedical informatics are increasingly becoming embedded in the curricula to train health professionals and undergraduates[7–9]. A training gap exists, however, at the level of high school, whereby many high school graduates are under-prepared for, and unaware of, careers in biomedical informatics. Of further concern, levels of data and health literacy may be low in communities relying on out-of-date high school science curricula[10,11]. In response to these issues, several summer programs have been designed by universities to expose high school students to bioinformatics and biomedical informatics research[12–14], and the American Medical Informatics Association (AMIA) began the High School Scholars Program to provide a dissemination venue for these students to report their research findings and network more broadly with trainees and professionals across the U.S.[15]. Although these programs have been very successful, they reach a limited number of students and can be biased toward those who are already prepared and high achieving. To prepare a wide range of students for biomedical informatics careers and equip them with fundamental skills in health and data literacy, teaching in the high school setting is necessary.

Early exposure to science, technology, engineering, and mathematics (STEM) careers and academic preparedness in K-12 are critical factors for success in STEM at the undergraduate level and beyond[16]. Programs seeking to broaden participation in STEM have thus attempted to enhance teaching of, and provide exposure to, STEM-related areas to students of all ages[17,18]. For example, introduction of bioinformatics curricula has shown significant gains in cognitive traits among high school students, as well as an increased interest in STEM careers[19]. Structured summer and afterschool research programs can also be highly effective STEM-strengthening interventions for high school students and may help increase the participation and diversity within the biomedical workforce[20,21]. The viability of incorporating training of STEM, and biomedical informatics specifically, into high school curricula is challenged by the lack of curricular resources and teacher knowledge[10]. Therefore, professional development activities initiated through collaborations between educators and scientists are necessary to effectively improve student proficiency in these areas[22]. Moreover, due to the complexity and continuous evolution of biomedical informatics approaches, there

is a general agreement that new curricula should be inquiry-based, tied to STEM content already taught, and support real-world problem solving[10,23]. Any new material must also impart basic biomedical informatics skills through simple, nonburdensome integration into the existing high school curricula. To imitate real-world scientific practice, mobile learning and technologies (e.g., portable sensors and mobile apps) offer an inexpensive way to collect large-scale data and embed activities into the student's context that enable authentic experimentation and participation[24,25].

Over half of Philadelphia, PA consists of Environmental Justice areas—census tracts where 20 percent or more of individuals live in poverty and/or 30 percent or more of the population is minority[26]. Residents of these areas have historically suffered a disproportionate burden of pollutant exposures from sources that include oil refineries, trash incinerator plants, and vehicular exhaust from major roadways, putting them at increased risk for various diseases, including asthma. Exposure to environmental pollutants such as fine particulate matter ($PM_{2.5}$) has been associated with increased risk of asthma exacerbations[27,28], and Philadelphia is among the most polluted cities in the U.S[29]. Over 20 percent of children in Philadelphia have asthma[30,31], with hospitalizations occurring at a rate of 59.1 per 10,000[32]. Asthma prevalence and hospitalizations disproportionately affect Black and Hispanic residents and those who are socioeconomically disadvantaged[32,33]. Because many Philadelphia children have asthma or know of people in their communities who do, asthma serves as a relatable case study for students to understand how the environment (air pollution), genetics, and lifestyle factors (smoking) contribute to disease. Further, biomedical informatics approaches can readily be applied to studies of asthma and its various risk factors which facilitates teaching of biomedical informatics in different high school classes (environmental science, biology).

We initiated a professional development project entitled "K-12 Bioinformatics" in 2019 to educate Philadelphia area high school teachers about data analysis and mobile learning, and partner with them to expose high school students to health and environment-related issues using biomedical informatics knowledge and current technologies. Through this project, we sought to impart skills that will aid students to investigate real-world health problems and inspire them to take action in their local communities, while also generating interest in future biomedical informatics and STEM careers. Here, we describe some of the materials we created to facilitate teaching of biomedical informatics to high school teachers, and report how teachers used these materials during the academic year.

Methods

Summer Institute for Teacher Professional Development. During the summer of 2019, three environmental science and three biology teachers from the School District of Philadelphia attended a three-week professional development course to bring current science research into the secondary classroom through a problem-based learning curriculum in biomedical informatics. Three teachers identified as African American and three identified as White. The schools they taught in ranged in race and ethnic diversity, each qualified for U.S. federal Title I Funds (i.e., low income families made up at least 40 percent of the enrollment), and the percent of students per school deemed *proficient or advanced* in the Keystone Exam ranged from 2 to 100 percent. In partnership with researchers from the University of Pennsylvania Graduate School of Education and Perelman School of Medicine, teachers learned to build and integrate mobile technologies into classroom activities and designed a problem-based learning unit for classroom implementation during the 2019-2020 school year. The course work consisted of in-person lectures that introduced concepts such as bioinformatics and air pollution, hands-on trainings with a "teacher-as-student" pedagogy such as investigating and interpreting air quality data using tools, modifying problem-based learning units for their existing curricula, and pilot testing their units with high school students invited to participate in the summer program.

Low-Cost Pollution Sensor Assembly. We designed and assembled low-cost $PM_{2.5}$ and carbon monoxide (CO) portable sensors using commercially available Android components. The prototype design was improved by the Penn Electronic Design shop who ordered custom printed circuit boards to facilitate assembly of 100 sensors. Each pollution sensor kit included a sensor, Android smartphone and 9V batteries (Figure 1A). The sensors paired via Bluetooth with an Android smartphone to record pollution measures with an app created with App Inventor, a block-based programming language for building Android apps (Figure 1B). Sensor measurements, phone location, time, de-identified student IDs, and other general information captured by the app was saved to a Google Sheet (Figure 1C). Details regarding the sensor components and assembly, as well as the related code to capture measures are available at https://github.com/HimesGroup/k12bioinformatics.

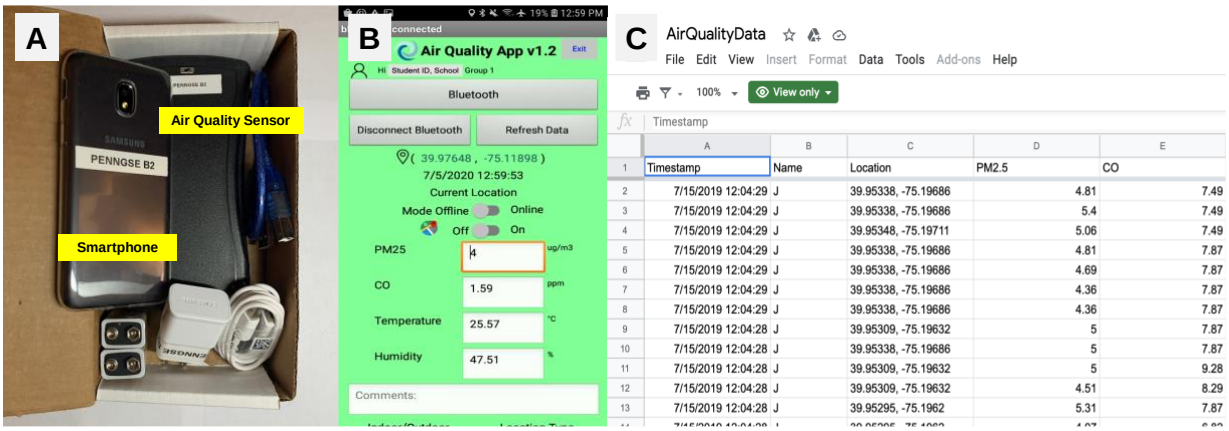


Figure 1. Low-cost pollution sensor measurements. (A) Air quality sensor kit included sensor, smartphone, cables, and 9V batteries. (B) Air quality app created with App Inventor collects measures of $PM_{2.5}$, CO, temperature, humidity, and geographic data. (C) A Google sheet was used to store data from all sensors distributed to a school.

Regulatory Monitor Air Pollution Data. To provide context for low-cost sensor pollution measures, data from Environmental Protection Agency (EPA) regulatory monitors was obtained for eight diverse U.S. cities (Philadelphia, PA, New York, NY, Los Angeles, CA, Miami, FL, Pierre, SD, Billings, MO, Standing Rock, NM and Portland, OR). Daily average $PM_{2.5}$ measures for September 2017 were computed from AirData[34] hourly estimates for all available monitors within each city. Monthly averages of $PM_{2.5}$ and CO for each city based on measures at a single geocoordinate during the years 2007-2017 were obtained with the R package *pargasite*[35].

Gene Expression Microarray Data. We searched for a gene expression microarray study related to the effects of cigarette smoking in the Gene Expression Omnibus (GEO) and selected one that measured differences in gene expression in macrophages from 13 cigarette smokers versus 11 non-smokers (GEO accession number GSE8823)[36]. The RAVED pipeline was used to analyze this dataset (https://github.com/HimesGroup/raved)[37], which included obtaining quality control metrics using outlier scoring methods from the *arrayQualityMetrics* R package[38], transforming the raw data using the robust multi-array average (RMA)[39] method, and performing differential expression analysis with the *limma* R package[40]. The Benjamini-Hochberg approach was used to correct for multiple comparisons, and an adjusted p-value <0.05 was considered significant.

Web Application Development. Prior to the summer institute for high school teacher professional development, we designed prototypes of three web-based tools that would provide an interactive user-interface to teach biomedical informatics concepts applied to topics that are highly relevant to students in the Philadelphia area, namely asthma and environmental exposures (e.g., air pollution and smoking). During the summer institute, as teachers designed problem-based mobile learning modules for use during the school year, the apps were updated based on teacher feedback to meet their needs. The apps, available at http://www.k12bioinformatics.org/, assisted with teaching of biomedical informatics concepts: 1) an exploratory data analysis tool for basic data visualization, 2) an app to visualize air pollution measures from low-cost sensors and regulatory monitors, and 3) an app demonstrating steps of gene expression microarray analysis. The apps were created with the R *Shiny* package[41] and deployed on a DigitalOcean droplet containing RStudio Connect and RStudio Server Pro. Various R packages, including *leaflet*[42] and *ggiraph*[43] were used to create the apps that displayed interactive plots and maps. Full code for the apps is available at https://github.com/HimesGroup/k12bioinformatics.

Results

Exploratory Data Analysis.

Description of App. The goal of this app was to help convey concepts related to descriptive statistical analysis. Users can upload either their own tabular data file in comma-separated-value (csv) format or use an example file available in the app. Tabular data is parsed by the app to identify columns that contain categorical and continuous variables, and subsequently, visualize univariate (e.g., bar plot for categorical variable, histogram for continuous variable) and bivariate (e.g., split bar plots, split box plots) distributions of user-selected variables. Students can learn to interpret

these visualizations using their own spreadsheets without the need to code. This app was intended for use with, for example, the phenotype file associated with the gene expression microarray study.

Use During the School Year. Students initially tested the app by uploading the example data file to understand types of variables and learn how to interpret bar plots, histograms and box plots. Subsequently, they utilized their own data to generate summary statistics and visualizations. For example, in one lesson, teachers provided students with large air quality datasets from two regulatory monitor sites (i.e., Torresdale Station and Car-Barn Montgomery I -76), and students uploaded the dataset and made plots (e.g., bar plot of mean values and box plot of all values) to compare differences between the two sites and examine air quality measures over time. Teachers guided students to identify what was familiar and what was confusing. Additionally, teachers asked students to respond to questions such as "What can we learn from a box plot that we can't learn from a bar plot and vice versa?" to ensure students learned proper usage of each visualization. For final student project reports, teachers instructed students to use the air quality data they collected along with Philadelphia asthma rate data to generate plots with this app.

Air Pollution Analysis and Visualization.

Description of App. The goal of this app was to help describe characteristics and visualize the geospatial distribution of low-cost air pollution sensor measures recorded by the students at all participating schools. Additionally, this app provides regulatory monitor pollution data that students can use as gold-standard pollution measures to contrast with those of low-cost sensors. The "EPA Measures Map" tab shows average $PM_{2.5}$ estimates in September 2017 for each of the eight cities, with the highest value for Portland, OR (10.5 µg/m^3) and lowest for Standing Rock, NM (5.6 µg/m^3) (Figure 2A). Interactive scatter plots of monthly average $PM_{2.5}$ and CO levels for a user-selected location and time period that provide information regarding pollution trends in the eight cities across a 10-year period are provided in the next tab. Seasonal variation in $PM_{2.5}$ and CO monthly average estimates can be observed, with differing trends based on location. For example, monthly $PM_{2.5}$ estimates in Philadelphia across the years 2007 to 2017 were consistently higher in the period from May to August while such trends were generally absent in Portland for the same time period (Figure 2B). The "Crowdsourced Map" and "Crowdsourced Summary Plots" tabs enable users to access, process, and analyze crowdsourced sensor data. Specifically, students can compute daily averages of sensor measures and compare them with the available regulatory monitor data and visualize in an interactive map of Philadelphia the estimates recorded for a user-selected pollutant, date range, school and sensor (Figure 2C). The overall characteristics of the selected data can be studied through univariate (box plot and histogram) plots for each pollutant as well as bivariate (scatter) plots comparing $PM_{2.5}$ and CO distributions.

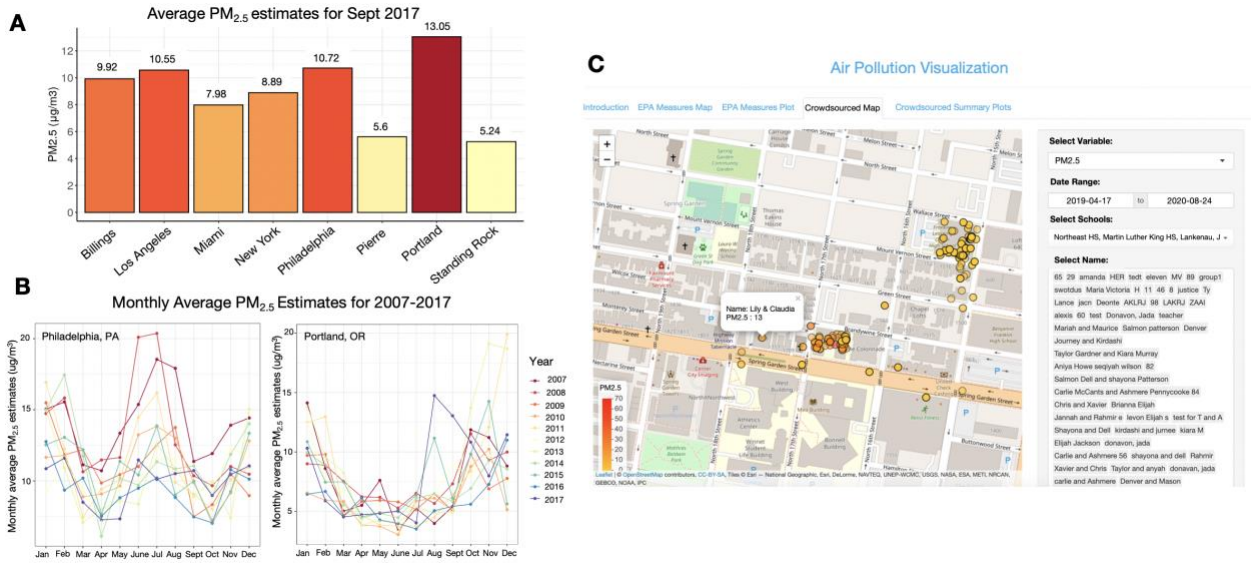


Figure 2. Features of air pollution visualization app. (A) Bar plot of average EPA AirData $PM_{2.5}$ estimates for September 2017 in eight U.S. cities. (B) Monthly $PM_{2.5}$ estimates for Philadelphia, PA and Portland, OR from 2007-2017. (C) Interactive map of crowdsourced $PM_{2.5}$ and CO sensor data acquired by students and teachers of six Philadelphia high schools.

Use During the School Year. During the school year, 153 students and teachers from six School District of Philadelphia high schools participated in the air pollution problem-based learning units. On average, 15 sensors were distributed to each of the six schools and managed using technology tracker sheets. In classrooms, one sensor package was assigned to a group of 2-4 students with distributed roles (e.g., sensor carrier, phone carrier, map navigator, observer and cartographer). Students were introduced to concepts related to air pollution, types of pollutants, EPA regulatory monitors in the U.S., and how the low-cost sensors worked. For the data collection activities, teachers first set classroom expectations (e.g., carry yourself as though you are a scientist) and then students gathered outdoor measures for 30-40 minutes. The research team provided on-site support when teachers introduced the sensors to the students and 1-2 backup sensor packages were available in case of technical difficulties. During the school year, 2001 measures of $PM_{2.5}$ and CO were successfully recorded. Using the app, students visualized data for their own schools and compared the pollutant levels across dates and geolocations to identify the most polluted areas among the surveyed neighborhoods in Philadelphia, recognizing that low-cost sensors are limited compared to research grade or regulatory monitors. Additionally, students downloaded the AirData $PM_{2.5}$ pollution estimates for the eight U.S. cities provided in the app and calculated the daily average estimate in each city for the month of September 2017. They compared these estimates with the *pargasite*-acquired estimates displayed in the app and explored seasonal trends across the years 2007-2017 via interactive scatter plots.

Gene Expression Microarray Analysis.

Description of App. The goal of this app was to allow students to explore the workflow of a gene expression microarray analysis without having to write code to do so. The "Sample Characteristics" tab explores the phenotype data information (smoking status, age, ethnicity and sex of the sample donors) associated with the study through univariate and bivariate plots of user-selected variables. The "Quality Control" analysis tab includes normalization of gene expression raw data using the RMA method, outlier identification through both log_2-transformed/normalized intensity distribution box plots (Figure 3A) and intensity curves of all samples in the dataset, and dimensionality reduction with principal component analysis (PCA) plots to visualize clustering patterns and batch effects for user-selected variables (Figure 3B). Under the "Differential Expression Results" tab, the top 50 significant differentially expressed gene probes (adjusted p-value < 0.05) are displayed in tabular form along with a volcano plot, a box plot comparing normalized read counts of smokers versus non-smokers for a user-selected gene (Figure 3C), and a heatmap of a user-selected number of top ranking probes (Figure 3D).

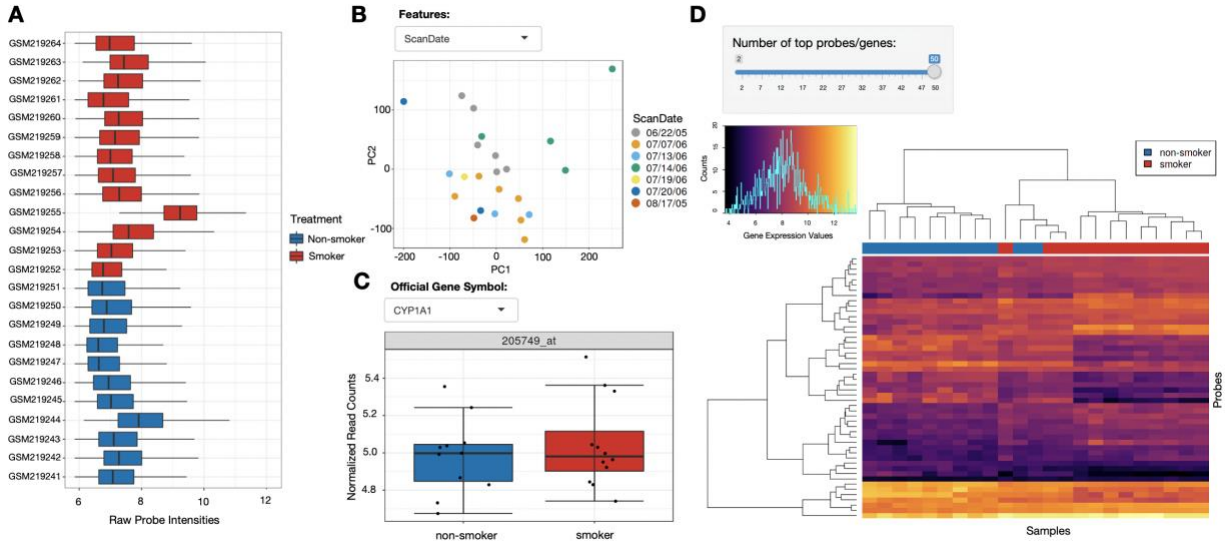


Figure 3. Features of gene expression microarray analysis app. (A) Log_2-transformed/normalized intensity distribution box plots of each sample. (B) Principal Component Analysis (PCA) plot colored according to a user-selected variable. (C) Box plots comparing normalized read count distributions of a user-selected gene *CYP1A1* in smokers versus non-smokers. (D) Heatmap based on a user-selected top number of probes/genes.

Use During the School Year. This app was not used by teachers during the school year, as they thought too many concepts would have to be introduced that were beyond the scope of their current curriculum. However, one teacher had students use the "Sample Characteristics" tab as a homework assignment in which students were asked to make an inference by writing a claim-evidence-reasoning essay on smoker/non-smoker data. The teacher asked students to

identify what they noticed and think of possible research questions based on this data set. Teacher prompts to support students making a claim and inference included, "Can you hypothesize answers to those questions?" and "What evidence do you have to support these inferences?".

Discussion

Despite its increasing scientific and practical importance, biomedical informatics remains a complex field and activities imparting related theoretical knowledge coupled with informatics skills and scientific thinking are atypical in current school tasks. Due to its relatively recent emergence and the continuous evolution of its scope, optimally integrating biomedical informatics in the high school scientific curricula is not straightforward. A previous study that evaluated existing U.S. secondary school science standards for bioinformatics found low and vague representation of bioinformatics across various topics, the lowest of which was "Human Genome Project/genomics and computer use" (implemented in only 4 of 49 U.S. states and the District of Columbia)[44]. There are major challenges to incorporating bioinformatics, and biomedical informatics more broadly, into high school lessons: 1) no guidance is provided on how to teach bioinformatics topics among existing frameworks[44], 2) teachers often lack any prior experience in bioinformatics research or its instruction[45], 3) schools lack sufficient computational infrastructure (e.g., computers, internet access, qualified IT personnel)[45], and 4) the majority of students do not have basic programming skills. To help overcome these challenges, biomedical informatics should be introduced as a practical education elective compatible with the time frame, cognitive level and available resources of the target population, while requiring minimal technical support and providing long-term training to the teachers[45]. Several external educational outreach programs offering bioinformatics training and research opportunities to high school students exist across the U.S.[13,14], but most engage directly with the students on an individual basis, and very few provide professional development and training strategies for the teachers[19]. Furthermore, public-facing education programs that engage with a mobile-learning curriculum designed around community issues of low-income minority neighborhoods to boost youth participation in local science campaigns and policymaking are still few in number[24].

The K-12 Bioinformatics program is the first of its kind in the Philadelphia area to provide advanced scientific seminars and professional training to educators, while equipping them with open-source, freely available web-based interactive tools that aid in imparting biomedical informatics training to high school students via effective visualizations and concise analysis workflows. We chose web-based resources that are akin to the facilities available at the participating public schools with majority low-income enrollments. Additionally, the resources such as databases and analysis tools that help build upon the data literacy of the teachers were procurable and nonintimidating for usage. Our initiative included problem-based mobile learning activities focused on locally contingent health hazards that aimed to empower students to address real-world problems through technological and data-driven investigations and propel evidence-based arguments with community stakeholders.

While planning our initial summer training session, we experienced minor setbacks, especially in designing the classroom activity using the air pollution sensors. Assembling the sensors in-house was more cumbersome than expected: initial sensor prototypes took 75 minutes to build, including soldering wires, gluing buttons using epoxy, and testing devices. We had considered including sensor assembly with the school activity but opted to work with a local electronics design shop to speed up production and improve the fabrication process. A major issue was experienced by the biology teachers who deemed the gene expression app, designed specifically to incorporate bioinformatics, as too complicated and elaborate for usage. Moreover, due to an existing heavy academic course load, the biology teachers found that the integration of bioinformatics-related modules was too cumbersome and time-consuming. In contrast, the environmental science teachers found it easier to incorporate additional materials related to air pollution measures into their existing curricula. As the summer institute progressed, the biology teachers decided to use the air pollution visualization app for their classroom activity sessions too. Lastly, during the school year, we faced an unanticipated challenge: the school district's firewall blocked the server where the shiny apps were hosted because it was integrated into the primary k12bioinformatics.org website without a named domain (i.e., with an IP address only). Communicating with the appropriate high school's IT team and the school district to identify and resolve this issue delayed the implementation of the new unit with the first teacher's class.

Our first summer training session in 2019 was successful, and teachers provided helpful feedback that was used to improve the program. To minimize teacher cognitive load and ensure long-term effectiveness, we introduced redesigned activities into our professional development program for the second cohort of teachers that included a fully annotated teacher guide armed with student-facing instructional presentations elucidating relevant scientific information translated for high school populations and familiar readymade curricular resources. We also extended the

summer program duration to a longer period with fewer daily hours. Due to COVID-19-related public health measures, the 2020 summer institute was held virtually. Publicly available video lectures introducing concepts such as bioinformatics, public health informatics, air pollution, and genetics were recorded in advance for asynchronous delivery, and the program was delivered via the edX platform. The new group of eight teachers from seven School District of Philadelphia high schools along with five teachers from the first session will implement newly developed units into their schools over the 2020-2021 school year, which includes collecting indoor air quality estimates at home and performing data analysis. Additionally, teachers are designing novel teaching strategies to implement mobile learning units that are appropriate to the circumstances faced at their schools.

In ongoing work, we are improving the interface of our apps to add more background content for coherence and help ease individual differences in contextualizing and processing biomedical informatics-specific knowledge and skills. We are also re-evaluating multiple aspects of the existing apps to build more on the foundations of data analysis as a concept, rather than fast-track through advanced topics. Increasing the compatibility of the material by matching the information base with the cognitive level of the participating teachers and students will help enhance the usability of the tools, making them more impactful in fulfilling the overall goals of the initiative. To ensure that our materials can be disseminated more widely and given that sensors may be unavailable or their use cost prohibitive in some settings, we are creating modules that do not rely on sensors. As computer programming becomes a requirement in more high schools across the country, expanding teaching modules to include more coding will enable the design of higher-level training materials in biomedical informatics.

Conclusion

We launched the K-12 Bioinformatics professional development project to educate high school teachers about biomedical informatics and partner with them to expose high school students to health and environment-related issues using biomedical informatics knowledge and current technologies. We successfully implemented a problem-based mobile learning unit focused on measuring neighborhood air pollution with low-costs sensors and relating these measures to asthma, thereby providing a hands-on biomedical informatics research experience that can potentially set in motion increased youth engagement in health and environmental hazards at a community level and generate interest in pursuing higher education in biomedical informatics.

Acknowledgements

We would like to thank Phil Bowsher and RStudio for providing licenses to RStudio Server Pro and RStudio Connect, which facilitated the web application development and deployment. We thank Miguel E. Hernandez and Alexander Santos from the Penn Electronic Design Shop for helping with low-cost air pollution sensor design and assembly. This work was supported by National Science Foundation award DRK12 1812738.

References

1. Kulikowski CA, Shortliffe EH, Currie LM, Elkin PL, Hunter LE, Johnson TR, et al. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. J Am Med Inform Assoc JAMIA. 2012 Dec;19(6):931–8.

2. Hersh W. The health information technology workforce: Estimations of demands and a framework for requirements. Appl Clin Inform. 2010;1(2):197–212.

3. Investing in America's data science and analytics talent. Business Higher Education Forum and PWC; [cited 2020 Aug 25]. Available from: https://www.bhef.com/sites/default/files/bhef_2017_investing_in_dsa.pdf

4. Missed opportunities? The labor market in health informatics, 2014. Burning Glass; [cited 2020 Aug 25]. Available from: https://www.burning-glass.com/wp-content/uploads/BG-Health_Informatics_2014.pdf

5. Wolff A, Gooch D, Montaner JJC, Rashid U, Kortuem G. Creating an understanding of data literacy for a data-driven society. J Community Inform. [cited 2020 Aug 25]; Available from: http://www.ci-journal.net/index.php/ciej/article/view/1286

6. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Viera A, Crotty K, et al. Health literacy interventions and outcomes: An updated systematic review. Evid ReportTechnology Assess. 2011 Mar;(199):1–941.

7. Florance V. Training for informatics research careers: History of extramural informatics training at the national library of medicine. In: Berner ES, editor. Informatics Education in Healthcare. London: Springer London; 2014. p. 27–42. (Health Informatics). Available from: http://link.springer.com/10.1007/978-1-4471-4078-8_3

8. Zhan YA, Wray CG, Namburi S, Glantz ST, Laubenbacher R, Chuang JH. Fostering bioinformatics education through skill development of professors: Big genomic data skills training for professors. Ouellette F, editor. PLOS Comput Biol. 2019 Jun 13;15(6):e1007026.

9. Moore JH, Boland MR, Camara PG, Chervitz H, Gonzalez G, Himes BE, et al. Preparing next-generation scientists for biomedical big data: Artificial intelligence approaches. Pers Med. 2019 May;16(3):247–57.

10. Machluf Y, Gelbart H, Ben-Dor S, Yarden A. Making authentic science accessible—the benefits and challenges of integrating bioinformatics into a high-school science curriculum. Brief Bioinform. 2017 Jan;18(1):145–59.

11. Smith PS. What does a national survey tell us about progress toward the vision of the NGSS? J Sci Teach Educ. 2020 Aug 17;31(6):601–9.

12. Teen Research and Education in Environmental Science (TREES) by Center of Excellence in Environmental Toxicology (CEET) at University of Pennsylvania. [cited 2020 Aug 25]. Available from: http://ceet.upenn.edu/training-career-development/summer-programs/teen-research-and-education-in-environmental-science/

13. Dutta-Moscato J, Gopalakrishnan V, Lotze M, Becich M. Creating a pipeline of talent for informatics: STEM initiative for high school students in computer science, biology, and biomedical informatics. J Pathol Inform. 2014;5(1):12.

14. Stanford Institute of Medicine summer Research program (SIMR). [cited 2020 Aug 25]. Available from: https://simr.stanford.edu

15. Unertl KM, Finnell JT, Sarkar IN. Developing new pathways into the biomedical informatics field: the AMIA high school scholars program. J Am Med Inform Assoc. 2016 Jul;23(4):819–23.

16. Honey M, Pearson G, Schweingruber HA, National Academy of Engineering, National Research Council (U.S.), editors. STEM integration in K-12 education: Status, prospects, and an agenda for research. Washington, D.C: The National Academies Press; 2014. 165 p.

17. Museus SD, Palmer RT, Davis RJ, Maramba DC, Ward K, Wolf-Wendel L. Racial and ethnic minority students' success in STEM education. San Francisco, Calif.: Hoboken, N.J: Jossey-Bass Inc.; Wiley Periodicals; 2011. 140 p. (ASHE higher education report).

18. Fisher AJ, Mendoza-Denton R, Patt C, Young I, Eppig A, Garrell RL, et al. Structure and belonging: Pathways to success for underrepresented minority and women PhD students in STEM fields. PloS One. 2019;14(1):e0209279.

19. Kovarik DN, Patterson DG, Cohen C, Sanders EA, Peterson KA, Porter SG, et al. Bioinformatics education in high school: implications for promoting science, technology, engineering, and mathematics careers. CBE Life Sci Educ. 2013;12(3):441–59.

20. Salto LM, Riggs ML, Delgado De Leon D, Casiano CA, De Leon M. Underrepresented minority high school and college students report STEM-pipeline sustaining gains after participating in the Loma Linda University Summer Health Disparities Research Program. PloS One. 2014;9(9):e108497.

21. Chittum JR, Jones BD, Akalin S, Schram ÁB. The effects of an afterschool STEM program on students' motivation and engagement. Int J STEM Educ. 2017;4(1):11.

22. Pierret C, Sonju JD, Leicester JE, Hoody M, LaBounty TJ, Frimannsdottir KR, et al. Improvement in student science proficiency through InSciEd out. Zebrafish. 2012 Dec;9(4):155–68.

23. Shuster M, Claussen K, Locke M, Glazewski K. Bioinformatics in the K-8 classroom: Designing innovative activities for teacher implementation. Int J Des Learn. 2016 Feb 3 [cited 2020 Aug 25];7(1). Available from: https://scholarworks.iu.edu/journals/index.php/ijdl/article/view/19406

24. Taylor KH, Silvis D, Kalir R, Negron A, Cramer C, Bell A, et al. Supporting public-facing education for youth: Spreading (not scaling) ways to learn data science with mobile and geospatial technologies. Contemp Issues Technol Teach Educ 193. [cited 2020 Aug 25]; Available from: https://citejournal.org/volume-19/issue-3-19/current-practice/supporting-public-facing-education-for-youth-spreading-not-scaling-ways-to-learn-data-science-with-mobile-and-geospatial-technologies

25. Herodotou C, Villasclaras-Fernández E, Sharples M. The design and evaluation of a sensor-based mobile application for citizen inquiry science investigations. In: Rensing C, de Freitas S, Ley T, Muñoz-Merino PJ, editors. Open Learning and Teaching in Educational Communities. Cham: Springer International Publishing; 2014 [cited 2020 Aug 25]. p. 434–9. (Lecture Notes in Computer Science; vol. 8719). Available from: http://link.springer.com/10.1007/978-3-319-11200-8_38

26. Department of Environmental Protection, State of Pennsylvania. PA Environmental Justice Areas. Available from: https://www.dep.pa.gov/PublicParticipation/OfficeofEnvironmentalJustice/Pages/PA-Environmental-Justice-Areas.aspx

27. Orellano P, Quaranta N, Reynoso J, Balbi B, Vasquez J. Effect of outdoor air pollution on asthma exacerbations in children and adults: Systematic review and multilevel meta-analysis. PloS One. 2017;12(3):e0174050.

28. Mirabelli MC, Vaidyanathan A, Flanders WD, Qin X, Garbe P. Outdoor PM2.5, ambient air temperature, and asthma symptoms in the past 14 days among adults with active asthma. Environ Health Perspect. 2016;124(12):1882–90.

29. American Lung Association. State of the air 2019. [cited 2020 Aug 25]. Available from: http://www.stateoftheair.org/assets/sota-2019-full.pdf

30. Bryant-Stephens T, West C, Dirl C, Banks T, Briggs V, Rosenthal M. Asthma prevalence in Philadelphia: Description of two community-based methodologies to assess asthma prevalence in an inner-city population. J Asthma. 2012 Aug;49(6):581–5.

31. Mangione S, Yuen EJ, Balsley C. Asthma prevalence in children: A survey of 57 Philadelphia middle schools. (Asthma: guidelines, delivery system and public health). Chest; (No 4). Report No.: Vol 122.

32. Department of Public Health, City of Philadelphia. Philadelphia Community Health Assessment: Health of the City 2019. [cited 2020 Aug 25]. Available from: https://www.phila.gov/media/20191219114641/Health_of_City_2019-FINAL.pdf

33. Bryant-Stephens T. Asthma disparities in urban environments. J Allergy Clin Immunol. 2009 Jun;123(6):1199–206.

34. Pre-generated data files. United States Environmental Protection Agency. [cited 2020 Aug 25]. Available from: https://aqs.epa.gov/aqsweb/airdata/download_files.html

35. Greenblatt RE, Himes BE. Facilitating inclusion of geocoded pollution data into health studies. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2019;2019:553–61.

36. Kazeros A, Harvey B-G, Carolan BJ, Vanni H, Krause A, Crystal RG. Overexpression of apoptotic cell removal receptor MERTK in alveolar macrophages of cigarette smokers. Am J Respir Cell Mol Biol. 2008 Dec;39(6):747–57.

37. Kan M, Shumyatcher M, Diwadkar A, Soliman G, Himes BE. Integration of transcriptomic data identifies global and cell-specific asthma-related gene expression signatures. AMIA Annu Symp Proc. 2018;2018:1338–47.

38. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinforma Oxf Engl. 2009 Feb 1;25(3):415–6.

39. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010 Oct 1;26(19):2363–7.

40. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.

41. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J, RStudio, et al. Shiny: Web application framework for R. 2019 [cited 2020 Aug 25]. Available from: https://CRAN.R-project.org/package=shiny

42. Cheng J, Karambelkar B, Xie Y, Wickham H, Russell K, Johnson K, et al. Leaflet: Create interactive web maps with the JavaScript "leaflet" library. 2019 [cited 2020 Aug 25]. Available from: https://CRAN.R-project.org/package=leaflet

43. Gohel D, Panagiotis S, Bostock M, Kokenes S, Schull E. Ggiraph makes 'ggplot' graphics interactive. [cited 2020 Aug 25]. Available from: https://davidgohel.github.io/ggiraph/

44. Wefer SH, Sheppard K. Bioinformatics in high school biology curricula: A study of state science standards. Schulz B, editor. CBE—Life Sci Educ. 2008 Mar;7(1):155–62.

45. Machluf Y, Yarden A. Integrating bioinformatics into senior high school: design principles and implications. Brief Bioinform. 2013 Sep 1;14(5):648–60.

Developing High Performance Secure Multi-Party Computation Protocols in Healthcare: A Case Study of Patient Risk Stratification

Xiao Dong, PhD[1], David A. Randolph, ME[1], Chenkai Weng, BS[3], Abel N. Kho, MD, MS[2], Jennie M. Rogers, PhD[3], Xiao Wang, PhD[3]

[1]Center for Clinical and Translational Science, University of Illinois College of Medicine, Chicago, Illinois, USA; [2]Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA; [3]Department of Computer Science, Northwestern University, Evanston, Illinois, USA

Abstract

We demonstrate that secure multi-party computation (MPC) using garbled circuits is viable technology for solving clinical use cases that require cross-institution data exchange and collaboration. We describe two MPC protocols, based on Yao's garbled circuits and tested using large and realistically synthesized datasets. Linking records using private set intersection (PSI), we compute two metrics often used in patient risk stratification: high utilizer identification (PSI-HU) and comorbidity index calculation (PSI-CI). Cuckoo hashing enables our protocols to achieve extremely fast run times, with answers to clinically meaningful questions produced in minutes instead of hours. Also, our protocols are provably secure against any computationally bounded adversary in a semi-honest setting, the de-facto mode for cross-institution data analytics. Finally, these protocols eliminate the need for an implicitly trusted third-party "honest broker" to mediate the information linkage and exchange.

Introduction

Patient risk stratification is an important task in population health management. The idea is to assign patients in a population to segments based on how much care they are expected to require. With these strata in mind, a health system can allocate resources appropriately to reduce overall cost while improving overall quality of care. Indeed, a recent study[1] found that 5% of the Medicare patients incur nearly half of the program's entire budget. Risk stratification promises not only to reduce this disproportionately high cost, but also to provide more effective and customized care for the patients who need it most.

Given its potential benefits, it is not surprising that risk stratification (or simply identifying high-risk patients) is an active area of research[2]. Two features from clinical records are often leveraged for this classification task[3,4,5]: *comorbidity* (the presence of multiple chronic diseases) and past *high utilization* of healthcare resources (especially expensive emergency department (ED) visits and inpatient hospital stays). Applying these features is complicated when a patient is being seen by multiple healthcare providers who have independent patient data repositories, which lead to siloed perceptions of comorbid conditions and resource utilization.

Such data fragmentation could potentially create significant challenges in quantitative assessment. For example, Brannon et al.[3] complain that an individual healthcare institution is unlikely to identify high utilizers accurately because the ED visits for such patients may be registered at multiple providers and obtaining all ED visits may be impossible due to the difficulty in information exchange among different healthcare providers and matters of patient privacy. Such concerns have been confirmed via a city-wide project carried out in Houston[6], where evidence have shown that high utilizers are indeed visiting multiple healthcare institutions to seek emergency care.

Similarly, complete and accurate data collection is essential to carry out analyses about comorbidity. Several studies[7,8,9] indicate a given hospital's electronic health record (EHR) is unlikely to include all pre-admission comorbidities, since some comorbidities reside within other hospitals' EHR systems. Mitigating such a situation requires cross-institution linkage to capture a complete picture of comorbidities. One study[10] conducted in Italy has attempted to validate a novel comorbidity score using a multisource approach, leveraging cross-hospital linkages enabled by a national beneficiary identifier. More often, however, cross-institution linkage may be challenging for a variety of reasons. For example, Quan et al.[11] report underestimation of comorbidity due to the unavailability of reliable unique identifiers. Lichtensztajn[12] comments on additional burdens in multi-site comorbidity studies, including legal, resource, and procedural barriers to cross-institution record linkage.

At present, the most common method is performing privacy preserving record linkage (PPRL) via a trusted honest broker[13]. Such a method has been used to assemble a city-wide data set from six different institutions to identify high

utilizers in Chicago[4], where linkages are established via matching hash tokens. However, such centralized data aggregation has drawbacks. It requires the participation of a trusted third party and the transmittal of data by all participating institutions to a central data store. This is inherently time consuming and demands considerable resources, along with a complicated multi-institution policy arrangement. Moreover, its centralized approach has security implications[14].

To address these shortcomings, we propose using secure multi-party computation (MPC) approach. Long viewed as a purely theoretical topic and impractical to apply in practice, MPC has recently emerged as a viable alternative thanks to advances in cryptography. Here we apply Yao's protocol[15], based on garbled circuits and oblivious transfer, which we accelerate using state-of-the-art Cuckoo hashing technique to optimize performance. Due to the complicated computation such as comorbidity index involving both logical and arithmetic operations, circuit-based MPC is much more efficient than fully homomorphic encryption schemes[16]. Here, we demonstrate our approach can reach extraordinary speed when studying patient risk stratification in a cross-institution setting. Some prior works[17,18] focused on cross-institution cohort characterization using circuit-based MPC only addressed the fundamental task of secure patient matching – which is a generic application of private set intersection (PSI), devoid of any context for clinical implications and applications.

Other investigators pursuing techniques for privacy-preserving data analytics for healthcare have experimented with deploying MPC in a few healthcare applications. Shi et al[19] demonstrated the viability of performing logistic regression using garbled circuits for EHR data. Several cases also achieve privacy-preserving genomics comparison for similar patient query or disease diagnoses[20]. In addition, cryptographers and privacy experts come together at the annual iDASH Workshop to showcase the latest findings in privacy-preserving analytics – including MPC. These approaches are domain specific, and they only run efficiently in their initial context. On the other hand, they do not readily generalize to additional use cases. Researchers are also identifying tools and techniques to support SQL queries over MPC applied to clinical data research network[21,22]. These systems support a broader class of workloads, although their generality may leave room for significant reduction in query runtime. Our investigation here attempts to achieve both generalizability and high performance at the same time by implementing two building blocks each can be optimized with great efficiency and flexibility to address common use cases in clinical data research. The first building block is a component that performs secure patient matching, and the second is an analytics component that securely computes functions specific to the analytics task at hand, for example calculating comorbidity index.

Background

In this section, we review high utilizers and comorbidity index – two metrics that are essential in patient risk stratification, and we also introduce the key technical components used in our method.

High utilizers refer to a group of patients who disproportionately consume healthcare resources. In general, providers categorize high utilizers in high-risk strata both because of their own health conditions but also the high expenses they typically incur. Although there has no formally agreed-upon definition, we adopt a commonly used criterion that high utilizers have ever visited an ED four or more times within a year[3,4]. Because not including ED visits from other healthcare organizations could potentially render true high utilizers undetected, jointly screening high utilizers across institutions will produce more accurate risk stratification.

Comorbidity refers to the presence of multiple chronic diseases within a single patient. Assessing comorbidity has significant implications for mortality, healthcare utilization, and patient outcomes. The most prevalent comorbidity measurement is the Charlson comorbidity index[23]. In this measurement, several chronic conditions are assigned scores ranging from 1 to 6 according to disease severity. The scores are then aggregated into a single value to measure overall comorbidity. The most common approach leverages algorithms based on ICD (International Classification of Disease) codes[24,25] from the hospital administrative databases. The corresponding scores for each disease are aggregated into a final index value, which is used to represent the patient's overall health status. During patient risk stratification, patients with higher comorbidity index score are put into higher risk strata. Because missing comorbid conditions inevitably lower the overall score, computing comorbidity index across institutions promises more accurate risk stratification.

Secure multi-party computation (MPC) is a cryptographic technique that allows multiple parties to compute a function jointly without revealing anything beyond the output. First introduced by Andrew Yao[15] within the context of Millionaires' Problem, MPC was later generalized by Goldreich, Micali and Wigderson to the multi-party setting[26]. To use an MPC protocol, we need to first represent the function to compute as a Boolean circuit or arithmetic circuit[27]. Until recently, MPC was not considered practical due to the heavy computation and the large amount of

communication required to evaluate even modest functions (simple circuits). In the past decade, the exploitation of hardware acceleration and a series of algorithmic optimizations including free-XOR, half-gates, and OT-extension have improved the efficiency of the garble circuits by several orders of magnitude[28,29,30]. Various specific applications of MPC have been considered practical to use in the real world. In the area of privacy-preserving machine learning[31], MPC has been used between multiple parties who want to train a model based on their combined dataset without exposing any party's data. Also, MPC can achieve the distribution of highly secret information, such as the private key for digital signature[32].

Private set intersection (PSI) is an important application of secure multi-party computation. It allows two parties to compute the intersection of their sets without revealing any elements that are not in the intersection or indeed without exposing any information beyond the intersection. PSI has found a number of practical applications. For example, an instant message (IM) service provider can use PSI to explore a clients' personal network without being able to learn any of the clients' cell phone contacts. Banks can take advantage of PSI to detect malicious mortgage fraud without sharing their clients' data. In healthcare, the most common PSI application is to identify the common patients from different healthcare provider's databases. There exist different PSI protocols that are realized by various techniques, including naïve hash-based PSI, public-key-based PSI, circuit-based PSI and OT-based PSI. Current state-of-the-art PSI protocols are based on Oblivious PRF (OPRF)[33] and Cuckoo hashing[34].

Cuckoo hashing[35] is a hash scheme that avoids hash collision of input values in a hash table. The initialization phase requires the preparation of *n* hash functions $h_1, h_2, \dots, h_n$ and *m* empty bins $B[1, \dots, m]$. To insert an input *x* into the hash table, first hash *x* using all hash functions. Then check if any bin of $B[h_1(x)], \dots, B[h_n(x)]$ is empty. If so, insert the item *x* into that bin, otherwise evict one of the occupying items x' in any bin listed above and insert *x* instead. Then re-enter x' the same way as *x*. Recursively try to insert and evict items based on the hash values until finally all items are accessible through one of the hash tables or an infinite loop is detected.

Method

Our protocol is a combination of Yao's garbled circuits and state-of-the-art PSI protocols. Although the problem to solve in our setting is similar to PSI, there are significant differences. In PSI, we only need to find the common elements. For our task, we need first to find the common elements (i.e., patients visit both healthcare institutions) as in PSI, but we then must compute a function defined for the specific clinical use case we are trying to address (i.e., Unifying data from different institutions for their common patients to identify those who have high Charlson index scores or have visited an ED for at least 4 times within a year). Our protocol consists of following steps:

Input

S^A, S^B: two private patient datasets held by institutions *A* and *B*.

$S^A = (pat_1^A, x_1), \dots (pat_i^A, x_i), \dots (pat_p^A, x_p), |S^A| = p$

$S^B = (pat_1^B, y_1), \dots (pat_j^B, y_j), \dots (pat_q^B, y_q), |S^B| = q$

$F(x_i, y_j)$: use case specific function for $pat_i^A = pat_j^B$

h_1, h_2, h_3: 3 hash functions (public)

Output

$F(x_i, y_j)_{pat_i^A = pat_j^B}$ *for all* pat_i^A *from* S^A, *all* pat_j^B *from* S^B

Preparing *T and T′* for PSI:

Institution *A* uses Cuckoo hashing to generate table *T of size m, m=1.5p*

Institution *B* uses bin-and-ball hashing to generate table *T′ of m* bins

Institution *B* pads all the bins using σ, the maximum bin size

Executing circuit-based PSI and use case specific protocols:

A takes item T_k *in T, where* $h(pat_i^A) = T_k$ *and* $h \in \{h_1, h_2, h_3\}$

B takes bin T_k' *in* T'

PSI protocol checks if $T_k \subseteq T_k'$

If *true* then identify from T_k' *s.t.* $h(pat_j^B) = T_k$

Obtain use case payloads x_i, y_j for pat_i^A and pat_j^B

Execute *CI* or *HU protocols* to evaluate $F(x_i, y_j)_{pat_i^A = pat_j^B}$

Algorithm 1. Cross-institution patient risk stratification MPC protocol.

Algorithm 1 above outlines the protocol to perform the task of patient risk stratification in a cross-institution setting. The inputs include two private datasets from two institutions A and B. The dataset has two parts, the first part is patient data which is used by PSI to link patients; the second part is a payload from the institution's EHR tailored to serve as input to the specific use case's function. (In the two use cases discussed here, these two functions' outputs could be used to perform enhanced risk stratification across two institutions). Two protocol components, High Utilizer (HU) and Charlson Index (CI), introduced below, are used to identify high utilizers or to compute the Charlson comorbidity index. The inputs also include three public hash functions h_1, h_2, h_3 that both institutions have agreed upon in advance.

Prior to executing PSI protocol to link common patients, A and B prepare their inputs using three public hash functions. Assuming A initiates the MPC request, and B responds to the request, A uses Cuckoo hashing to obtain table T, and B uses ordinary bin-and-ball hashing to obtain table T'. Assuming A's patient data set has p patients, the size of T *and* T' are both set to m ($m = 1.5p$). Such choices of $1.5p$ and *three* hash functions[36] are made to ensure the Cuckoo hashing schema only has a failure rate of about $1/2^{40}$ to enter an infinite loop. Using h_1, h_2, h_3, Institution A hashes all its patient records to a Cuckoo hash table T of size m. Here, Cuckoo hashing guarantees each location will have at most one item, and since $m = 1.5p$, one third of T will be empty. Institution B, using the same set of hash functions, hashes all its patient records and obtain a set of bins T_1',..., T_m'. The PSI protocol checks if the patient in T matches any patients at the same location in T', which is carried out inside garbled circuits by performing bit-wise comparison between the items in T and T'.

Figure 1 illustrates PSI process using a concrete example. Assuming the i^{th} patient from institution A and first patient from institution B (*patient 1* highlighted in red) are indeed the same person, i.e., this patient visits providers at both institutions. A uses the Cuckoo hashing to obtain a table T, and B uses bin-and-ball hashing to obtain table T' of the same size. Notice in this example, the i^{th} patient's position in the Cuckoo hash table is obtained using the second hash function h_2. This indicates the i^{th} patient from institution A was evicted once from the position generated by h_1, whereas the j^{th} patient from A has not been evicted since first being hashed by h_1. The same patient's record also exists at institution B (as patient 1), and, because all three hashes are used, one of the items from the bin at the same bin location (also generated by h_2) is guaranteed to match the corresponding entry in A's Cuckoo hash table, as indicated by the red arrow.

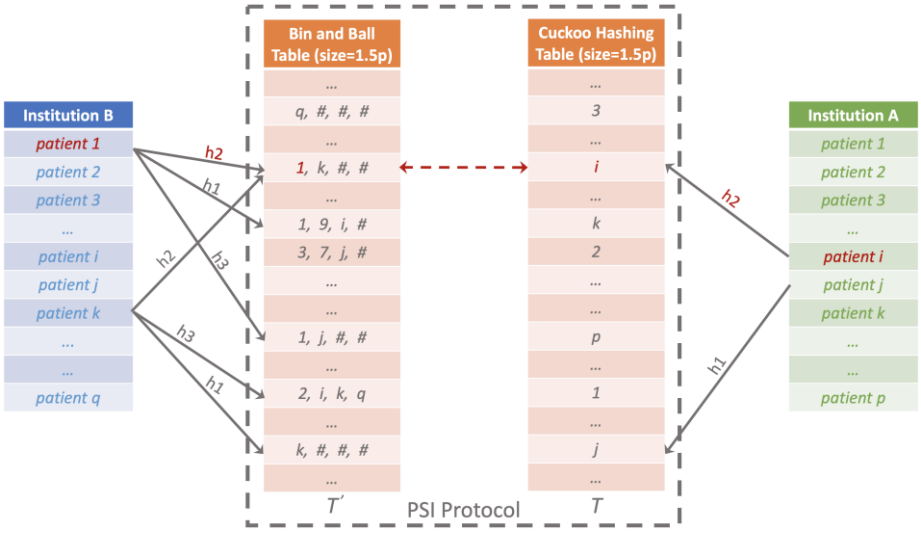


Figure 1. Cuckoo hashing-based PSI protocol.

One caveat here is B's bin sizes may leak private information to A. For instance, an empty bin for B may reveal to A its own patient in the same table location never visit any providers from institution B. To mitigate this risk, all the bins are padded to have the maximum size of B's bins. In the above figure, this is illustrated using "#" to pad each bin to have maximum size of four, assuming the largest bins has four patients mapped in it. This ensures our protocol reveals nothing about either institution's patient records other than which patients match.

The use-case protocol components CI and HU are customized to evaluate specific functions. In CI, additional use-case payload is included to record the comorbid conditions for each patient. The CI circuit component is designed to calculate the dot product of a weight vector of 17 disease categories and each patient's disease vector jointly obtained

from two institutions. The 17 weights of values of 1, 2, 3 or 6 are defined in the Charlson comorbidity index. Using garbled circuits, the disease data payloads for the same patient from two different institutions will be unified together using Boolean union operator. Similarly, using the HU circuit components, the ED visit payloads for the same patient are added together. To keep the size of the payload small, we ignore multiple ED visits to the same provider in the same month. Given this, we scan each (sliding window) 12-month period's combined ED episode to see if the total number of ED visits is ever greater than or equal to four times. If it is, the patient is identified as a high utilizer.

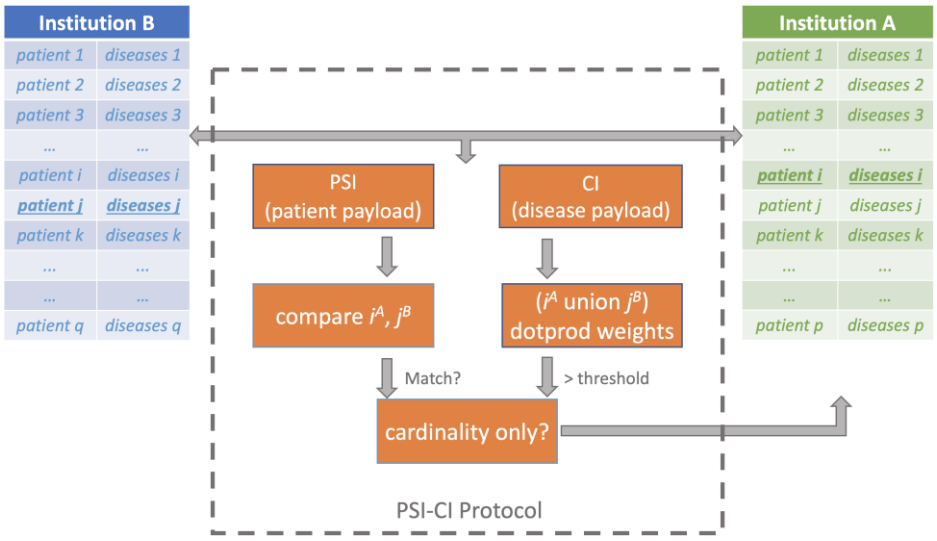


Figure 2. Illustration of Charlson comorbidity index protocol PSI-CI.

Figures 2 and 3 illustrate our two protocols PSI-CI and PSI-HU, respectively. Note each protocol has a PSI component as well as a data analytics component. The PSI component applies on patient data payload and the analytics component (e.g. CI and HU) applies on the use case specific payload. Figure 2 and 3 illustrate the point in time when the protocol performs computation on the i^{th} patient from A and j^{th} patient from B. In addition, each protocol has a regular mode as well as a cardinality-only mode, where only the total number of patients who meet the defined criteria will be the output, instead of a record for each identified patient. By convention, we append a suffix "CA" when this mode is invoked (e.g., PSI-CI-CA and PSI-HU-CA). Using different thresholds, for example a Charlson index score of 9 for high mortality risk and 2 for low mortality risk, the PSI-CI-CA protocol can calculate the size of each risk strata. Similarly, using 6 for super-utilizer, 4 for high-utilizer and 1 for average utilizer, the PSI-HU-CA can characterize the patient risk strata in terms of resource spending. The regular versions of PSI-CI and PSI-HU are used to when the requesting institution needs to identify the individual patients.

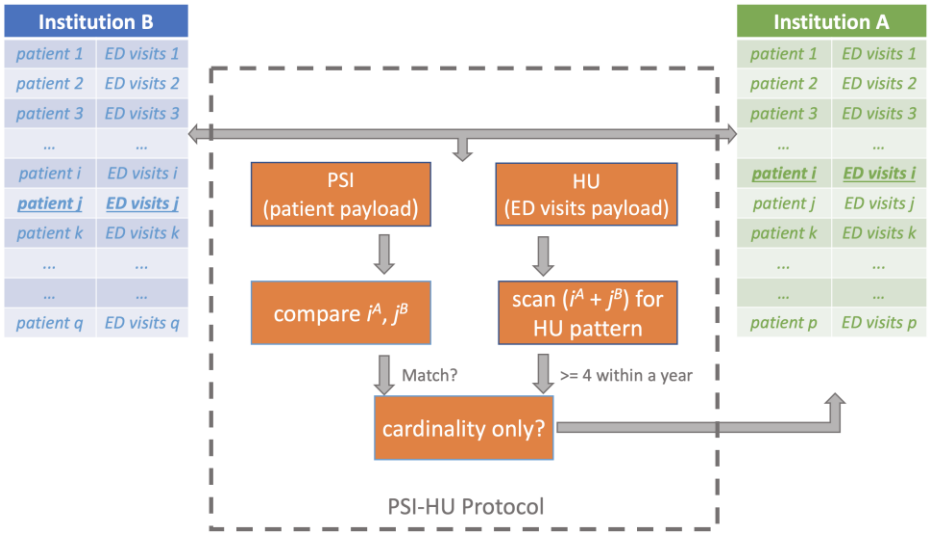


Figure 3. Illustration of high utilizer identification protocol PSI-HU.

In Figures 2 and 3, the protocols only return results to the requesting party A. Under properly arranged data sharing agreements, the outputs can also be shared with the responding party B. In addition, the protocols can be tuned to also output the actual values of Charlson index itself, or the exact total number of ED visits during the 12-month high utilization period. Such patient level results offer finer details in the risk stratification analysis. Furthermore, PSI-CI and PSI-HU could be combined together to derive strata to reflect both disease severity and healthcare resource spending.

Our protocols reveal no additional information about either institution's patient records beyond the mutually agreed-upon outputs. The protocols are secure in the semi-honest setting with static corruption. In a semi-honest setting, we assume the parties involved are curious but not malicious. The parties may try to infer additional information, but they will not intentionally break the protocol or intentionally mislead the computation to produce incorrect results. Semi-honest mode is the de-facto threat model in cross-institution clinical data analytics since full trust cannot be granted to parties outside the covered entities. However, reputational and legal costs should discourage overtly malicious behavior by the parties.

The security of our protocols is promised by the underlying two-party computation protocol, which is the classical Yao's garbled circuits. The protocols return accurate results under these assumptions. To implement the protocol, we use the EMP-toolkit[37], which provides a suite of tools for executing two-party and multi-party computation protocols efficiently. This toolkit includes state-of-the-art protocol and implementation optimizations. We use AES-based hash function to compute the Cuckoo hash table for improved efficiency.

We implement our computation in EMP functions, where all operations are overloaded to execute cryptography operations instead of computation in clear. To avoid any side-channel leakage, all computation has to be oblivious, that is, the behavior of the computation cannot depend on the payload data. Usually, this will reduce the efficiency of the computation since we need to incur the worst-case running time regardless of the computation. To alleviate this slowdown, we perform the optimization as follows: for each bin from T and T', we first perform PSI to see whether there are matching patients, if yes then compute the use case function only *once* using the matched patients' payloads, otherwise merely compute the same function using the first patient's payloads from each bin. This optimization significantly reduces the number of comparisons while keeping the correctness and security.

Data Preparation. In order to simulate a real-world multi-institution setting, we use Synthea[38] to create a large-scale, realistically synthesized dataset. The Synthea platform deploys publicly available health statistics and clinical practice guidelines to simulate large scale electronic healthcare data, including modeled interactions between patients, providers, payers, etc. over the lifespan of synthetic patients. Synthea applies public census and health data to generate patient population with realistic demographics and leverages expert-curated disease packages to model disease progression throughout the patient's lifetime. The synthetic data produced by Synthea has no privacy or security restrictions.

We use Synthea to generate a Chicago city-wide database consisting of approximately 2.7 million patients and ~141 million encounters. Records of 1.3 million ED visits over three years are processed to test the performance of our protocols. To simulate patients arriving at EDs in different institutions for different visits, we took the longitudes and latitudes for the ED departments from seven large academic hospitals in Chicago. After measuring the distance from those ED departments to the synthesized patients' addresses at ED visits, an ED is chosen randomly from the two closest EDs. The encounter records are then split into seven data sets, each representing the institution's own ED encounter data sets during the 3 year period. Datasets for two institutions that are geologically very close are chosen to test the performance of our protocols. These two datasets contain 120,666 and 109,072 patients respectively. Using these two datasets, we generate patient payloads for both PSI-CI and PSI-HU in order to allow direct performance comparison between them as shown in the next section.

To prepare the patient (PSI) payload, each patient is represented by a 64-bit integer. Although the synthetic data contains an SSN field that is an excellent choice for cross-institution patient matching and it only takes 32-bit to represent, we consider the factor that participating institutions of MPC may want to avoid using SSN directly due to its sensitive nature, even though only AES encrypted bits of SSN leaves the institutional boundary during the protocol execution. Moreover, SSN values are not reliably available in practice. The 64-bit integer is obtained using the first 64 bits of the SHA256 hashes generated from patients' name, SSN, birthday. The combination of such demographics data fields ensures the hash tokens can uniquely identify patients while allowing patient payload to be reasonably lightweight. The chance of a single collision occurs in our data is thus approximately $120{,}000/2^{32} = 0.003\%$, where $1/2^{32}$ is derived from the birthday paradox.

To construct the use case payload for the PSI-HU protocol, we encode patients' presence at ED each month. This results in a Boolean vector of length 36 for the last three years for each patient. (This renders additional visits to the same institution within the same month invisible. Using additional bits would allow for a more accurate accounting, at the expense of greater memory needs.) For the PSI-CI protocol, we use a Boolean vector of length 17 to encode the presence of each disease. Here we used the 17 comorbid conditions in the modified index[24] instead of the 19 conditions in the original index[23]. At the time of this writing, Synthea only contains 10 conditions with the highest morbidity in the US, and many conditions required in Charlson index calculations – such as chronic pulmonary disease, rheumatologic disease, peptic ulcer disease – are not currently included as Synthea disease modules. Therefore, instead of using the disease output from the Synthea, we impute the distribution of 17 conditions from published disease percentage statistics[11]. Further details can be seen at our demo website (see footnote of this page).

Results

Table 1 below shows the main performance results for our protocols. Four main categories – running time, circuit size, network traffic and memory – are evaluated for PSI-CI and PSI-HU. Here we test both the full and cardinality-only versions for each protocol. Using large-scale realistic datasets, the protocols complete in around 3 minutes and 7 minutes, respectively. The memory usages are 2.9 GB for PSI-CI and 3.5 GB for PSI-HU, acceptable on even basic consumer systems. PSI-HU requires slightly more memory since its payload is larger (36 bits versus 17 bits for the use case payload). Network communication depends only on the size of the circuit, which primarily depends on the complexity of the use case specific function. Hence, the high utilizer protocol PSI-HU also requires more network resources than the PSI-CI protocol because the HU circuit function is more complicated. Additional overhead in running time, circuit size, and communication cost are present in the cardinality-only protocols because of the additional counter components required in the circuits.

Table 1. Performance testing results for our secure computation protocols, max padding size is 11.

Protocol	**Running Time** (Seconds)	**Circuit Size** (Number of AND Gates)	**Network Traffic** (GB)	**Memory** (GB)
PSI-CI	193	163,608,000	5.7	2.9
PSI-CI-CA	203	173,751,634	6.0	2.9
PSI-HU	422	378,098,088	12.7	3.5
PSI-HU-CA	432	388,241,722	13.0	3.5

In order to prevent information leak, all the bins in institution B's bin-and-ball table are padded to the maximum size, as described above. The maximum bin size is determined by B's patient data payload and hashing functions. Larger patient datasets require larger padding sizes. Using our hashing functions and data sets where institution A has 120,666 patients and B has 109,072 patients, the maximum padding size is 11. For the above experiments, we use simulated average network delay and bandwidth of 96 *ms* (roundtrip) and 50 *MB/s* in a WAN setting. We adopt these simulation parameters from another study[33] that uses such settings to evaluate a PSI system. Our experiments were performed on a system with an Intel i7-8750H 2.2 GHz CPU and 16 GB memory. The software for the presented protocols, including the synthetic dataset and programs to reproduce our results, can be found on GitHub[†].

Discussion

In this paper, we demonstrate the practicability and flexibility of state-of-the-art MPC approaches for real-world clinical use cases. Our approach meets requirements for real-world deployment, as it is performant using affordable computing resources. Our most complex protocol completes in only a little over seven minutes, running over a large dataset. Compared to two early stage circuit-based clinical informatics applications for the simpler tasks of PSI[17] and joint cohort discovery[18], our results achieved improvements of an order of magnitude. Table 2 captures the performance of our most complex protocol and the published results from these two earlier garbled circuits systems. Our performance advantage is obtained in spite of using patient data sets that are more than 10 times larger, having higher precision patient payloads, adding analytics workload, and being simulated in WAN setting.

[†] https://github.com/dongxiao/MPC-RiskStratification

Table 2. Comparison between our highly optimized protocol and two previous circuit-based MPC systems.

	Running Time	Data Size	Patient Payload	Use Case	Network	Security
PSI-HU-CA	432 seconds	121K,109K	64-bit	36-bit add, scan	WAN	128 bits (AES)
Dong et al.[18]	~2 hours	10K, 10K	36-bit	N/A	LAN	128 bits (AES)
Chen et al.[17]	~3 hours	10K, 10K	36-bit	N/A	LAN	80 bits

Our performance improvement is largely due to our use of Cuckoo hashing. With Cuckoo hashing, the number of required patient comparisons is drastically reduced. In the previous work, naïve patient-by-patient comparisons between the two data sets implies a complexity of $O(N^2)$, where N is the number of patients at an institution. In our work, the complexity is $O(1.5\sigma N)$, where σ is the maximum padding size. The maximum padding size σ grow roughly logarithmically as the dataset grows. Using Cuckoo hashing, we cut the running time by a factor of $2N/3\sigma$. In our experiments, the patient comparison takes place 1,799,688 times, whereas in the naïve method of pairwise comparison that needs to take place 109702×120666 times, therefore using our Cuckoo hashing scheme the patient comparison workload is reduced by a factor of 7355 compared to naïve methods used in those two previous works. Obviously, our performance here is aided somewhat by simplifying the patient linkage (PPRL) step to PSI. To be practical, this would require the existence of unique identifiers for patients used by both institutions. As discussed above, universal identifiers are rarely available. However, deterministic linkage implementations like the one suggested by Kho and associates[13] are amenable to optimization by Cuckoo hashing, as they require perfect matching of at least one of several concatenations of hashed PHI elements. In their initial implementation, they suggest four such concatenations, which would require four pairs of hash tables under our approach and raise the complexity to $O(4\times1.5\sigma N)$, which is still $O(N)$. But we must concede that methods for linking patients will be constrained under our optimization.

On the other hand, because all computable functions can be expressed as a Boolean circuit, our approach can be adapted flexibly to other clinical use cases beyond patient risk stratification. As shown in the results section, it only takes a few gigabytes memory per site to run a secure protocol on a relatively large dataset. This suggests our approach is ready to handle protocols using payloads about one order of magnitude larger, while still running on very affordable hardware. It is also worth noting the disease cohorts often seen in pre-clinical trial studies are much smaller – typically on the order of hundreds or thousands of patients. This indicates our approach is ready to provide circuit-based MPC solutions to many real-world multi-site clinical studies.

The security of our protocol is based on the underlying premise of Yao's garbled circuits, which is provably secure in a semi-honest setting. From a correctness perspective, it returns correct results all the time. The only caveat is that, as established by Demmler et al.[36], building a Cuckoo hash table with *three* hash functions along with the factor of *1.5* has a failure rate of about $1/2^{40}$. A failure occurs when a patient cannot fit into the Cuckoo hash table. In this extremely unlikely event (approximately 1 in a trillion), the protocol could easily re-generate the Cuckoo hash table with a new set of hash functions. Such an event has no apparent security or privacy implications.

For future work, we plan to develop new protocols to incorporate heterogeneous data domains to handle complex scenarios that not only involve traditional clinical data, but also data from genomics, mobile health, and social determinants of health. Since our protocols are memory efficient, this leaves a lot of room for expanding payloads to represent more complex clinical data elements, such as medications and lab tests. EMP-toolkit boasts support for MPC involving multiple parties, so we plan to study its application to more complex multi-site clinical scenarios.

As the clinical data model becomes more sophisticated, it will inevitably require more memory friendly payload designs. So such space optimization promises to be another avenue for research.

Finally, we plan to compare our protocol to existing trusted third-party approaches. Abel and associates[13] report matching patients among seven institutions takes significant computational time. We would like to perform controlled experiments to compare our circuit-based approach directly with such standard PPRL systems. We note our protocol is provably secure in a semi-honest setting, whereas the centralized third-party approach has demonstrated security vulnerabilities in the same setting[14].

Conclusion

Having demonstrated their performance with measures important for patient risk stratification, we assert the protocols developed based on garbled circuits are ready for real world deployment in clinical informatics. In the context of discussing the adoption of blockchain, Kuo and Ohno-Machado[39] argue there are two ways for cutting-edge secure computing technology to make a real-world impact. First, sustained trust with policy and regulatory bodies must be

built. Second, early adopters must be encouraged, so they may serve to allay the fears of institutions to adopt who feel hesitant. Breakthroughs in cryptography have made circuit-based MPC a completely viable technology for healthcare. Well-defined clinical use cases and provably secure processing of patient data should help build trust with governing bodies in healthcare and may help streamline review processes. The time to champion the adoption of circuit-based MPC protocols in real-world multi-site clinical analytics is upon us.

Acknowledgement

This research was supported by NCATS CTSA grant UL1TR002003 and NCATS ACT grant UL1TR000005.

References

1. Peter S. 5 percent of Medicaid patients account for half of program's costs. The Hill. Retrieved from https://thehill.com/policy/healthcare/241491-5-percent-of-medicaid-patients-account-for-50-percent-of-costs. Published 2015.
2. Chong JL, Lim KK, Matchar DB. Population segmentation based on healthcare needs: a systematic review. Syst Rev. 2019;8(1):202.
3. Brannon E, Wang T, Lapedis J, et al. Towards a learning health system to reduce emergency department visits at a population level. AMIA Annual Symposium Proceedings. 2018;295-304.
4. Ghosh A, Jackson K, Balsley K, Kho AN, Walunas T. Identification of risk factors for high utilization of healthcare in diabetic patients using an integrated medical records database. AMIA Annual Symposium Proceedings. 2012;504.
5. Jean-Baptiste D, O-Malley AS, Shah T. Population segmentation and targeting of health care resources: findings from a literature review. Math Policy Res. 2017;(Working Paper 58):1-50.
6. Using Technology to Transform Health Care in Houston. Retrieved from https://www.pcictx.org/pcic-media/item/55-using-technology-to-transform-health-care-in-houston. Published 2017.
7. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for handling missing data in electronic health record derived data. eGEMs (Generating Evid Methods to Improv patient outcomes). 2013;1(3):7.
8. Wong ML, McMurry TL, Schumacher JR, et al. Comorbidity assessment in the national cancer database for patients with surgically resected breast, colorectal, or lung Cancer (AFT-01, -02, -03). J Oncol Pract. 2018;14(10):e631-e643.
9. Corser W, Sikorskii A, Olomu A, Stommel M, Proden C, Holmes-Rovner M. Concordance between comorbidity data from patient self-report interviews and medical record documentation. BMC Health Serv Res. 2008;8:1-9.
10. Corrao G, Rea F, Di Martino M, et al. Developing and validating a novel multisource comorbidity score from administrative data: A large population-based cohort study from Italy. BMJ Open. 2017;7(12):1-8.
11. Quan H, Li B, Couris CM, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol. 2011;173(6):676-682.
12. Lichtensztajn DY, Giddings BM, Morris CR, Parikh-Patel A, Kizer KW. Comorbidity index in central cancer registries: The value of hospital discharge data. Clin Epidemiol. 2017.
13. Kho AN, Cashy JP, Jackson KL, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Informatics Assoc. 2015;22(5):1072-1080.
14. Dong X, Randolph DA, Rajanna S. Enabling privacy preserving record linkage systems using asymmetric key cryptography. AMIA Annual Symposium Proceedings. 2019;380-388.
15. Yao AC. Protocols for secure computations (Extended Abstract). 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982).
16. Gentry C. Fully homomorphic encryption using ideal lattices. In: Proceedings of the Annual ACM Symposium on Theory of Computing. ; 2009.
17. Chen F, Jiang X, Wang S, et al. Perfectly secure and efficient two-party electronic-health-record linkage. IEEE Internet Comput. 2018;22(2):32-41.
18. Dong X, Randolph DA, Wang X. The feasibility of garbled circuits for cross-site clinical data analytics. AMIA Informatics Summit Proceedings. 2020;788-789.
19. Shi H, Jiang C, Dai W, et al. Secure Multi-pArty Computation Grid LOgistic REgression (SMAC-GLORE). BMC Med Inform Decis Mak. 2016.
20. Jagadeesh KA, Wu DJ, Birgmeier JA, Boneh D, Bejerano G. Deriving genomic diagnoses without revealing patient genomes. Science (80-). 2017.

21. Bater J, Elliott G, Eggen C, Goel S, Kho A, Rogers J. SMCQL: Secure querying for federated databases. In: Proceedings of the VLDB Endowment 10.6 (2017): 673-684.
22. Bater J, He X, Ehrich W, Machanavajjhala A, Rogers J. Shrinkwrap. Proceedings of the VLDB Endowment 12.3 (2018): 307-320.
23. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J Chronic Dis. 1987.
24. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol. 1992.
25. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005;43(11):1130-1139.
26. Goldreich O, Micali S, Wigderson A. How to play any mental game. STOC 1987: Proceedings of the nineteenth annual ACM symposium on Theory of computing; 1987.
27. Beaver D. Efficient multiparty protocols using circuit randomization. CRYPTO 1991. In: Lecture Notes in Computer Science, vol 576. Springer, Berlin, Heidelberg.
28. Kolesnikov V, Schneider T. Improved garbled circuit: Free XOR gates and applications. IICALP 2008. In: Lecture Notes in Computer Science.
29. Zahur S, Rosulek M, Evans D. Two halves make a whole reducing data transfer in garbled circuits using half gates. EUROCRYPT 2015. In: Lecture Notes in Computer Science 2015.
30. Ishai Y, Kilian J, Nissim K, Petrank E. Extending oblivious transfers efficiently. CRYPTO 2003. In: Lecture Notes in Computer Science 2003.
31. Mohassel P, Zhang Y. SecureML: A system for scalable privacy-preserving machine learning. In: Proceedings - IEEE Symposium on Security and Privacy. ; 2017.
32. Lindell Y. Fast secure two-party ECDSA signing. In: Lecture Notes in Computer Science. Annual International Cryptology Conference; 2017.
33. Kolesnikov V, Kumaresan R, Rosulek M, Trieu N. Efficient batched oblivious PRF with applications to private set intersection. In: Proceedings of the ACM Conference on Computer and Communications Security. ; 2016.
34. Pinkas B, Schneider T, Segev G, Zohner M. Phasing: Private set intersection using permutation-based hashing. In: Proceedings of the 24th USENIX Security Symposium. ; 2015.
35. Pagh R, Rodler FF. Cuckoo hashing. J Algorithms. 2004.
36. Demmler D, Rindal P, Rosulek M, Trieu N. PIR-PSI: Scaling private contact discovery. Proc Priv Enhancing Technol. 2018.
37. Wang X, Malozemoff AJ, Katz J. EMP-toolkit: Efficient multiparty computation toolkit. [Internet]. Available from: https://github.com/emp-toolkit.
38. Walonoski J, Kramer M, Nichols J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Informatics Assoc. 2018.
39. Kuo T, Ohno-Machado L. Blockchain in biomedical, healthcare and genomic applications, Workshop 25, AMIA Annual Symposium, November 17, 2019. Washington DC.

Evaluating Organizational Readiness for Change in the Implementation of Telehealth and mobile Health Interventions for Chronic Disease Management

Ibukun E. Fowe, MBChB[1], MSGH[2], PhD student[3]
OHSU-PSU School of Public Health[4]

Introduction

Advancements in digital technology innovations have resulted in an increasing ability for Telehealth and mobile health (mHealth) technologies to provide efficient and effective healthcare [1]. This has resulted in a growing interest in the use of these technologies to develop solutions that can improve the management of chronic conditions [1,2]. A high chronic disease burden within the US, an increasingly aging population, high costs of chronic disease care, challenges with timely and effective access to care, the need to obtain a holistic picture of a patient's health outside of the health care setting, and a fragmented health care system have led to an increase in interest toward the use of these technologies [1,3,4]. Furthermore, given the disproportionate burden of chronic conditions among older adults and the projected increase in the number of Americans age 65 and older to 98 million by year 2060, there is an increased need for innovative approaches to improving the management of chronic diseases within the US [5–7].

Based on this, Telehealth and mHealth solutions that can facilitate increased self-care and self-management through the use of remote patient monitoring (RPM) technologies among the older adult US population are being developed. This is also projected to be able to potentially address the dis proportionate burden of care giving on family members and a projected decrease in care giver population given the increasingly aging population [6–8]. Although chronic conditions are prevalent among older adults, they tend to begin in early adulthood, and this makes it important to begin to address these issues in early adulthood. These challenges have reiterated the need to proactively provide avenues for self-management and remote monitoring of these chronic conditions that affect adults and older adult US populations through the use of innovative digital health technologies [1,7].

Telehealth and mHealth interventions have been observed to enable patients' self-monitoring, self-management and provider enabled remote monitoring [9]. Prior to the COVID-19 pandemic a good number of Telehealth and mHealth interventions were focused on chronic conditions that involved the cardiovascular, respiratory, neurological and endocrine systems such as hypertension, Chronic Obstructive Pulmonary Disease (COPD), Alzheimer/dementia, and diabetes. Currently, due to the ongoing COVID-19 pandemic there has been an increased interest in the use of Telehealth and mHealth interventions for the management of acute conditions in a bid to maintain access to health care services while reducing the potential for the spread of the virus caused by SARS-CoV-2. Also, policy makers at the global and national level such as the WHO and the US Center for Medicaid Services (CMS) have reiterated the importance of Telehealth and mHealth services to reduce the spread of the virus and maintain access to care irrespective of a patient's location [10,11]. Based on this, there has been a lessening of restrictions on Telehealth and mHealth services in the US and other parts of the globe with HCOs rising up speedily to meet the upsurge in demand for virtual visits, remote symptom tracking and evaluation and RPM. Based on these, Telehealth and mHealth interventions show promise for maintaining access to care during the pandemic as well as after the pandemic, hence, these seemingly temporary changes raise some pertinent questions, to clinicians, patients, hospital administrators and other stakeholders, one of which is how organizations can sustain this trend post COVID-19 pandemic by developing and implementing sustainable Telehealth and mHealth interventions that are best suited for their health organizations, staff populations, and their level of Telehealth or mHealth organizational readiness. It is important for HCOs to be aware of organizational factors that are key to enabling sustained adoption and implementation of Telehealth and mHealth interventions so as to be able to design or adopt Telehealth or mHealth interventions that their HCO is ready for and has the capacity to maintain and sustain its use over an appreciable period of time.

Studies have shown that Organizational Readiness for Change (ORC) is key to organizations' ability to successfully adopt, implement and sustain innovative technology solutions such as Telehealth and mHealth interventions [12–14]. ORC has been defined as a multi-level and multi-faceted construct that involves organizational members' shared resolve and commitment to implement a specific change or a set of changes and their shared belief in their collective capability to do so [15]. ORC also refers to the degree to which members of an organization are prepared psychologically and behaviorally to implement organizational change [14]. ORC varies based on how much organizational members value the change and how they perceive determinants of implementation capability such as availability of resources, task demands, and situational factors [14]. A high level of ORC is more likely to result in organizational members

initiating change, exerting greater effort, exhibiting greater persistence, and displaying more cooperative behavior, which results in a more effective and successful implementation [13,15,16].

One of the key Implementation Science (IS) frameworks that has been used to evaluate ORC is the Consolidated Framework for Implementation Research (CFIR). The CFIR was developed by unifying several IS theories, and provides a pragmatic and unifying framework for assessing organizational readiness factors that have been observed to be key to the successful adoption or implementation of new interventions [17,18]. The CFIR offers a comprehensive, unifying taxonomy of constructs related to the intervention (e.g. Telehealth or mHealth intervention), the organization's inner and outer settings, the characteristics of involved individuals (such as members of staff or implementers), and the implementation process [17]. These CFIR domains interact in rich and complex ways to influence implementation effectiveness and are useful for planning or evaluating implementation processes.

Despite the importance of evaluating ORC prior to and during the implementation of Telehealth and mHealth interventions, prior studies have been largely focused on patients' or end users' acceptance of Telehealth or mHealth technology, the effectiveness or efficiency of the Telehealth or mHealth technologies in aiding remote or virtual diagnoses and treatment, and the ability of the technologies to enable RPM, and aid patients' self-management [19,20]. Few studies have evaluated organizational factors in the implementation of Telehealth or mHealth interventions within US HCOs that can influence readiness for change in the implementation of these technologies.

Given the increasing burden of chronic diseases, particularly among the adult and older adult US population and the increasing need for enhanced and enabled self-management of chronic conditions in this population, this systematic review of literature was conducted to understand organizational factors that can enable ORC in the implementation of Telehealth and mHealth interventions for chronic disease management among US adults and older adults. Study findings were organized and analyzed using the CFIR this is with a view to enabling implementers to have an understanding of organizational factors that can increase ORC and thereby increase the potential for successful implementation and sustained adoption. Findings from this review can benefit HCOs as they decide on Telehealth and mHealth interventions to adopt or implement. It can also assist HCOs to know what organizational factors to look out for when evaluating their readiness for change by presenting them with an up to date and comprehensive review of key organizational factors that impact the adoption and implementation of Telehealth and mHealth interventions as reported in the academic literature. This can inform and guide HCOs in the development of strategies for promoting the adoption of these tools and enable them to realize their potential benefits.

Methods

This study conducted a systematic review of study articles focused on evaluating organizational factors that are important in evaluating ORC in the implementation of Telehealth and mHealth interventions for chronic disease management using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline. This review focused on Telehealth or mHealth interventions designed for US adults and older adults for the management of their chronic conditions by HCOs such as hospitals or physician practices. The CFIR was used in this study to organize and analyze identified organizational factors. The CFIR given its robust set of constructs offered an avenue to clearly elicit organizational factors that influenced readiness for change in the implementation of Telehealth and mHealth interventions within HCOs. The CFIR's domain and constructs was used to evaluate identified internal and external organizational facilitators and barriers to implementation within the selected articles in order to elicit how these factors affected ORC [17].

Eligible studies for this review included peer-reviewed articles that 1) were written in English language; 2) were focused on the implementation or deployment of community facing Telehealth or mHealth interventions by healthcare related organizations; and, 3) involved the description of organizational facilitators and barriers to implementation and change in the implementation of home or community facing Telehealth or mHealth interventions for chronic disease management. Study population included adults ages 18 years and older with one or more chronic conditions who lived in their homes or in communities. This study excluded articles that were 1) non-original studies or generic reports that were not focused on home or community facing Telehealth or mHealth interventions; 2) articles or reviews that were focused on patients' or providers' perspective of the effectiveness or efficiency of the Telehealth or mHealth intervention; and 3) studies that were conducted outside of the US.

Electronic databases and search engines such as PubMed, CINAHL, SCOPUS and Web of Science were searched for potentially eligible articles. Furthermore, experts in the use of electronic databases for conducting systematic review related searches were contacted and asked to assist in database searches as well as suggest additional relevant articles that had not been included. Duplicate articles were excluded at each stage of the search process. An initial search was conducted in November 2019, with follow up searches carried out in February 2020. Search strategy included keywords and terms such as "Telehealth, telemedicine, mHealth, ehealth, chronic disease, condition, or illness," which were searched in any combination. These search terms were identified during a preliminary search of the literature focused on discovering the various terms used in articles related to utilizing IS frameworks for evaluating organizational readiness for change in the implementation of Telehealth or mHealth interventions for chronic disease management. Specific implementation science related keywords were however not included in the search in order to broaden the scope of the search. Also, the search was not limited to any particular time frame in order to enable an exhaustive search. Filters were used in all searches to exclude non-English articles, or articles that were not peer-reviewed.

All titles and abstracts were reviewed by the author using the systematic review software – Covidence, and irrelevant publications were excluded. Then both the author and a study colleague using Covidence conducted a full text review of the selected articles. Study articles were included in the study when the author and the study colleague were in agreement. When they were not in agreement, both the author and the study colleague conducted a second review and subsequently made a decision. If there was doubt, the author conducted another review and subsequently made a decision. Data extraction was conducted with the use of a data extraction form that was developed by the author for the purpose of this review. The data extraction form included parameters that were related to study details such as study design, setting, population, intervention, implementation findings, and organizational factors that affected the implementation and adoption of Telehealth and mHealth interventions. Identified organizational barriers and facilitators of the implementation of Telehealth and mHealth interventions were then organized and analyzed using CFIR's five domains and constructs in order to elicit how these factors affected ORC in the implementation of Telehealth and mHealth interventions for chronic disease management among US adults and older adults. The data extraction was guided by the aims of the review, which is focused on understanding 1) organizational factors that influenced the implementation of Telehealth and mHealth interventions for chronic disease management among US adults and older adults, and, 2) how the elicited organizational factors influenced ORC using the CFIR's five domains and constructs. The identified organizational factors were sorted according to the CFIR's domains and assessed based on CFIR's constructs that corresponded with or matched each elicited factor.

Results

The search generated 10,162 references, and 3753 duplicates were excluded. 6212 irrelevant studies were excluded based on title abstract screening due to the inconsistencies in terminologies used in the literature on Telehealth and mHealth implementation. Also, few numbers of US original studies were identified that were focused on organizational factors that affected implementation or ORC in the implementation of Telehealth or mHealth interventions. 197 full studies were selected for full text evaluation. Following evaluation by the first author, and a study colleague, 13 studies met all the inclusion criteria and were included in the study. (See Figure 1 beneath for details on the selection process).

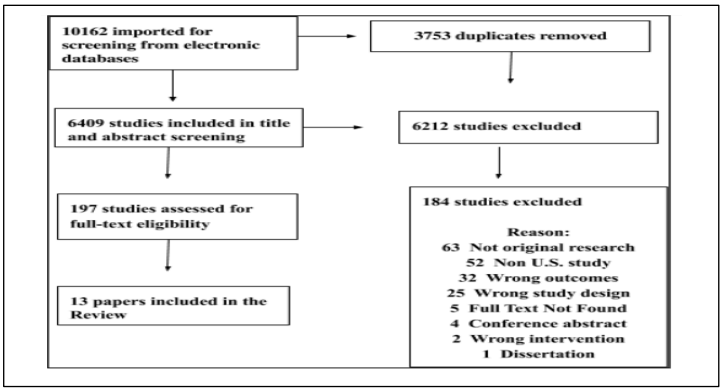


Figure 1. – Systematic review flow diagram. (Prisma Chart showing the selection process)

All included studies were published between 2004 and 2019. Nine of the studies used qualitative research methods [21–29], three studies used mixed methods approaches [30–32], and one study used quantitative surveys [33]. Four of the included studies were pre-implementation or feasibility assessment studies [24,28,30,32], three studies evaluated pilot interventions [21,25,33], four studies were intervention or program evaluation studies and spanned from early to late phases of implementation [23,26,27,29], one study was an end of intervention evaluation [31], while one study was a prospective study [22].

Seven of the studies involved Telemedicine, Telemonitoring, or videoconferencing interventions [23–25,27,29,31,32]. Three studies involved Telehealth interventions [22,26,33]. Two studies involved mHealth interventions [21,30], and one intervention was a Telehealth kiosk [28]. Five studies occurred in clinics, primary care settings or ambulatory practices that were affiliated with larger academic centers [21,23,24,30,33], while three studies involved Veterans Health Affairs (VHA) networks of medical centers and clinics [27,29,32]. Three studies involved home care programs [25,26,31]. One study involved a senior living facility; and one study involved grant funding agencies and two recipient clinical sites [22,28].

Five of the studies used IS frameworks or organizational theories and frameworks to organize their study, or to guide the analysis and interpretation of implementation study findings and discuss study or implementation outcomes. One study used the diffusion of innovations framework to guide a pilot evaluation which was aimed at implementing telehealth services in a clinic within a large academic center [33]. A second study used the CFIR to evaluate barriers and facilitators to the implementation and spread of a video-conferencing intervention and to distinguish between high and low performing sites [29]. A third study used the conceptual framework on interorganizational relations and resource dependency to organize data collection for an end of program evaluation of a home care Telemedicine program [31]. A fourth study used the Weiner Organizational Theory of Implementation Effectiveness to conduct a pre-implementation study and needs assessment of a Telemedicine program that involved a network of integrated VHA sites [32]. Lastly, a fifth study used the Realist evaluation approach and user-task-context usability framework to understand the factors associated with a failed pilot in a Federally Qualified Health Center [21].

Six studies reported the use of implementation strategies prior to or during the implementation of interventions. Some of the reported organizational implementation strategies reported included conducting a needs assessment, establishing a needs-based practice protocol, and staff training [33]. Others include, a feasibility assessment of potential recipients and managers; an evaluation of staff acceptance of telemedicine in the early implementation phase; the use of the health belief model to frame participants' perceptions of the intervention and identify barriers and facilitators, and the use of onsite champions, awareness outreach, staff education, and hands-on training [22,24,27,28]. Technical implementation strategies included simple app design, feasibility testing, and the use of an accompanying practice model to enable ease of integration into existing workflows [30].

All of the studies included, identified, organizational barriers, facilitators or challenges that affect the implementation of Telehealth or mHealth intervention. Nine studies identified barriers and facilitators in the implementation of mHealth or Telehealth interventions. One study identified facilitators mainly and used them in the eventual intervention implementation post pilot phase [33]. Another study identified barriers and opportunities, while two studies were focused on identifying implementation challenges mainly [24,25,27].

This study defined ORC as the degree to which organizational structures, environmental settings and members seemed ready to support and facilitate the Telehealth or mHealth intervention. Based on this, the study identified organizational factors that are relevant in the evaluation of ORC in the implementation of Telehealth and mHealth interventions for chronic disease management among US adults and older adults using the CFIR framework. Emerging CFIR constructs in this review were defined based on the facilitators and barriers identified across the studies and then organized and evaluated using the CFIR framework [18]

The next section begins with a table showing how the elicited barriers and facilitators were organized and their influence on ORC that was elicited based on the CFIR (Table I). This will be followed by an enumeration of the elicited facilitators and barriers of ORC in each CFIR domain. The discussion section will then describe the importance of these factors in assessing ORC. Finally, based on the elicited facilitators and barriers of Telehealth or mHealth interventions in this review, specific recommendations to implementers on assessing ORC are discussed.

Table I. Using the CFIRs domains and constructs to elicit ORC.

CFIR Domains	Definition based on this review	Identified CFIR Constructs	Effects on ORC
Intervention Characteristics	Specific factors related to the Telehealth or mHealth intervention that affect its design, implementation and how it is received by end-users.	Design quality and packaging, complexity, cost.	Assesses the fit of the intervention with organizational infrastructure, and recipient populations.
Outer setting	Features of the external environment or context of an organization that can influence successful implementation.	External policy and incentives, Patients' needs and resources, and cosmopolitanism.	Assesses external support for the intervention.
Inner setting	The organizational context in which the intervention exists.	Structural characteristics, networks and communications, culture, implementation climate and implementation readiness.	Assesses organizational structure's readiness for the intervention.
Process	Course of actions that need to be carried out to achieve the set organizational or individual goals for the intervention.	Planning, engaging, executing, and reflecting and evaluating.	Assesses readiness for smooth intervention implementation and operation.
Characteristics of Individuals	Specific features of individuals that are involved in the implementation and use of the intervention refers to individual related factors of the intervention's implementation.	Knowledge and beliefs about the intervention, self-efficacy, individual stage of change, individual identification with the implementing organization, and other personal attributes.	Assesses workforce readiness for the intervention and the expected level of ease of change.

i. *Intervention characteristics*

The intervention is the specific mHealth or Telehealth solution that is designed to address a health issue or to achieve a desired health related or health care outcome. Intervention characteristics are the specific peculiarities of the Telehealth or mHealth intervention that is related to the intervention's design and can affect how and where it can be implemented [34]. Based on the studies included in this review, the identified CFIR related intervention characteristics constructs are design quality and packaging, complexity, and cost [34].

Design quality and packaging related facilitating factors identified in the review include simplicity of design, ease of use and easy to use clinical guidelines [30]. Barriers included device usability challenges for providers and patients, lack of usability testing, poor fit between the intervention and end users, and a cumbersome installation process [21,26]. Complexity inhibiting factors in this review include easier to use technology platforms, and an informed patient-clinician interaction [26,30]. Complexity related barriers include, time consuming processes, difficulties with accessing the intervention, the need to hire new staff, the need to acquire new equipment, and the need to change existing work schedules or job duties to accommodate the intervention, [21,25,27,29]. Cost related facilitating factors for this domain include cost minimizing implementation strategies such as the possibility of the use of existing resources, minimal implementation and labor costs, and cost-effective service agreements for technology use [26,33]. Identified barriers include budgetary limitations, high cost of implementation, high cost of maintaining technology equipment, and inadequate data for cost benefit analysis [26,27,31].

ii. *Outer setting*

The outer setting relates to features of the external environment or context of an organization that can influence successful implementation [17]. Three outer setting CFIR constructs identified in this review include external policy and incentives; patient needs and resources; and cosmopolitanism.

External policy and incentives related facilitating factor identified in this review includes cooperative regional licensure agreements [22]. Barriers identified include lack of, inconsistent or limited reimbursements for the intervention, and fiscal constraints from Medicare policy [21,22,26,31]. Other barriers include limited external funding, and a lack of strong organizational structures for administrative oversight by Federal agencies providing funds for innovative technology interventions [22,31]. Patients' needs and resources related facilitators in the review includes population-specific knowledge and skills by implementing staff such as nurses that directly interface with patients that are using

the innovation; quick response to patients' needs within 24 hours; end-user training; and reduced cost of access (e.g. reduced transportation costs for patients) [26,28,30,33]. Identified barriers include concerns or skepticism on how the collected data will be utilized, concerns on being able to utilize the intervention appropriately, concerns that technology complexity may limit patients' understanding and ability to use, and lack of patient motivation [27,28,32]. Cosmopolitanism refers to how much an organization is networked with other external organizations in order to be able to potentially access or harness resources for growth [35]. Identified facilitators based of this domain include adequate telecommunications and technology infrastructure irrespective of urban or rural location [30,31]. Barriers identified include poor telecommunications infrastructure in rural areas with less economic advantage than urban counterparts, low revenue, and low recovery of implementation costs due to low number of users in rural areas [31].

iii. *Inner setting*

The inner setting refers to "the organizational context in which the intervention exists [36]. Based on this review, the inner setting elicited the highest number of CFIR constructs and implementation facilitators and barriers. Five inner setting constructs were identified, and these include structural characteristics, networks and communications, culture, implementation climate and implementation readiness.

Structural characteristics refers to the age, size, maturity level, and social architecture of the organization [35]. Identified structural facilitators in this review include the existence of other programs that are supportive of the intervention[32]. Identified barriers include other programs that are not in alignment with the intervention; existing incompatible payment structures (e.g. a lack of internal reimbursement for the intervention), limited staff capacity, a lack of fit between the intervention, organizational protocols and approaches, and available staff time to use or implement the intervention [21,32]. Networks and communications facilitating factors identified across the selected studies include reducing information overload due to the intervention and creating data management plans and platforms that improve communication between intervention providers, physician practices or HCOs, and patients [23]. Identified barriers include a lack of awareness of the need to coordinate communication and integration between the intervention provider and physician practice [25]. Cultural facilitators include an existing culture of research and innovation within the implementing organization, and an alignment between the project or intervention and the mission, and scope of practice of the implementing site [22,32]. Cultural barriers include intervention policies and procedures that are not in alignment with already established policies and procedures of implementing site, and a lack of culture of involving implementing staff in contributing to implementation strategies which can result in lack of staff ownership [31].

The implementation climate is also an inner setting construct that refers to the organization's capacity for a change, the shared receptivity of participants to an intervention, and the extent to which an intervention's use is supported, expected, or rewarded within an organization [18]. Implementation climate related facilitating factors identified in this review include staff ownership of intervention, optimal staff training, and the availability of highly skilled staff [31]. Others include the availability of site champion(s) for the intervention, compatibility of the intervention with the organization's mission and structure, and the presence of other programs that are supportive of the intervention [22,32]. Staff or people related implementation climate barriers include lack of training for implementing staff, low computer literacy of staff, and a lack of clinician end user input or involvement in decision making about the technology [26,31]. Technical challenges include a lack of related technical and software support, licensure issues and the need for technology related infrastructural changes for the intervention to be implemented [22,32].

Readiness for implementation refers to the level of preparedness of an organization to implement an intervention. Facilitators identified in this domain include, the availability of necessary resources for the implementation (such as office space, IT staff, equipment, and a trained and ready work force), the ability to integrate the intervention easily into existing workflows, a basic understanding of the intervention by key stakeholders, the presence of other organizational programs that support the implementation, and a lack of interoperability challenges with technology[32]. Barriers include inadequate technology infrastructure such as poor connectivity which restricts the use of the intervention, a lack of integration of the intervention into existing workflows or a lack of clearly defined approaches for clinicians to integrate the intervention into existing workflows [21,31,32].

iv. *Process*

This involves a course of action that needs to be carried out to achieve set organizational goals for the intervention [34]. Identified CFIR constructs in this domain include planning, engaging, executing, reflecting and evaluating.

Planning related facilitators identified include identifying the right patient population for the intervention, defining the role of implementing staff, and developing population specific guidelines and protocols [23,25,33]. Engagement refers to the involvement of key stakeholders in the design and implementation process. Engagement related facilitators include early stakeholder engagement, considering and including clinician perspectives in implementing the intervention, obtaining intervention buy-in at all leadership levels, developing an informed patient-clinician interaction for the intervention, and leadership and upper management buy-in on technology maintenance costs [22,23,26,28–30,32]. Executing refers to factors associated with the real time implementation of the intervention. Execution related facilitators identified in the review include the ease of integration into existing workflows, and data management infrastructure that reduce the burden of information overload [23,29,30,32]. Identified barriers include technological and technical barriers to use, the need to train staff on how to use the intervention, low referral by physicians due to increased workload or dissatisfaction with the implementing platform, a lack of consideration for physicians' off hours, and workflow integration challenges [22,24–27,32]. Reflecting and evaluating refer to actions related to taking stock of implementation processes, successes, and failures to enable improvements. Facilitators identified in the review include engaging in reflection and evaluation that can inform adjustments such as conducting surveys [29,30].

v. *Characteristics of Individuals*

This refers to the characteristics of the individuals that are involved in the implementation and use of the intervention, or the personal or individual related factors of the intervention's implementation that can have an impact on how well the intervention is delivered or received by end-users. Identified CFIR constructs in this domain based on this review include knowledge and beliefs about the intervention, self-efficacy, individual stage of change, and individual identification with the implementing organization.

Knowledge and beliefs about the intervention facilitator identified in this review include the individuals' positive attitude and value toward the intervention, while the barrier for this construct identified in this review is negative attitudes and impressions about the intervention, and low value toward the intervention particularly by implementing site staff or patient end-user populations [29]. Self-efficacy facilitator identified in the review include a good understanding of the intervention by staff of implementing sites, while identified barriers include lack of understanding of the intervention by staff, extra need for staff or referring provider training on technology use, or appropriate referrals [27,29,32]. Individual stage of change related facilitator identified include health providers' acceptance of the intervention and the process, while barriers include staff's resistance to the technology intervention [22,23]. Individual identification with the organization facilitator identified includes staff ownership of the intervention [25].

Discussion

A good number of implementation failures have been attributed to the failure of organizational leaders to establish the level of ORC prior to the onset of implementation [37,38]. Change amenable organizations are not immune to implementation failures, this is because some organizations that are generally change amenable may not be ready for a particular change at a particular point in time. Hence, change management leaders and researchers emphasize the need to establish organizational readiness for a specific change and have recommended the use of frameworks such as the CFIR to assess and prepare for it [15,39]. Studies have also observed that a high level of organizational readiness results in organizational members investing more effort in the change process, showing lesser resistance, and displaying a higher level of persistence in overcoming setbacks or impediments [40].

An evaluation of the external and internal facilitators and barriers of ORC for Telehealth and mHealth interventions as done in this study, proffers an avenue for HCOs seeking to implement new Telehealth or mHealth interventions for this populations to understand factors that are important in assessing ORC and achieving a successful implementation. Using the CFIR framework in this review to evaluate these factors ensured a systematic identification and evaluation of key barriers and facilitators to implementation across the studies included in the review. It also helped to critically assess and interpret the facilitators and barriers identified. An evaluation of the CFIR domain and construct, with the practical examples based on the implementation studies that were identified in the review can be useful for HCOs that seek to implement Telehealth or mHealth for these populations within the US. The availability of such actionable information before or during implementation can assist with assessing the level of organizational readiness of an implementing organization as well as having access to the necessary information that planners and implementers can use to guide each step of the implementing process and increase the likelihood of a successful implementation [41]. Based on this, specific recommendations on facilitators and barriers to ORC in the implementation of Telehealth and mHealth interventions for US adults and older adults based on the CFIR framework are discussed.

Recommendations for implementers based on facilitators and barriers of ORC identified in each CFIR domain

Intervention characteristics such as design quality, cost, and complexity have an influence on successful implementation, and hence are key in assessing ORC. Sufficient consideration of the end-user, the socio-technical infrastructure, and budgets or other funding commitments of the implementing site in designing and allocating costs for the intervention will help implementing organizations to assess their level of ORC as regards a particular intervention. A lack of consideration of these factors prior to the onset of designing the intervention will lead to the design of an intervention that is incongruent with the implementing site's infrastructure, end-user population, and budget. Hence, this is a key measure of ORC, and can increase the chances of a more successful intervention.

An understanding of how outer setting constructs such as external policies and incentives (e.g. regional licensure agreements), and external payment or reimbursement policies affect Telehealth or mHealth implementation is key in evaluating ORC. This can assist organizations to understand how well their external environments support the intervention. It can also enable them plan ahead to address and mitigate barriers to successful implementation due to the lack of readiness of some of these factors prior to implementation. Furthermore, an evaluation of the specific needs and peculiarities of the recipient or patient population, such as the specific needs of the adults or older adult population that are the recipients of the intervention can inform increased receptivity toward the intervention and this facilitates successful implementation. Failure to assess the needs of the recipient population or the peculiarities of this population prior to implementation implies a low level of ORC and can lead to implementation challenges. Furthermore, a lack of adequate technology or telecommunications infrastructure due to rural location of the implementation site can also result in poor implementation, and also implies a low level of ORC. Assessing ORC based on these outer setting factors is critical to successful implementation.

Given that the inner settings represent the environment within which the intervention is implemented, inner setting factors are important measures of ORC and are particularly important for implementation success. Within the inner setting domain, the level of alignment of internal payment structures and arrangements, the level of staff or work-force readiness for the intervention, the level of skilled readiness to engage recipient populations such as older adult populations, the level of alignment of the intervention with existing workflows and cultural factors that facilitate successful implementation such as staff involvement in planning or designing the intervention are critical in assessing ORC. A good consideration of these factors in the planning phase shows an appreciable level of ORC for the Telehealth or mHealth intervention and can facilitate successful adoption and implementation.

An evaluation of process related factors is crucial to implementation success and in evaluating ORC in the implementation of Telehealth and mHealth interventions for adults and older adult populations that might need some specific accommodations to be able to successfully use the intervention. Developing population specific guidelines and protocols that take into consideration the needs of these populations, as well as engages them early on in the design process are important components of achieving ORC. A proper assessment of workforce needs, and challenges related to the intervention is also important in assessing ORC.

Characteristics of implementing individuals is key to implementation and in assessing ORC. Attention to this domain by implementers has the potential to result in positive attitudes and intervention buy-in. A good level of acceptance, ownership, confidence and self-efficacy by implementing staff toward the intervention is indicative of an appreciable level of ORC. Hence, assessing the characteristics of involved individuals based on the highlighted factors is key in assessing ORC and achieving a successful implementation.

Given the broad scope of elicited facilitators and barriers of ORC in this review, study limitations include challenges with aligning implementation strategies and terminologies across studies due to differences in the terminologies used across the included studies. Also, the broad and flexible nature of the CFIR which enables its use in different scenarios and settings resulted in some overlap between identified domains and constructs, and clear boundaries a times could not be well defined between the domains and constructs. Furthermore, although the inner setting elicited the largest sets of facilitators and barriers across the included studies, it was challenging to quantifiably assess its level of importance, relative to other domains in assessing ORC. Future studies that are focused on examining the prevalence and importance of specific implementation facilitators and barriers of ORC in the implementation of Telehealth and mHealth interventions across the literature can be helpful.

Conclusion

An understanding of the organizational barriers and facilitators that are important in the implementation of Telehealth and mHealth interventions for chronic disease management among US adults and older adults is key in assessing ORC for these interventions in this population. Contextual factors such as the internal and external facilitators of implementation are key in assessing ORC and can be organized, evaluated, and analyzed using the CFIR Framework. An appropriate level of consideration of the identified factors in this review can guide implementers on understanding and gauging the readiness level of their organization prior to implementation, which has the potential to result in a more successful implementation experience with sustained adoption.

References

1. Agnihothri S, Cui L, Delasay M, Rajan B. The value of mHealth for managing chronic conditions. Health Care Manag Sci. 2020 Jun 1;23(2):185–202.
2. Bhavnani SP, Narula J, Sengupta PP. Mobile technology and the digitization of healthcare. Vol. 37, European Heart Journal. Oxford University Press; 2016. p. 1428–38.
3. Salmond SW, Echevarria M. Healthcare transformation and changing roles for nursing. Orthop Nurs. 2017;36(1):12–25.
4. Raghupathi W, Raghupathi V. An empirical study of chronic diseases in the united states: A visual analytics approach. Int J Environ Res Public Health. 2018 Mar;15(3).
5. Buttorff C, Ruder T, Bauman M. Multiple chronic conditions in the United States. 2008.
6. Colby SL, Ortman JM. Population estimates and projections current population reports. 2015.
7. Matthew-Maich N, Harris L, Ploeg J, Markle-Reid M, Valaitis R, Ibrahim S, et al. Designing, implementing, and evaluating mobile health technologies for managing chronic conditions in older adults: a scoping review. JMIR mHealth uHealth. 2016 Jun 9;4(2):e29. 5
8. CDC. Adult caregivers in the United States: characteristics and differences in well-being, by caregiver age and caregiving status. 2013.
9. Williams V, Price J, Hardinge M, Tarassenko L, Farmer A. Using a mobile health application to support self-management in COPD: a qualitative study. Br J Gen Pract. 2014 Jul 1;64(624):e392–400.
10. CMS. Medicare telemedicine health care provider fact sheet. CMS TeleHealth Fact Sheet. 2020
11. DHHS. Telehealth: Delivering care safely during COVID-19 | HHS.gov. 2020
12. Jacob C, Sanchez-Vazquez A, Ivory C. Social, organizational, and technological factors impacting clinicians' adoption of mobile health tools: systematic literature review. Vol. 8, JMIR mHealth and uHealth. JMIR Publications; 2020
13. Jennett P, Yeo M, Pauls M, Graham J. Organizational readiness for telemedicine: implications for success and failure. J Telemed Telecare. 2003;9 Suppl 2.
14. Shea CM, Jacobs SR, Esserman DA, Bruce K, Weiner BJ. Organizational readiness for implementing change: A psychometric assessment of a new measure. Implement Sci. 2014 Jan 10;9(1):7
15. Weiner BJ. A theory of organizational readiness for change. Implement Sci. 2009;4(1):67.
16. Armenakis AA, Harris SG, Mossholder KW. Creating readiness for organizational change. Hum Relations. 1993;46(6):681–703.
17. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci. 2009;4(1).
18. Hill JN, Locatelli SM, Bokhour BG, Fix GM, Solomon J, Mueller N, et al. Evaluating broad-scale system change using the Consolidated Framework for Implementation Research: challenges and strategies to overcome them. BMC Res Notes. 2018 Aug 4;11(1):560.
19. Grigsby J, Kaehny MM, Sandberg EJ, Schlenker RE, Shaughnessy PW. Effects and effectiveness of telemedicine. Vol. 17, Health Care Financing Review. Centers for Medicare and Medicaid Services; 1995.
20. Bujnowska-Fedak M, Grata-Borkowska U. Use of telemedicine-based care for the aging and elderly: promises and pitfalls. Smart Homecare Technol TeleHealth. 2015 May 7;3:91.
21. Thies K, Anderson D, Cramer B. Lack of adoption of a mobile App to support patient self-management of diabetes and hypertension in a federally qualified health center: Interview analysis of staff and patients in a failed randomized trial. JMIR Hum Factors. 2017 Oct 3;4(4):e24
22. Siciliano M, Redington L, Lindeman D, Housen P, Enguidanos S. Lessons from the trenches: adopting medication technology within agencies serving older adults. Ageing Int. 2014 Oct 3;39(3):259–73.

23. Pecina JL, Vickers KS, Finnie DM, Hathaway JC, Takahashi PY, Hanson GJ. Health care providers style may impact acceptance of telemonitoring. Home Health Care Manag Pract [Internet]. 2012 Dec 3 ;24(6):276–82.
24. Sultan M, Kuluski K, McIsaac WJ, Cafazzo JA, Seto E. Turning challenges into design principles: telemonitoring systems for patients with multiple chronic conditions. Health Informatics J [Internet]. 2019 Dec 1;25(4):1188–200.
25. Vest BM, Hall VM, Kahn LS, Heider AR, Maloney N, Singh R. Nurse perspectives on the implementation of routine telemonitoring for high-risk diabetes patients in a primary care setting. Prim Heal Care Res Dev. 2017 Jan 1;18(1):3–13.
26. Radhakrishnan K, Xie B, Jacelon CS. Unsustainable home telehealth: a Texas qualitative study. Gerontologist. 2016 Oct 1;56(5):830–40.
27. Hopp F, Whitten P, Subramanian U, Woodbridge P, Mackert M, Lowery J. Perspectives from the Veterans Health Administration about opportunities and barriers in telemedicine. J Telemed Telecare. 2006 Dec 1;12(8):404–9.
28. Courtney KL, Lingler JH, Mecca LP, Garlock LA, Schulz R, Dick AW, et al. Older adults' and case managers' initial impressions of community-based telehealth kiosks. Res Gerontol Nurs. 2010;3(4):235–9.
29. Stevenson L, Ball S, Haverhals LM, Aron DC, Lowery J. Evaluation of a national telemedicine initiative in the Veterans Health Administration: factors associated with successful implementation. J Telemed Telecare. 2018 Apr 1;24(3):168–78.
30. Rudin RS, Fanta CH, Qureshi N, Duffy E, Edelen MO, Dalal AK, et al. A clinically integrated mHealth app and practice model for collecting patient-reported outcomes between visits for asthma patients: implementation and feasibility. Appl Clin Inform. 2019;10(5):783–93.
31. West VL, Milio N. Organizational and environmental factors affecting the utilization of telemedicinein rural home healthcare. Home Health Care Serv Q. 2004 Dec 9 [cited 2020 Aug 24];23(4):49–67.
32. Shaw RJ, Kaufman MA, Bosworth HB, Weiner BJ, Zullig LL, Lee SYD, et al. Organizational factors associated with readiness to implement and translate a primary care based telemedicine behavioral program to improve blood pressure control: The HTN-IMPROVE study. Implement Sci. 2013 Sep 8;8(1).
33. Vinson M, McCallum R, Thornlow D, Champagne M. Design, Implementation, and Evaluation of Population-Specific Telehealth Nursing Services. [Internet]. Nurse Economics. 2011 [cited 2020 Aug 24].
34. Rojas S, Ashok M, Morss D. Table A, Consolidated Framework for Implementation Research (CFIR) domains and constructs. AHRQ. 2014;
35. CFIR. Constructs – The Consolidated Framework for Implementation Research.
36. Lash SJ, Timko C, Curran GM, McKay JR, Burden JL. Implementation of evidence-based substance use disorder continuing care interventions. Psychol Addict Behav. 2011 Jun;25(2):238–51.
37. Kotter J. Leading Change: Why transformation efforts fail. Havard Bussines Review. 1995
38. Gagnon M-P, Attieh R, Ghandour EK, Légaré F, Ouimet M, Estabrooks CA, et al. A systematic review of instruments to assess organizational readiness for knowledge translation in health care. Jeyaseelan K, editor. PLoS One. 2014 Dec 4;9(12):e114338.
39. Lehman WEK, Greener JM, Rowan-Szal GA, Flynn PM. Organizational readiness for change in correctional and community substance abuse programs. J Offender Rehabil. 2012 Feb;51(1–2):96–114.
40. Fuller B, Rieckmann T, Nunes E, Miller M, Arfken C. Organizational readiness for change and opinions toward treatment innovations. J Subst Abus Treat. 2007;33(2).
41. Keith RE, Crosson JC, O'Malley AS, Cromp DA, Taylor EF. Using the consolidated framework for implementation research (CFIR) to produce actionable findings: a rapid-cycle evaluation approach to improving implementation. Implement Sci. 2017 Feb 10;12(1):15.

Acknowledgment

This is to acknowledge Kyra Mendez, BSN, RN, PhD Candidate, Johns Hopkins University School of Nursing who contributed to the full text screening of the articles included in this review.

[1] MBChB - Bachelor of medicine and surgery
[2] MSGH – Master of Science in Global Health
[3] OHSU-PSU School of public health Portland, Oregon
[4] Oregon Health and Science University – Portland State University School of Public Health Portland Oregon

More Generalizable Models For Sepsis Detection Under Covariate Shift

Jifan Gao, MS[1], Philip L. Mar, MD[2], Guanhua Chen, PhD[1]
[1]University of Wisconsin, School of Medicine and Public Health
[2]Saint Louis University, School of Medicine

Abstract

Sepsis is a major cause of mortality in the intensive care units (ICUs). Early intervention of sepsis can improve clinical outcomes for sepsis patients[1,2,3]. Machine learning models have been developed for clinical recognition of sepsis[4,5,6]. A common assumption of supervised machine learning models is that the covariates in the testing data follow the same distributions as those in the training data. When this assumption is violated (e.g., there is covariate shift), models that performed well for training data could perform badly for testing data. Covariate shift happens when the relationships between covariates and the outcome stay the same, but the marginal distributions of the covariates differ among training and testing data. Covariate shift could make clinical risk prediction model nongeneralizable. In this study, we applied covariate shift corrections onto common machine learning models and have observed that these corrections can help the models be more generalizable under the occurrence of covariate shift when detecting the onset of sepsis.

Introduction

Sepsis is a life-threatening complication of infection, which can cause a cascade of changes that damage multiple organs and sometimes even leads to death[1]. About 6 million people die from complications of sepsis each year[7]. Early intervention with fluid resuscitation and antibiotics greatly improves the chance of survival for sepsis patients. However, the detection of sepsis is challenging because sepsis is a very heterogeneous syndrome[4]. Rule-based scoring systems have been widely used in hospitals for identifying sepsis[8,9,10]. In recent years, with the prevalence of electronic health records (EHRs), many attempts have been made to build machine learning models for sepsis detection, and some of them outperform rule-based models. Lyra et al.[11] used 40 features for early prediction of sepsis using random forests. Mitra et al.[4] and Mao et al.[5] studied several machine learning models using only six vital signs, including heart rate, respiratory rate, SpO2, temperature, systolic blood pressure, and diastolic blood pressure, and found machine learning models outperform rule-based scoring systems in sepsis detection task for ICU patients.

Generalizability refers to the ability of machine learning models to make correct predictions on data collected from a different source that is not included in the training data[12]. A generalizable model should perform well for both training data and testing data; however, when there is data shift, it is difficult to ensure generalizability for machine learning models. Data shift is defined as when the population characteristics on which the model was developed is different from the population characteristics on which the model is applied[13,14]. There are three types of data shift: covariate shift, prior probability shift, and concept shift. Covariate shift is associated with the change of distributions of the predictors; prior probability shift is related to the change of the outcomes; concept shift refers to the change of the underlying relationship between the predictors and the outcomes. Studies have shown that data shift can hurt the generalizability of machine learning models. For example, Hwang et al.[15] trained a model to detect abnormal chest radiographs and the model's specificity at a fixed threshold varied from 0.566 to 1.000 when validated using external data from different sites. Nestor et al. observed an AUROC drop of 0.29 for mortality prediction when models were trained on historical data and tested on future data. In this study, we focused on mitigating the model performance deterioration caused by covariate shift.

In common machine learning models, there is presumption that the joint distribution of covariates/predictors and the outcome variable is the same in training and testing sets. However, this assumption is violated when there is covariate shift, which refers to the situations where the underlying relationship between covariates and outcomes stays the same but training and testing sets follow different covariates distributions[13]. In clinical care scenarios, covariate shift is likely to occur due to temporal or geographical differences in populations. Some research has been done to demonstrate the need of covariate shift correction on clinical risk prediction tasks such as mortality and readmission[16,17]. When the covariate information of the testing data is also available, a general framework called importance sampling (IS)

weights has been proposed to remedy the impact of covariate shift. In particular, we would assign different weights to training samples based on their similarity to the test samples[18]. Higher weights are assigned to those samples which are common in the test set, and lower weights are assigned to those which rarely occur in the test set. The optimal sample weight for a subject in the training set with its covariates vector equals to $\mathbf{x}$ can be written as a density ratio as:

$$r(\mathbf{x}) = \frac{p_{test}(\mathbf{x})}{p_{train}(\mathbf{x})} \tag{1}$$

where $p_{test}(\mathbf{x})$ and $p_{train}(\mathbf{x})$ are the densities of covariates $\mathbf{x}$ associated with the testing and training set. To see why density ratio is useful, we show that when the covariate shift takes place, the expected (squared) prediction error of a model on its testing set can be written as:

$$\begin{aligned} E_{test}(y - f(\mathbf{x}))^2 &= \int\int (y - f(\mathbf{x}))^2 p_{test}(\mathbf{x}) p_{test}(y|\mathbf{x}) d\mathbf{x} \\ &= \int\int (y - f(\mathbf{x}))^2 r(\mathbf{x}) p_{train}(\mathbf{x}) p_{test}(y|\mathbf{x}) d\mathbf{x} \\ &= \int\int (y - f(\mathbf{x}))^2 r(\mathbf{x}) p_{train}(\mathbf{x}) p_{train}(y|\mathbf{x}) d\mathbf{x} \\ &= E_{train}[r(\mathbf{x})(y - f(\mathbf{x}))^2] \end{aligned} \tag{2}$$

Note that second equality holds since the conditional probability $p_{train}(y|\mathbf{x})$ and $p_{test}(y|\mathbf{x})$ are assumed to be equal under covariate shift. The quantity $E_{train}[r(\mathbf{x})(y - f(\mathbf{x}))^2]$ can be empirically estimated given the training data and the density ratios. The above equation indicates that to minimize the prediction error given a testing set, we should use density ratios as sample weights during the training stage.

The IS framework is very general such that any methods that would produce non-negative weights estimation are applicable for correcting covariate shift. However, very few studies quantify how different combinations of weight estimating methods and risk prediction models could impact model performance. To fill this gap, we compared three commonly used machine learning models combined with different density ratio estimations methods for mitigating covariate shift problems in sepsis detection. We used routinely measured covariates in the EHR, including heart rate, respiratory rate, SpO2, temperature, systolic and diastolic blood pressures to build machine learning models to detect sepsis's onset using the training data. We reported the model performance on the testing data under covariate shift. In particular, the final models output a probability that sepsis is taking place at the current hour. During the training process, we applied two categories of approaches for density ratio estimation: direct approaches and indirect approaches. For direct approaches, we used two methods called Kernel Mean Matching (KMM)[19] and Relative unconstrained Least-Squares Importance Fitting (RuLSIF)[20] to estimate the density ratios for training samples. For indirect approaches, we built probabilistic classifiers to separate training and test samples[21] and then derived the density ratios from the classifiers' output using formulas described later. We built sepsis prediction models using three commonly used machine learning models: logistic regression, random forests, and neural networks. Then, we applied the density ratios during the training process and examined the effect of the density ratios on these machine learning models.

Methods

In this study, we used a public available EHR data called eICU Collaborative Research Database (eICU), which is a relational database that contains 200,859 admissions and 139,367 patients between 2014 and 2015 from 335 units at 208 hospitals in the US[6]. The eICU dataset stores clinical data such as diagnosis, vital signs, lab tests, and drug admissions. The dataset is de-identified to comply with the Health Insurance Portability and Accountability Act (HIPAA). In this study, we only made use of six vital signs, which are frequently available in EHRs. We applied the inclusion criteria to keep patients who are at least 18 years or older, have at least one measurement for each of the six selected vital signs, and have at least 3 hours of data before the onset. With these criteria applied, there are 1431 patients with sepsis encounters in the cohort in the study. To make sure the case numbers and controls are balanced, we randomly picked 1600 patients without sepsis encounters instead of using all patients without sepsis encounters. To maintain independence among samples, we randomly picked one sepsis onset for each sepsis patient

and one random time point for each non-sepsis patient. We used the latest sepsis-3 criteria[3] to define the onset of sepsis. According to the latest sepsis-3 criteria, sepsis is defined as life-threatening organ dysfunction when organs are injured by a dysregulated response to infection. Organ dysfunction is defined as an increase of the Sequential Organ Failure Assessment (SOFA) Score[10] by 2 or greater after an infection. The onset of sepsis is defined in our analysis by the first use of antibiotics or microbial sampling. Measured values of patients were divided into 1-hour segments. If a patient had multiple values in one type of measurement within an hour, we used their mean to represent the value of this hour for the measurement. When one measurement was missing at a given hour, it was filled with the patient's last measured value to the missing hour. When the patient did not have any measurement prior to the missing hour, it was filled with the next available measurement.

Although the eICU dataset contains data from multiple medical centers, there is no information about the patients' accurate admission times and locations due to the HIPAA regulations. Therefore, we could not develop an approach to obtain the covariate shift based on temporal or geographical differences. Instead, we developed approaches to mimic covariate shift scenarios. In particular, we trained classifiers to detect sepsis when randomly splitting the dataset with 5-fold cross-validation and extracted the mean of covariates importance from the classifiers. Using random forests classifiers, we found that systolic blood pressure and diastolic blood pressure have the highest predictive power among all covariates (see Table 1). We split the dataset into two parts based on cluster membership output from spectral clustering using systolic blood pressure and diastolic blood pressure. Distributions of normalized systolic and diastolic blood pressures of training and test data are displayed in Figure 1. As a result, there is a covariate shift between training and testing in terms of systolic blood pressure and diastolic blood pressure. We chose these two variables instead of other variables because if a variable is not critical for predicting the outcome, distribution changes in this variable will not impact risk prediction.

Table 1: Covariates importance by random forests

Covariate	**Importance**
Systolic blood pressure	0.217
Diastolic blood pressure	0.232
Heart rate	0.176
Respiratory rate	0.140
SpO2	0.087
Temperature	0.147

We applied two categories of approaches to compute the density ratio of training samples: direct estimation and indirect estimation. Kernel Mean Matching (KMM) and Least-Squares Importance Fitting (RuLSIF) are used as direct approaches to estimate density ratios. KMM is proposed by Huang et al.[19] and the main idea of KMM is to match the moment of $r(\mathbf{x})p_{train}(\mathbf{x})$ and $p_{test}(\mathbf{x})$ using kernel functions. The optimization problem of KMM can be solved by quadratic programming. RuLSIF is proposed by Yamada et al.[20] and its main idea is to estimate the relative density ratio by minimizing the squared loss. The optimal solution of RuLSIF can be obtained analytically.

Besides the direct estimation approaches mentioned above, we also built probabilistic classifiers to calculate the density ratio based on the classifiers' prediction. The classifiers were trained to predict the probability of a sample coming from the test set given its covariates $\mathbf{x}$, which is denoted as p(sample is from test$|\mathbf{x}$). The density ratio can be computed as:

$$r(\mathbf{x}) = \frac{p(\text{sample is from test set}|\mathbf{x})}{(1 - p(\text{sample is from test set}|\mathbf{x}))} \tag{3}$$

Then the ratio was normalized in order to avoid extremely large values. In this study, we chose two commonly used machine learning models, logistic regression with L1 and L2 penalty and random forests, to generate the density ratio indirectly.

Both parametric and non-parametric models were built to detect the onset of sepsis with/without density ratio correction using the training data. We chose random forests[22] as a representative of non-parametric models and chose the neural networks and logistic regression with L1 and L2 penalty[23] as representatives of parametric models. For

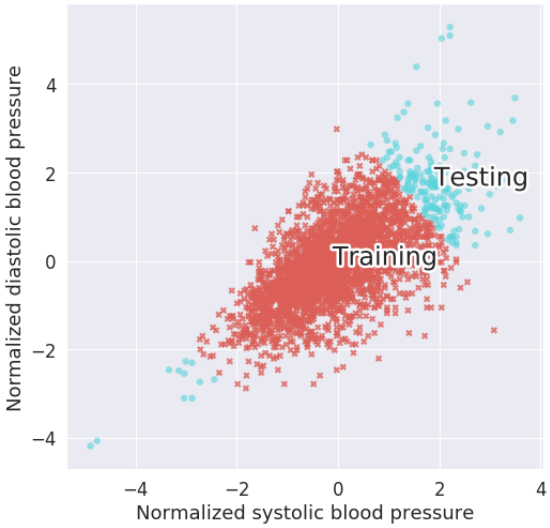


Figure 1: Training and testing data sets. Each red point represents a training sample and each blue points represents a testing sample.

parameters tuning, we performed 5-fold cross-validation to find optimal hyper-parameters with the highest AUROCs. Building these models is essentially a weighted classification with each sample weighted by density ratio estimation. We evaluated the model performance on predicting the onset of sepsis using the testing data. We reported the Area Under the ROC curve (AUROC) as the discrimination metric (a larger value is favored) and the Brier score as the calibration metric (a smaller value is favored), which are two necessary measures to assess the performance of a clinical risk prediction model[24]. As shown in the literature[25,26], discrimination accuracy is critical for evaluating risk prediction model performance, but it does not assess the accuracy of individual risk predictions. Hence, we reported both metrics to provide a more comprehensive assessment.

Table 2: Models' AUROCs under different covariate shift correction methods

	Logistic Regression	**Random Forests**	**Neural Networks**
Training 5-CV	0.801	0.824	0.816
Upper bound	0.790	0.820	0.802
Without correction	0.673 [0.669, 0.675]	0.773 [0.757, 0.791]	0.694 [0.667, 0.723]
With KMM ratio	0.684 [0.679, 0.692]	0.791 [0.758, 0.813]	0.706 [0.676, 0.736]
With RuLSIF ratio	0.707 [0.700, 0.710]	0.778 [0.754, 0.803]	0.715 [0.686, 0.748]
With RF ratio	0.676 [0.671, 0.681]	0.773 [0.749, 0.794]	0.715 [0.681, 0.750]
With LR ratio	0.680 [0.675, 0.684]	0.782 [0.757, 0.799]	0.711 [0.683, 0.745]

Results

Table 2 and Table 3 show the numerical performance of models when predicting the occurrence of sepsis. The first rows are the models' performance on the training data with 5-fold cross validation. We also put the covariates and labels from the testing data together with the training data. The second rows display the models' performance using 5-fold cross validation with all these samples. Therefore, the second rows in the two tables can be considered as the upper bounds of correction methods on a dataset with covariate shift because the covariate shift disappears with such

a dataset. The third rows show the models' performance on the testing set without any sample weights applied when models were trained on the training set. The last four rows show the models' performance on the testing set with different correction methods when models were trained on the training set. The logistic regression improved when density ratios were applied as evidenced by a higher AUROC with RuLSIF (0.707) when compared to an AUROC of 0.673 without any corrections. The improvement upon random forests was limited with respect to the indirect density ratio estimation approaches. In addition, AUROC improvement was also observed in the neural networks with the density ratio corrections, but the Brier scores were not improved with the incorporation of KMM and RuLSIF ratios, which indicates that the improvement in AUROC is not necessarily tied together with the improvement in the Brier scores.

Table 3: Models' Brier scores under different covariate shift correction methods

	Logistic Regression	**Random Forests**	**Neural Networks**
Training 5-CV	0.181	0.173	0.174
Upper bound	0.186	0.174	0.180
Without correction	0.250 $[0.248, 0.253]$	0.185 $[0.181, 0.189]$	0.210 $[0.196, 0.236]$
With KMM ratio	0.211 $[0.204, 0.220]$	0.182 $[0.173, 0.191]$	0.236 $[0.208, 0.259]$
With RuLSIF ratio	0.224 $[0.220, 0.231]$	0.184 $[0.177, 0.191]$	0.234 $[0.209, 0.257]$
With RF ratio	0.233 $[0.230, 0.238]$	0.185 $[0.180, 0.191]$	0.207 $[0.185, 0.233]$
With LR ratio	0.235 $[0.232, 0.239]$	0.182 $[0.176, 0.188]$	0.209 $[0.187, 0.233]$

Figure 2 shows the derived training samples density ratios. Sample points with darker colors are assigned with more substantial weights, whereas sample points with lighter colors are assigned with smaller weights. Since the testing data are mainly located near the edges of the plot, the results from Figure 2 indicate that samples which are similar to the test set are assigned heavier weights by both direct and indirect covariate shift corrections.

To track the change of shifted covariates' impact on the models, we obtained the Shapley values of systolic blood pressures and diastolic blood pressures. The Shapley values are computed through game theories and can explain the covariates' contribution to a model's predictions[27,28]. A covariate with a large Shapley value indicates that it significantly influences the model given other covariates. Figure 3 shows the Shapley values of systolic blood pressures and diastolic blood pressures before and after the corrections, as well as on the models which were learned using the testing set (denoted as "Testing" group). For the logistic regression models, the Shapley values are close to the model trained on the testing data, especially with the RuLSIF ratios. For the random forests and neural networks, the values are far from the "Testing" group, even though they are different from the model without any correction.

Discussion

In this work, we applied covariate shift corrections to sepsis detection task for ICU patients. We found that when covariate shift takes place, assigning different weights to training samples based on their similarity to testing samples can improve the performance of machine learning models. Compared with some previous studies[4,5], we adopted much fewer measurements as covariates/features and achieved an AUROC close to 0.80 with covariates correction. Our work is potentially impactful on clinical practice, especially when we want to transit a risk prediction model trained using one healthcare system to another healthcare system where the population may have inherently different characteristics. Our results indicate that applying covariate shift corrections are likely to make the model more generalizable. Although our paper focused on detecting sepsis onset, the strategy of correcting the covariate shift is applicable to detecting and predicting other clinical risks without much modification.

Compared with the random forests model (a complex nonlinear model), we found that logistic regression is more sensitive to the density ratio corrections and achieved a bigger improvement in discrimination and calibration. We believe that the random forests model is more complex and less likely to suffer from the model misspecification.

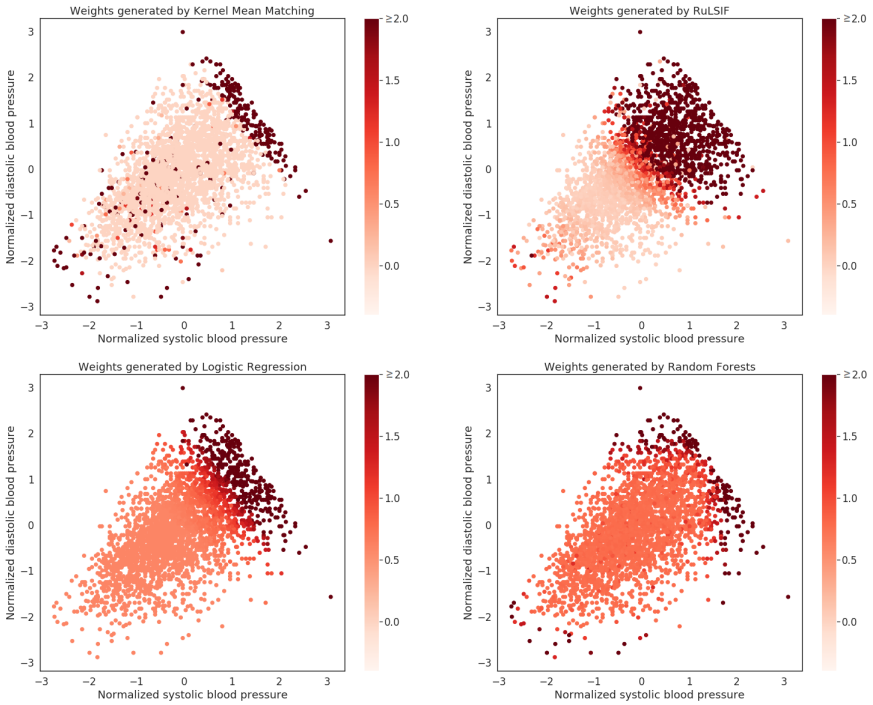


Figure 2: Visualization of training sample weights

Random forests could achieve decent performance even before the correction (AUROC larger than 0.75 and Brier scores smaller than 0.20 for random forests before corrections compared with AUROC smaller than 0.70 and Brier scores larger than 0.25 for logistic regression). In contrast, the logistic regression is accurate only when the linear functions of vital signs can well approximate the logit of being sepsis cases. Hence, the density ratio can alleviate the model misspecification problem by encouraging the model to perform well on the testing data instead of training data. Similar to the random forests, Brier scores of the neural network models did not improve much after density ratio correction. One possible explanation is that the neural networks with nonlinear activation layers are flexible nonlinear models, making them suffer less from misspecification than logistic regression. These results indicate that parametric/simple models are more likely to benefit from the covariate shift corrections than complex models.

The Brier scores of the neural networks show that density ratio correction does not necessarily improve both discrimination and calibration of the machine learning models. One explanation is that the hyper-parameters are optimized through cross-validation using AUROC (a discrimination metric); hence the corrections may have less impact on calibration metrics such as the Brier score. On the other hand, the variable importance of covariates (in particular, the shifted covariates) could also change after applying density ratio correction. Such a change of variable importance is accompanied by the improvement of the models' performance. For example, the AUROC of the logistic regression model increased from 0.673 to 0.707 with the RuLSIF ratio applied, and the corresponding Shapley values of the shifted covariates are relatively close to the values of the model trained on the testing set. In contrast, the performance of the random forests and neural networks model have minimal improvement after correction as the Shapley values of the density ratio corrected models are very different from the model trained on the testing data.

Table 4: Model performance (AUROC) with prior shift

	First Test set with 45.9% positive	Second Test set with 55.6% positive
Random Forests	0.836	0.836
Logistic Regression	0.809	0.806
Neural Networks	0.814	0.820

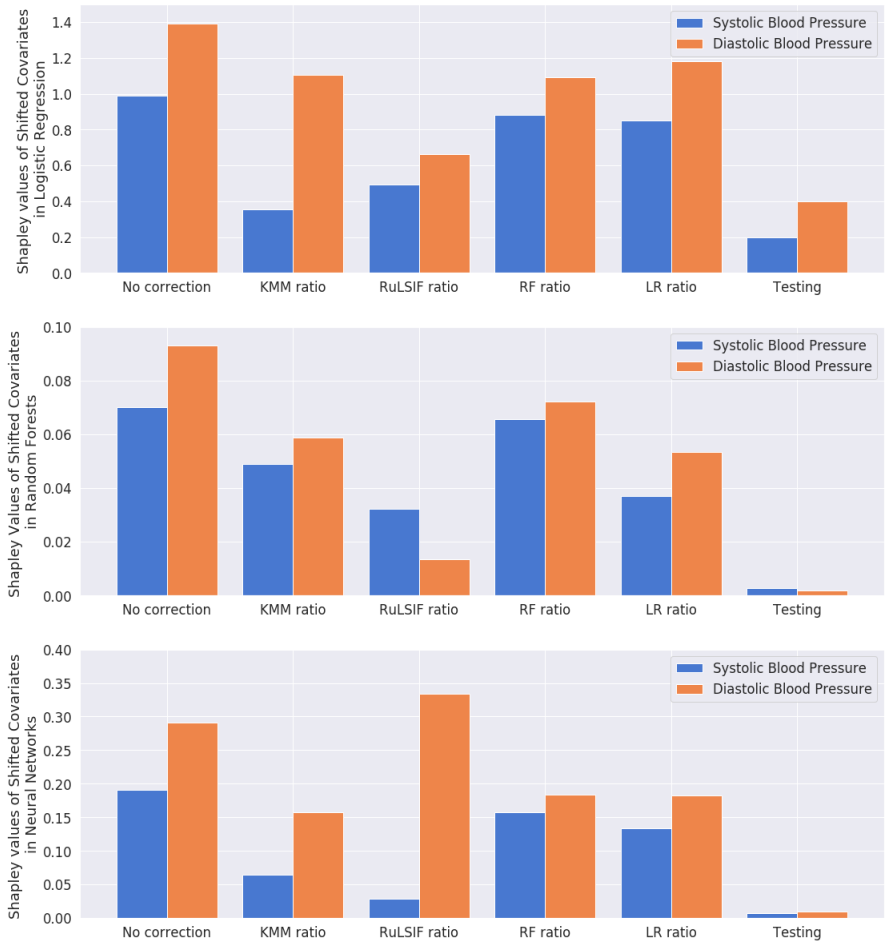


Figure 3: Change of feature importance (Shapley values) and coefficients before and after the corrections. The Shapley values are computed through game theories and can explain the covariates' contribution to a model's predictions. A covariate with a large Shapley value indicates that it significantly influences the model given other covariates

In this paper, we focused on the effect of the covariate shift. However, other types of data shift can also happen in practice. Based on Equation (2), we have shown that in theory, the density ratio correction is useful for covariate shift, while density ratio correction may not be necessary/useful for prior shift and concept shift. In particular, the prior shift can lead to case/control imbalance, which could have a mild impact on model generalizability. Resampling based approaches that are useful for handling imbalanced classification could potentially remedy the problem caused by prior shift. On the other hand, density ratio correction is not helpful for concept shift at all. We have conducted simulations to support these two conclusions as follows. To create a prior shift, we used 1958 of the original training samples, where 46.9% are positive, to train classifiers then tested the classifiers on two testing set: one is the 451 samples from the original training samples where 45.9% are sepsis cases, the other is the remaining 450 samples from the original training samples where 55.6% are sepsis cases. The AUROC performance displayed in Table 4 demonstrates that the prior shift does not substantially affect model performance for our setting. To create a concept shift, we used 1958 samples as the training set and two sets with precisely the same sample covariates as testing. Then, we replaced the outcome labels of the second testing set with labels generated from a model that is independent of the sample covariates. Such a setting guarantees that both testing sets have no covariate shift and only the second testing set has a concept shift. We tested the covariate shift correction, and the results in Table 5 show that the density ratio cannot mitigate the performance drop caused by the concept shift. How to remedy the problem caused by concept shift is an open question and interesting future direction.

Table 5: Model performance (AUROC) with concept shift

	Logistic Regression	Random Forests	Neural Networks
Training 5-CV	0.801	0.824	0.808
Upper bound	0.720	0.787	0.745
Wihtout correction	0.394	0.381	0.393
With KMM ratio	0.364	0.383	0.376
With RuLSIF ratio	0.373	0.391	0.334
With RF ratio	0.379	0.438	0.334
With LR ratio	0.385	0.383	0.386

Over the past two decades, most sepsis-related studies have used the Sepsis-2 criteria as the gold standard for sample labeling cases of sepsis[4,5,11]. The Sepsis-2 criteria uses vital signs such as heart rates and body temperature to define the onset of sepsis[29]. This definition may lead to data leakage if we use the same measurements in both outcome labeling and feature construction. Therefore, we used the latest Sepsis-3 criteria as the gold standard in this project. Besides, previous studies[4,5,11] only used AUC to measure model performance. On the one hand, AUC has desirable properties such as scale-invariance and being concordance based; on the other hand, AUC is a discrimination metric that cannot reflect calibration. For this reason, we also reported Brier score (a calibration metric) together with the AUROC to better characterize the models' performance improvement.

Conclusion

We built parametric and non-parametric models to detect the onset of sepsis for ICU patients. Direct and indirect approaches are applied to compute density ratios for covariate shift correction. We found that both parametric and non-parametric models benefit from the density ratios but the simpler parametric models such as the logistic regression are more sensitive to the covariate shift corrections.

Acknowledgement

This study was partially funded through a Fall Research Competition grant from OVCRGE at University of Wisconsin-Madison and through Patient-Centered Outcomes Research Institute (PCORI) Awards (ME-2018C2-13180). The views in this paper are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

References

1. Sakr Y, Jaschinski U, Wittebole X, Szakmany T, Lipman J, Ñamendys-Silva SA, et al. Sepsis in intensive care unit patients: worldwide data from the intensive care over nations audit. In: Open forum infectious diseases. vol. 5. Oxford University Press US; 2018. p. ofy313.
2. Rhee C, Kadri SS, Danner RL, Suffredini AF, Massaro AF, Kitch BT, et al. Diagnosing sepsis is subjective and highly variable: a survey of intensivists using case vignettes. Critical Care. 2016;20(1):89.
3. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama. 2016;315(8):801–810.
4. Mitra A, Ashraf K. Sepsis prediction and vital signs ranking in intensive care unit patients. arXiv preprint arXiv:181206686. 2018;.
5. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ open. 2018;8(1).
6. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific data. 2018;5:180178.
7. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine. 2006;34(6):1589–1596.

8. Subbe C, Slater A, Menon D, Gemmell L. Validation of physiological scoring systems in the accident and emergency department. Emergency Medicine Journal. 2006;23(11):841–845.
9. Rangel-Frausto MS, Pittet D, Costigan M, Hwang T, Davis CS, Wenzel RP. The natural history of the systemic inflammatory response syndrome (SIRS): a prospective study. Jama. 1995;273(2):117–123.
10. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al.. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Springer-Verlag; 1996.
11. Lyra S, Leonhardt S, Antink CH. Early Prediction of Sepsis Using Random Forest Classification for Imbalanced Clinical Data. In: 2019 Computing in Cardiology (CinC). IEEE; 2019. p. 1–4.
12. Dexter GP, Grannis SJ, Dixon BE, Kasthurirathne SN. Generalization of Machine Learning Approaches to Identify Notifiable Conditions from a Statewide Health Information Exchange. AMIA Summits on Translational Science Proceedings. 2020;2020:152.
13. Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset shift in machine learning. The MIT Press; 2009.
14. Moreno-Torres JG, Raeder T, Alaiz-RodríGuez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern recognition. 2012;45(1):521–530.
15. Hwang EJ, Park S, Jin KN, Im Kim J, Choi SY, Lee JH, et al. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA network open. 2019;2(3):e191095–e191095.
16. Nestor B, McDermott M, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. arXiv preprint arXiv:190800690. 2019;.
17. Curth A, Thoral P, van den Wildenberg W, Bijlstra P, de Bruin D, Elbers P, et al. Transferring clinical prediction models across hospitals and electronic health record systems. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2019. p. 605–621.
18. Sugiyama M, Krauledat M, MÃžller KR. Covariate shift adaptation by importance weighted cross validation. Journal of Machine Learning Research. 2007;8(May):985–1005.
19. Huang J, Gretton A, Borgwardt K, Schölkopf B, Smola AJ. Correcting sample selection bias by unlabeled data. In: Advances in neural information processing systems; 2007. p. 601–608.
20. Yamada M, Suzuki T, Kanamori T, Hachiya H, Sugiyama M. Relative density-ratio estimation for robust distribution comparison. Neural computation. 2013;25(5):1324–1370.
21. Bickel S, Brückner M, Scheffer T. Discriminative learning for differing training and test distributions. In: Proceedings of the 24th international conference on Machine learning; 2007. p. 81–88.
22. Breiman L. Random forests. Machine learning. 2001;45(1):5–32.
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301–320.
24. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux P, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. Jama. 2017;318(14):1377–1384.
25. Davis SE, Lasko TA, Chen G, Matheny ME. Calibration drift among regression and machine learning models for hospital mortality. In: AMIA Annual Symposium Proceedings. vol. 2017. American Medical Informatics Association; 2017. p. 625.
26. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. Journal of the American Medical Informatics Association. 2017;24(6):1052–1061.
27. Kalai E, Samet D. On weighted Shapley values. International journal of game theory. 1987;16(3):205–222.
28. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in neural information processing systems; 2017. p. 4765–4774.
29. Kaukonen KM, Bailey M, Pilcher D, Cooper DJ, Bellomo R. Systemic inflammatory response syndrome criteria in defining severe sepsis. New England Journal of Medicine. 2015;372(17):1629–1638.

Applications of Aspect-based Sentiment Analysis on Psychiatric Clinical Notes to Study Suicide in Youth

Amy George, BA, BCS[1], David Johnson, Ms[1], Giuseppe Carenini, PhD[1], Ali Eslami, PhD[2,4], Raymond Ng, PhD[1], Elodie Portales-Casamar, PhD[3,4]

[1]Dept. of Computer Science, University of British Columbia, Vancouver, BC, Canada;
[2]Department of Psychiatry, University of British Columbia, Vancouver, BC, Canada;
[3]Dept. of Pediatrics, University of British Columbia, Vancouver, BC, Canada;
[4]BC Children's Hospital, Vancouver, BC, Canada.

Abstract

Understanding and identifying the risk factors associated with suicide in youth experiencing mental health concerns is paramount to early intervention. 45% of patients are admitted annually for suicidality at BC Children's Hospital. Natural Language Processing (NLP) approaches have been applied with moderate success to psychiatric clinical notes to predict suicidality. Our objective was to explore whether machine-learning-based sentiment analysis could be informative in such a prediction task. We developed a psychiatry-relevant lexicon and identified specific categories of words, such as thought content and thought process that had significantly different polarity between suicidal and non-suicidal cases. In addition, we demonstrated that the individual words with their associated polarity can be used as features in classification models and carry informative content to differentiate between suicidal and non-suicidal cases. In conclusion, our study reveals that there is much value in applying NLP to psychiatric clinical notes and suicidal prediction.

Introduction

Although suicide accounts for fewer than 10% of deaths in youth globally, it is still the second leading cause of death in youth[1]. Suicidal thoughts and ideations are even more prevalent among youth, ranging from 20 to 30%[2,3]. Suicidal ideations and behaviours are strongly associated with co-occurring mental disorders and high-risk behaviours[4,5]. Despite a growing body of research on intervention and improved social awareness, mental health concerns in youth are still prevalent and under-treated. Recent research has shown a 60% increase in pediatric emergency visits for mental health disorders and a 329% increase in visits for intentional self-harm between 2007 and 2016[6]. Research by Doan et al.[7] has shown that over 35% of youth surveyed with universal psychosocial screening in the emergency room warranted further psychiatric follow-up. There is clearly a gap that can be addressed before they are admitted for self-harm or suicide attempts.

The Child and Adolescent Psychiatric Emergency (CAPE) unit at BC Children's Hospital, Vancouver, Canada, specializes in providing emergency intervention and stabilization for youth in psychiatric crisis. Approximately 45% of patients are admitted annually to CAPE for suicidality, amounting to over 100 admissions annually of children deemed to be at substantial and acute risk for suicide. The rate of readmission to CAPE is approximately 30% which has remained consistent over several years. Clinical notes are written at admission and discharge by psychiatrists and typically include, although without following a formal template, the patients' background, mental health, family history, current circumstances, and more, providing a wealth of information that can be analysed to help understand key factors associated with suicidality. Such understanding may be applied more broadly to help flag patients potentially at risk and offer care before reaching a critical stage.

Sentiment analysis is a branch of Natural Language Processing (NLP) used most often to identify and quantify the sentiment, feelings or opinions associated with a topic. This branch of NLP is most often applied to the analysis of online content (Twitter, reviews, forums, etc.) in an attempt to understand users' opinions of a brand or topic. There have only been limited applications of NLP to psychiatric clinical notes in part because they tend to be long and cover a variety of topics (family history, social context, clinical observations…) making it hard to tease out the relevant information from the rest. Conversely, this is exactly why such approaches are particularly important to apply to mental health where most clinical documentation is done through long narratives and does not fit well into structured fields that are typical in case report forms[8]. Research has shown that NLP and Machine Learning (ML) techniques can be successfully applied to identify suicide related crises in clinical notes[9–12]. However, sentiment analysis has rarely

been applied to clinical notes since one might not expect clinicians to express sentiment in their documentation. A study by McCoy et al.[13] applied sentiment analysis to over 17,000 discharge notes looking at the correlation between a sentiment score for each note and readmission and mortality risk. The study used the Pattern Module[14] which comes with a predefined lexicon that has an assigned polarity for each word but is not tailored to the clinical field. Similarly, a study by Waudby et al.[15] performed a survival analysis by applying the Pattern Module to free-text nursing notes, and a study by Weissman et al.[16] looked at the construct validity of six different sentiment analysis methods on patient encounter notes, five of which were lexicon based and only one (CoreNLP by Stanford) was ML-based. These studies mostly focused on overall sentiment and how it correlates with patient outcomes such as mortality or readmission. They were also limited by the use of lexicon-based sentiment analysis techniques, which lack lexicons tailored to the clinical field. As noted by Holderness et al.[17], most off-the-shelf sentiment analysis tools are not tailored to clinical notes, do not incorporate any medical ontologies, and thus cannot identify clinical sentiment well.

There is an increasing number of ML-based sentiment analysis tools that remove the need to use generic lexicons in specialized domains. One such tool is a recently developed sentiment analysis software, ABSApp[18] (part of NLP Architect by Intel® AI Lab), which is a system for weakly-supervised aspect-based sentiment extraction. Aspects are the words that are the object of sentiment words within a sentence. Sentiment words convey whether the given aspect has a positive or negative polarity. ABSApp can extract aspects and sentiment words from an unlabeled dataset sentence by sentence, producing an aspect-level sentiment report across the dataset. To reduce redundancy, the report combines aspects by their aliases so that plural or other forms of the same word are not listed multiple times.

This study aims to investigate the utility of sentiment analysis using ABSApp to analyze psychiatric clinical notes specifically as it relates to suicidal risk. Our objective was to evaluate whether the use of a tailored lexicon combined with a quantification of the aspects at the topic level can enable classification of the notes related to suicide from other psychiatric crises.

Methods

With research ethics board approval (H18-01402; June 2018), we obtained 1,559 long-form clinical notes written by psychiatrists during encounters with patients at the Child and Adolescent Psychiatric Emergency (CAPE) unit of BC Children's Hospital, Vancouver, Canada, between January 1st, 2015 and May 5th, 2018. Of the 1,559 notes, 515 were labelled as related to suicide (thoughts, ideation, or attempt; "Suicidal dataset") according to ICD 10 codes, 151 were labelled as other psychiatric crises ("Non-suicidal dataset"), and the remaining 893 were not labelled and are excluded from the analysis, only used in the initial step to create the lexicon. The 666 files we included in our analysis represent 289 unique patients.

To create our tailored lexicon, we applied ABSApp's lexicon extraction feature to our entire dataset. It extracts the aspects based on a prebuilt lexicon of sentiment words with an assigned polarity (positive or negative, neutral is not included)[18]. In brief, new aspect and sentiment terms are extracted through a bootstrap process initiated with a seed lexicon of generic sentiment terms. In order to initialize the bootstrap process, we used the opinion lexicon that comes with ABSApp which contains around 6,800 sentiment terms along with their polarity. This ML-based lexicon extraction algorithm[19] enables the expansion of sentiment and aspect coverage to find sentiment words not already in the lexicon, thereby expanding sentiment and aspect coverage, which has been a limitation of previous studies using solely lexicon-based analysis. By running our full dataset of 1,559 files through ABSapp, we obtained a report of 700 aspects, along with their aliases, such as plural forms, as well as up to 20 examples of the context in which the aspects and sentiment pairs were found. With reference to a protocol[20] developed for the same software, we systematically reviewed all aspects to remove redundancies, analysed their nuance and categorized them to ensure relevance to the given field. This also included a detailed inspection of the context of the aspects within the example sentences. We only retained aspects where at least 50% (10/20) of the examples were deemed relevant to patient characteristics or care process. If fewer than 50% of examples were relevant, the aspect was deleted. Categorizations were reviewed by the supervisor, and disagreements or uncertainty were resolved by discussion until there was consensus. Aspects were also consolidated into aliases during this process. This refinement process resulted in a final lexicon of 330 aspects, which our patient partners then reviewed and provided feedback on.

We then ran ABSApp's sentiment extraction feature on the 666 labelled files using the edited lexicon to extract aspects, their contexts and their associated polarity in each instance. From this, we extracted all unique aspects that had been found, and counted the number of positive and negative sentiments for each unique aspect to get frequencies by polarity for both the Suicidal and Non-suicidal datasets.

We calculated the negativity proportion for each aspect found in both the Suicidal and Non-suicidal datasets by dividing the negative polarity count by the sum of the positive and negative polarity counts for each aspect in each dataset. For the aspect-level analysis, we performed Fisher's exact test on the positive and negative counts produced by each dataset for each aspect, and then corrected for multiple testing with Bonferroni. We also performed Fisher's exact test and Bonferroni correction on the aspect frequencies at the category level.

To understand if any individual aspects or categories were strongly associated with suicidality, we ran two classification models on the document-level data. We chose Logistic Regression and Random Forest Classification for their interpretability. We created a matrix using the 330 aspects as features and calculated a net polarity for each feature by summing the positive (+1) and negative (-1) polarity associated with each aspect found in each of the 515 and 151 files in the Suicidal and Non-suicidal datasets, respectively. We organized our files by patient encounters first so that admission and discharge notes for the same patient would not be split across training and testing data when dividing the datasets in folds. We then used random forests and logistic regression for the classification task and performed cross validation with 3-folds. We also ran the classifier on the whole dataset of 666 files and used the 'feature importance' function of the Random Forest Classifier to estimate which features are most important based on permutations. We ran these two models with default parameters[21] and did not perform any hyperparameter tuning as this experiment was meant to understand baseline performance. Finally, we shuffled the labels of the training data and reran the classification models with the 3-folds to confirm that the performance of our results was not a product of the disproportionate amount of Suicidal data we have to Non-suicidal data. Performance was measured by calculating the mean accuracy and mean Receiver-Operating Characteristic (ROC) curve and Area Under the Curve (AUC) across the cross-validation folds. All analyses were run using Python and scikit learn libraries[21].

Results

In order to develop a lexicon tailored to our dataset, we ran ABSApp to extract all aspects associated with sentiments from our unlabeled dataset of 1,559 psychiatric clinical notes. The output saturated at just over 700 aspects, which made manual refinement possible. We categorized the aspects using seven major risk factor domains that can be found in patient records associated with readmission of psychiatric patients: appearance, mood, interpersonal relationships, substance use, occupation, thought content, and thought process[14]. Since many of the aspects were related to either medications or disorders, we added these two categories as well. 360 aspects that did not fit into one of the nine categories and were considered unrelated to patient characteristics or care process were removed. Table 1 describes the number of unique aspects retained and their distribution by category, with the examples highlighting the three most frequent aspects in each category. The full lexicon is available upon request.

Table 1. Number of aspects and the top three examples in each category in our tailored lexicon.

Aspect category	**Count**	**Examples**
Appearance	19	Eye contact, gestures, tics
Disorders	82	Disorders, history, illnesses
Interpersonal relationships	58	Mother, parents, dad
Medications	30	SSRI, fluoxetine, effects
Mood	38	Felt, moods, behaviours
Occupation	16	Schools, grades, students
Substance use	13	Substances, drugs, medications
Thought content	43	Suicidal ideations, thoughts, intent
Thought process	31	Accessibility, speech, insight

To start investigating potential differences between our Suicidal vs. Non-suicidal datasets, we calculated the overall proportions of positive and negative sentiment contained across all clinical notes in each dataset. In the Suicidal dataset,

we identified a total of 5,954 instances of positive sentiment and 22,926 instances of negative sentiment associated with one of the aspects, as compared to 1,452 and 5,752 respectively in the Non-suicidal dataset. Despite a large difference in the quantity of sentiment instances found in the two datasets, the ratio of positive to negative instances of sentiment are almost the same in each dataset – 20% positive and 80% negative.

We next separated the aspects based on their category and counted the positive and negative sentiments contained in each category. Table 2 shows the break-down of negativity proportions in both datasets across all categories. It shows that negativity is variable with the biggest difference in the Thought Process category (13.91% difference), followed by Thought Content (9.46% difference). Three categories were significantly different based on a Fisher's exact test and Bonferroni correction: Thought Content (corrected p-value < 0.001), Thought Process (corrected p-value < 0.05) and Mood (corrected p-value < 0.05).

Table 2. Break-down of positive and negative sentiment counts in the Suicidal and Non-suicidal datasets.

Aspect category	**Suicidal dataset positive:negative count**	**Suicidal dataset negativity proportion**	**Non-suicidal dataset positive:negative count**	**Non-suicidal dataset negativity proportion**
Appearance	349: 309	46.96%	54: 67	55.37%
Disorders	938: 11,207	92.28%	294: 3,215	91.62%
Interpersonal relationships	1,451: 1,995	57.89%	371: 546	59.54%
Medications	312: 237	43.17%	66: 66	50.00%
Mood *	930: 2,665	74.13%	219: 823	78.98%
Occupation	228: 418	64.71%	45: 76	62.81%
Substance use	811: 1,297	61.53%	198: 323	62.00%
Thought content *	665: 4,539	87.22%	147: 514	77.76%
Thought process *	277: 324	53.91%	56: 118	67.82%
Total	**5,954: 22,926**	**79.38%**	**1,452: 5,752**	**79.84%**

There were a total 265 aspects found in both datasets out of the 330 in the lexicon. To investigate the differences between the Suicidal and Non-suicidal datasets at the aspect level, we performed a Fisher's exact test on each aspect, and then performed the Bonferroni correction to correct for multiple testing. After correction, the aspect "intent" remained statistically significant with a corrected p-value < 0.01 (Suicidal: 297:158; Non-suicidal: 21:39 for positive:negative counts respectively). The next closest aspect was "risk factors" with a corrected p-value of 0.08.

As a preliminary step toward assessing the value of this output in classification tasks, we investigated whether aspect polarity counts for each document could be used as informative features in logistic regression and random forest classifiers when applied to our datasets. We split the data proportionately into three folds and ran 3-fold cross-validation with the models. Then, we generated three new folds by shuffling which files were added to each fold and repeated the analysis. Finally, we took the mean of all six outputs for each model. The results are listed in table 3. The 3-fold cross-validation resulted in an accuracy of 80.70% for logistic regression and 83.69% for random forest. The top three features from the random forest classifier were "suicidal ideations", "autism", and "behaviors". Interestingly, "intent" also comes high in the list of top features, ranked 6th out of 330 aspects.

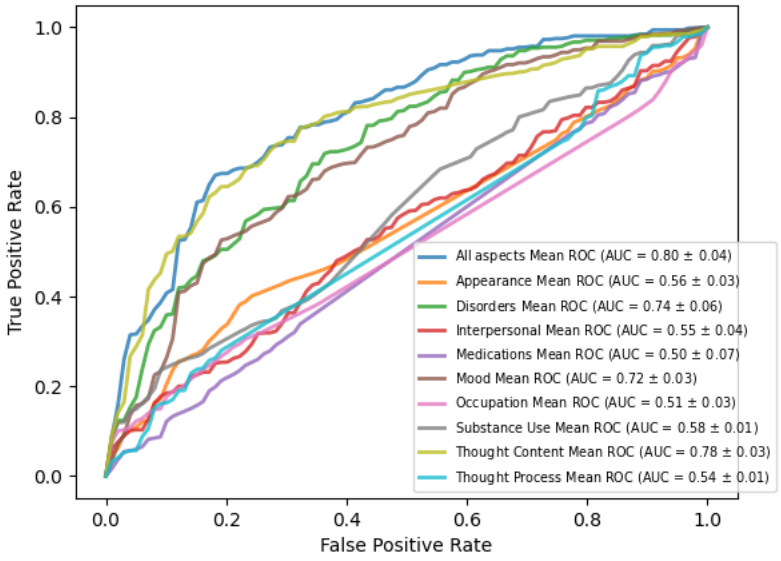
Table 3. Mean classification accuracy (%) from 3-fold cross validation using the aspects and their polarity as features in a logistic regression model and random forest classifier.

Aspect Features	Logistic Regression	Random Forest
All aspects	80.70%	83.69%
All aspects (50:50 dataset)	69.89%	80.35%
All aspects (training labels shuffled)	68.00%	75.09%
All aspects (50:50 dataset and training labels shuffled)	43.50%	53.27%
Appearance aspects only	77.06%	74.50%
Disorders aspects only	81.00%	82.32%
Interpersonal aspects only	76.02%	75.09%
Medications aspects only	76.92%	75.27%
Mood aspects only	77.87%	77.37%
Occupation aspects only	77.05%	76.20%
Substance use aspects only	77.22%	73.54%
Thought content aspects only	79.18%	76.37%
Thought process aspects only	76.47%	75.87%

Given that we have far more Suicidal documents than Non-suicidal, we were concerned that the accuracy was a product of the unbalanced dataset. When the dataset was reduced to a 50:50 split (151 files in each the Suicidal and Non-suicidal dataset), the accuracy dropped to 69.89% for logistic regression and 80.35% for random forest. Next, we shuffled the labels in the training data for the three folds, which resulted in a drop of accuracy to 68.00% and 75.09% for the full dataset and 43.50% and 53.27% for the 50:50 dataset. To see if any category of aspects plays a stronger role in the classification task, we used only the aspects from each category in turn as the models' features. Table 3 shows that the Disorders category contains the most informative features for the classification, with the highest accuracy, similar to the accuracy for the models using all aspects. All other categories resulted in similarly lower accuracies. As expected, shuffling performed the worst out of all.

Finally, to evaluate performance more thoroughly, we generated ROC curves for the models with all aspects and aspects from individual categories (Figure 1). We observe that the models including all aspects perform the best with AUC of 0.8 and 0.85 for the logistic regression and random forest classifier respectively. Three categories perform also fairly well on their own, Thought Content (AUC of 0.78 and 0.77 respectively), Disorders (AUC of 0.74 and 0.77 respectively), and Mood (AUC of 0.72 for both). The ROC curves demonstrate a low performance for all other categories (AUC ranging from 0.5 to 0.61).

a. Logistic Regression Model



b. Random Forest Classifier

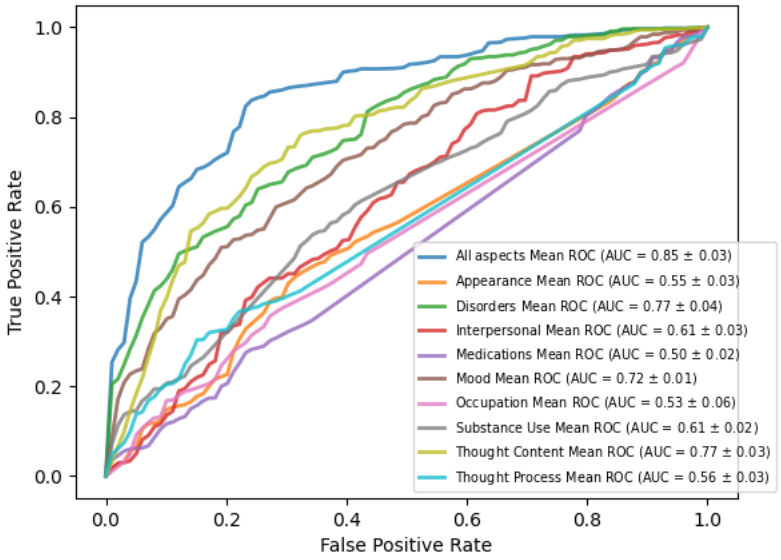


Figure 1. Mean Receiver-Operating Characteristic (ROC) curves from 3-fold cross validation using the aspects and their polarity as features in a logistic regression model (a) and random forest classifier (b).

Discussion

This exploratory study establishes a new lexicon generated by extraction from clinical notes using sentiment analysis tailored to the psychiatric and mental health fields. We identified 330 aspects relevant to the domain with attached sentiments, highlighting the fact that even though clinical notes are not expressing an individual's sentiments, the methodology still has relevance in the analysis of these notes. Our study not only looks at the polarity of sentiments within the notes like previous studies have done[13,15,16], but it also innovates by focusing on the specific aspects these sentiments are attached to. We also explored their relation to the mental health domain through categorization of the aspects to previously defined risk factors for readmission of psychiatric patients[17]. This categorization enables the user

to drill-down into the sentiment analysis results and facilitates interpretation. As a demonstration, looking at our small Suicidal vs Non-suicidal datasets, despite not identifying an overall difference in negativity (79.38% vs 79.84%), we observed variance within individual categories. The difference between the two datasets for Thought Content, Thought Process, and Mood were statistically significant.

Interestingly, looking at individual aspects, we observed that "intent" showed the most striking contrast between the two datasets, with many more negative sentiments in the Suicidal dataset. As "intent" is an aspect in the Thought Content category, this finding aligns with the fact that overall, Thought Content is attached to more negative sentiments in the Suicidal dataset and differs significantly from the Non-suicidal dataset.

As an additional way to explore the value of the sentiment analysis methodology applied to clinical notes, we investigated how informative the findings would be as features in classification models trained to differentiate between clinical notes related to suicidal patients vs other psychiatric crises. We selected two out-of-the-box classifiers to ensure that our results would not be biased by the tools selected, and we observed that the sentiment analysis data showed promising accuracy and AUC values in the classification task, similar to previous studies. For instance, a study by Le et al.[22] showed that Support Vector Machine (SVM) and other algorithms could be used to predict risk of inpatient self-harm with an accuracy of 0.69-0.77 from free-text narrative clinical notes by using symptom, sentiment and frequency dictionaries. Another study by Fernandes et al.[12] achieved precision of 82.8% for classifying suicide attempts with SVM on clinical notes using a manually curated list of features related to suicide. In our study, the tailored and categorized lexicon enabled a more in-depth investigation of which aspects are more informative than others both within and across categories. The Thought Content, Disorders, and Mood categories showed the best performance, which aligns with the fact that the top three features in the random forest classifier using all aspects belong to these three categories ("suicidal ideations" in Thought Content, "autism" in Disorders, and "behaviors" in Mood). It also aligns with the observation that the overall negativity score for the Thought Content and Mood categories were statistically different between the Suicidal and Non-Suicidal datasets. It is interesting to note that, although Disorders performed well both at the accuracy and AUC levels, it showed almost no overall negativity score difference at the category level between the Suicidal and Non-suicidal datasets. This shows that aggregating sentiment at the category-level may lose some important information and reinforces the value of the aspect-level analysis.

Our study reveals that there is much that can be gained by applying NLP techniques to psychiatric clinical notes. Although sentiment analysis may often be the tool of businesses trying to improve their brand image, this study demonstrates that it can be applied to a complex, nuanced, and sensitive task such as the analysis of psychiatric clinical notes. Not only can it be used to understand the overall polarity of a document or a dataset, but it can be used to classify complex data and extract words that may be associated with suicidality. Research has shown there is both a gap in care and space to address mental health concerns in youth[7]. Moving forward, we plan to apply our findings and techniques in building predictive models that could be used to screen youth and offer help in advance.

Conclusion

Despite the challenges that free-text clinical notes pose, we have shown that they can be an excellent resource and that aspect-based sentiment analysis and machine learning models are up to the task. We acknowledge that there are many limitations to this study. Namely, that the portion of our dataset we were able to use for classification is very small when compared to what machine-learning-based techniques are typically performed on. In addition, our data is very unbalanced, with only a few non-suicidal files, which makes it difficult to do cross-validation with more folds. To address these issues, we are in the process of labelling the remaining 893 notes left aside in this analysis to use as an independent dataset to test our model. Our data is also very sparse, as each document only contains a handful of the 330 features we analysed. As we continue this project, we may address the sparsity to try to improve accuracy of our models. Finally, as our data only spans three years to 2018, we are also working on getting the newest data from the last two years, which we hope will double our data size.

Our preliminary results are very encouraging despite working only with default parameters. Future steps will include tuning hyperparameters to improve performance, as well as a comparison of our sentiment-based method to a more classical bag-of-words or Naïve Bayes model.

In future work we hope to explore whether the differences between the aspect categories that we found may have some clinical relevance. Thought Content includes the aspect "intent", which is often how psychiatrists describe suicidality, so it is not surprising that it could be more negative in the suicide group. Other research has shown that behavioural disturbances and disordered thought process in the context of neurodevelopment disorders are

common[23,24]. This may partially explain why there is more negativity in the Thought Process category within the Non-suicidal dataset.

A possible future direction could be to apply these techniques to a broader set of outpatient notes – for example, when a patient is seeing their psychiatrist for follow up in the community, we might be able to flag that the patient is doing worse and suggest interventions before the patient reaches the point of presenting to the emergency room. Semantic analysis with lexicons fine-tuned to the mental health domain, when used in conjunction with longitudinal predictive models, may help predict an individual's risk for suicide as well as their risk for developing various mental health conditions.

Acknowledgements

We acknowledge financial support for this project from 1) the BC SUPPORT Unit Data Science and Health Informatics Methods Cluster (Award Number: DaSHI-002), which is part of British Columbia's Academic Health Science Network; and 2) the Evidence to Innovation Stimulus Award at BC Children's Hospital Research Institute. The BC SUPPORT Unit receives funding from the Canadian Institutes of Health Research and the Michael Smith Foundation for Health Research. We thank Sinead Nugent from the CAPE unit for the manual annotations and labelling of the datasets; Dr. Ali Mussavi Rizi from PHSA Data Analytics, Reporting & Evaluation for providing and ensuring continuous access to the data; as well as other students involved in other aspects of the project and who provided feedback on the manuscript, Rebecca Lin, Esther Lin, Cindy Ou Yang, and John-Jose Nuñez. Finally, we would like to give special thanks to our patient and family partners - Ariel Qi, Alison Taylor, Omar Bseiso, for their pointed questions, engagement, in-depth feedback, and suggestions on possible future directions.

References

1. Cha CB, Franz PJ, M Guzmán E, Glenn CR, Kleiman EM, Nock MK. Annual Research Review: Suicide among youth - epidemiology, (potential) etiology, and treatment. J Child Psychol Psychiatry. 2018;59:460–82.

2. Nock MK, Borges G, Bromet EJ, Cha CB, Kessler RC, Lee S. Suicide and suicidal behavior. Epidemiol Rev. 2008;30:133–54.

3. Evans E, Hawton K, Rodham K, Deeks J. The prevalence of suicidal phenomena in adolescents: a systematic review of population-based studies. Suicide Life Threat Behav. 2005;35:239–50.

4. Patton GC, Coffey C, Sawyer SM, Viner RM, Haller DM, Bose K, et al. Global patterns of mortality in young people: a systematic analysis of population health data. Lancet. 2009;374:881–92.

5. Georgiades K, Boylan K, Duncan L, Wang L, Colman I, Rhodes AE, et al. Prevalence and Correlates of Youth Suicidal Ideation and Attempts: Evidence from the 2014 Ontario Child Health Study. Can J Psychiatry. 2019;64:265–74.

6. Lo CB, Bridge JA, Shi J, Ludwig L, Stanley RM. Children's Mental Health Emergency Department Visits: 2007-2016. Pediatrics. 2020;145.

7. Doan Q, Wright B, Atwal A, Hankinson E, Virk P, Azizi H, et al. Utility of MyHEARTSMAP for Universal Psychosocial Screening in the Emergency Department. The Journal of Pediatrics. 2020;219:54-61.e1.

8. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. J Biomed Inform. 2018;88:11–9.

9. Ben-Ari A, Hammond K. Text Mining the EMR for Modeling and Predicting Suicidal Behavior among US Veterans of the 1991 Persian Gulf War. In: 2015 48th Hawaii International Conference on System Sciences. 2015. p. 3168–75.

10. Metzger M-H, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. International Journal of Methods in Psychiatric Research. 2017;26:e1522.

11. Hammond KW, Laundry RJ. Application of a Hybrid Text Mining Approach to the Study of Suicidal Behavior in a Large Population. In: 2014 47th Hawaii International Conference on System Sciences. 2014. p. 2555–61.

12. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing. Scientific Reports. 2018;8:7426.

13. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study. PLoS ONE. 2015;10:e0136341.

14. De Smedt T, Daelemans W. Pattern for python. J Mach Learn Res. 2012;13:2063–7.

15. Waudby-Smith IER, Tran N, Dubin JA, Lee J. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. PLOS ONE. 2018;13:e0198687.

16. Weissman GE, Ungar LH, Harhay MO, Courtright KR, Halpern SD. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. Journal of Biomedical Informatics. 2019;89:114–21.

17. Holderness E, Miller N, Cawkwell P, Bolton K, Meteer M, Pustejovsky J, et al. Analysis of risk factor domains in psychosis patient health records. J Biomed Semant. 2019;10:19.

18. Pereg O, Korat D, Wasserblat M, Mamou J, Dagan I. ABSApp: A Portable Weakly-Supervised Aspect-Based Sentiment Extraction System. arXiv:190905608 [cs] [Internet]. 2019 [cited 2020 Aug 27]; Available from: http://arxiv.org/abs/1909.05608

19. Qiu G, Liu B, Bu J, Chen C. Opinion Word Expansion and Target Extraction through Double Propagation. Computational Linguistics. 2011;37:9–27.

20. Johnson D, Chen Y, Dragojlovic N, Kopac N, Carenini G, Ng R. Estimating Patient Preferences Directly from Patient-Generated Text Using Aspect Based Sentiment Analysis. Paper submitted for publication.

21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON. :6.

22. Le DV, Montgomery J, Kirkby KC, Scanlan J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. Journal of Biomedical Informatics. 2018;86:49–58.

23. Sadock BJ, Sadock VA. Kaplan & Sadock's Concise Textbook of Clinical Psychiatry. Lippincott Williams & Wilkins; 2008. p. 756.

24. Jean S, Kim B, Donna R. Psychiatric Emergencies in Children and Adolescents: An Emergency Department Audit. Australas Psychiatry. 2006;14:403–7.

MAgEC: Using Non-Homogeneous Ensemble Consensus for Predicting Drivers in Unexpected Mechanical Ventilation

Stefanos Giampanis, PhD[1], Abhishaike Mahajan, BS[1],
Theodore Goldstein, PhD[1], Beau Norgeot, PhD[1]
[1]Anthem AI, Palo Alto, CA 94301

Abstract

We conduct exploratory analysis of a novel algorithm called Model Agnostic Effect Coefficients (MAgEC) for extracting clinical features of importance when assessing an individual patient's healthcare risks, alongside predicting the risk itself. Our approach uses a non-homogeneous consensus-based algorithm to assign importance to features, which differs from similar approaches, which are homogeneous (typically purely based on random forests). Using the MIMIC-III dataset, we apply our method on predicting drivers/causers of unexpected mechanical ventilation in a large cohort patient population. We validate the MAgEC method using two primary metrics: its accuracy in predicting mechanical ventilation and the similarity of the proposed feature importances to a competing algorithm (SHAP). We also more closely discuss MAgEC itself by examining the stability of our proposed feature importances under different perturbations and whether the non-homogeneity of the approach actually leads to feature importance diversity. The code to implement MAgEC is open-sourced on GitHub (https://github.com/gstef80/MAgEC).

Introduction

In this work, we propose a novel method for extracting variables of high clinical concern for a given outcome. Our method, called Model Agnostic Effect Coefficients (or MAgEC), uses a non-homogeneous[1], consensus based approach to predict the most important features contributing to the dependent variable of interest. This method is intuitive to understand, extremely fast, and has similar results to a competing algorithm. Furthermore, MAgEC is unique in its actual implementation. Typically, datapoint-specific feature importances are generated using a homogeneous set of models[2], which are typically some form of random forest. Indeed, while methods like SHAP[3] and LIME[4] are model-agnostic in terms of their application to any individual type of model, they are not cross-model compatible. On the other hand, MAgEC is model-agnostic to both individual types of models and a mixture of models, capable of combining the results from any arbitrary set of models.

MAgEC works using pseudo-counterfactuals; perturbing a single input variable at a time towards the cohort mean value, getting the model-derived expected likelihood of the outcome, and finding the difference of this perturbed likelihood to the baseline likelihood (which is derived using an unperturbed input). These differences, when done across all features for each model in a given set (logistic regression, neural network, etc), are defined as MAgEC coefficients. These coefficients can be then L2 normalized in order to be compared across models, which can then be combined using some pre-defined policy in order to achieve a final ranking. In this work, we use a naive sum between normalized feature coefficients of models to decide the ranking, but more sophisticated methods could be used as well.

To assess the applicability of MAgEC, we will specifically apply it to the problem of predicting possible drivers of unexpected mechanical ventilation (MV) within the next 3 hours, which is a prophylaxis procedure often used for ICU patients experiencing severe respiratory failure. Typically, the conditions that these patients face can range from neurological disorders causing them to be physically unable to breath, to underlying conditions such as pneumonia that have directly harmed lung capacity. Assessing the reasons behind a patient's need for MV is dependent on a range of factors, including clinical variables (e.g. systolic blood pressure), biological variables (e.g. creatinine levels), or patient variables (e.g. history of asthma). Building a holistic picture of these variables takes valuable time, and machine-learning models could potentially predict the reasons for the ventilation far faster than any team of clinicians. It is also likely that knowing the exact reasons behind an upcoming MV could prevent the MV from exacerbating or creating clinical issues in a patient[5].

We use data from the Medical Information Mart for Intensive Care (MIMIC)-III database[6], which contains a rich set of electronic health record information from thousands of patients, for this problem. Given the set of predictions and feature importances that MAgEC has made on this dataset (setup discussed in the methods section), we assess it on

two metrics of interest: its accuracy in predicting mechanical ventilation and the similarity of the proposed feature importances to a competing algorithm (SHAP). We also analyze the inner-workings of MAgEC through another two outcomes: examining the stability of our proposed feature importances under different perturbations and whether the non-homogeneity of this approach actually leads to feature importance diversity.

Related Work

In this paper, we also propose a novel method of discovering drivers for ventilation. While there are multiple current methods that are capable of performing similar tasks via model interpretability[3,4,7], these methods are not well-suited to our task. This is due to our method essentially attempting to generate counterfactuals, which are extremely useful in a clinical context ('would this person have needed MV had their arterial blood pressure been lower?'). Existing methods are not well suited to provide similar insights, because they either provide very small perturbations in the immediate vicinity of a feature by calculating partial derivatives which does not reflect the model's estimation of any clinically meaningful change, or they fit linear surrogate models which are not guaranteed to reflect the original model's prediction[4]. Our method quantifies the direction and magnitude of a clinical variable perturbation in a more realistic setting where a clinician may attempt to improve a potential outcome outside the linear regime of infinitesimally small feature changes.

On the subject of predicting the onset of MV, there are two related works[8,9]. The former paper specifically is highly similar to our own work, in that they are working to predict unexpected MV's (but not assign patient-specific feature importance). We follow a very similar patient stratification setup, in order to compare our results on predicting ventilation/non-ventilation with this work.

Finally, the usage of electronic health records to predict patient outcomes is quite common, and has been increasingly used for a large variety of clinical tasks. Common areas of application include ICU patient outcome prediction[10,11], chronic disease progression[12], and clinical text interpretation[13].

Methods

In this section we will describe the algorithmic details of what we will use to find the drivers of unexpected MV (coefficient generation, normalization, and consensus). We will then discuss the intuition behind our approach, and how it attempts to replicate medical panels. Finally, we will discuss our data sources and feature set.

Coefficient Generation

We begin with a dataset where a patient's data is represented in a tabular format, with a row representing each patient's covariates at each point in time and separately the outcome variable we are interested in predicting. The first step of our method is to train multiple supervised models on the dataset of interest. Next, each model is used to calculate the predicted outcome for each patient using the original/observed data. The predicted outcome is either a probability for classification tasks or a real number for regression tasks.

We proceed to perturb the value for one variable while keeping all other variables fixed at their original values. The perturbation is accomplished by shifting the observed value of the variable for each patient (or case) to either (a) an expert-determined clinically desired value or (b) a standard-deviation determined value calculated from the cohort used in the study (e.g. the cohort mean, or a 1SD movement towards the cohort mean). We use the latter in this study, and an example of the method can be seen in Figure 1. We then run each patient through the trained model with the single updated variable value and calculate how that single alteration affects the predicted outcome. For classification tasks, the estimated effect is calculated by comparing the difference in logits between the original prediction and the prediction made using the altered covariate value. For regression tasks, the estimated effect is calculated by comparing the difference in target values between the original prediction and the prediction made using the altered covariate value. We repeat this process for each variable and each time point (if time series models are used).

$$m_j^{(k)} = logit(f(x_1^{(k)}, ..., x_j^{(k)} = N_j, ..., x_n^{(k)})) - logit(f(x_1^{(k)}, ..., x_j^{(k)}, ..., x_n^{(k)})) \quad (1)$$

Equation 1 defines a MAgEC coefficient $m_j^{(k)}$ (which we will refer to as 'MAgECs' for the remainder of the paper) for each covariate x_j in the dataset, given a model f and a case/patient k. The value N_j represents a clinically "normal" value for x_j. Figure 1 illustrates scaling (z-normalizing) our dataset and computing MAgEC for a single case and covariate by shifting to the mean of the distribution. For binary covariates, such as 'gender' or a co-morbidity, the perturbation is done by flipping the value of the perturbed covariate (e.g. from 0 to 1 and vice versa).

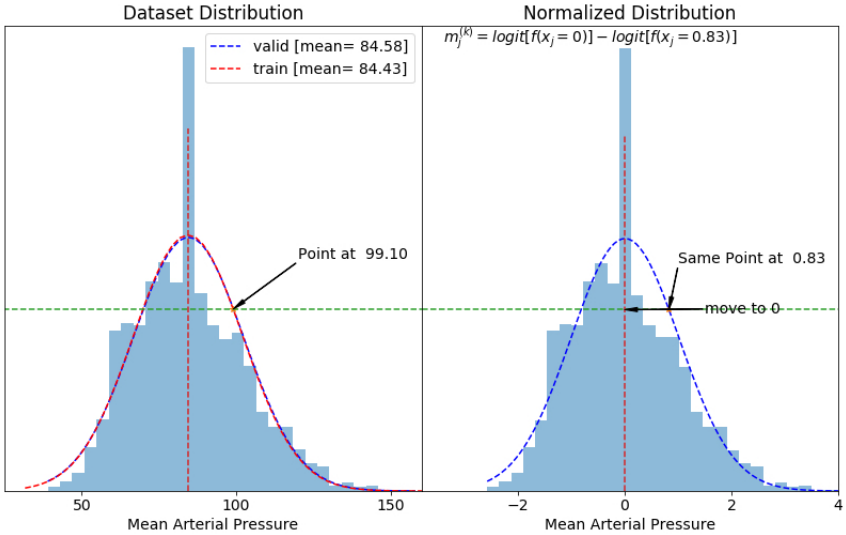


Figure 1: Mean Arterial Pressure distributions for train and validation datasets in our MIMIC-III cohort. An example of a higher than normal Mean Arterial Pressure case is shifted to the mean (at 0 for z-normalized distribution on the right) to compute MAgEC for that feature given a fitted model f. Cases with Mean Arterial Pressure bellow the measured 'normal' value are also set to 0 (for z-normalized values) when computing a MAgEC. A researcher is free to specify both the 'normal' value as well as the shifted value, e.g. shifting by 1 SD closer to the mean.

Normalization

A model's response to a perturbed covariate is a function of all of the covariates in the model, as seen in Equation 1. Even when modifying a covariate to a value that is not practically feasible (e.g. changing gender or age), one must consider that perturbation's direction and magnitude relative to a perturbation of another covariate that may be practically feasible (e.g. when setting a patient's glucose level to a normal value). In order to place all coefficients on equal footing and to account for differences in direction of effect we apply a Euclidean (L2) distance normalization, such that for any of the k cases in our dataset the Euclidean sum of all coefficients adds up to 1.

$$\tilde{m}_{jl}^{(k)} = \frac{m_{jl}^{(k)}}{\sqrt{\sum_{i=1}^{n} (m_{il}^{(k)})^2}} \tag{2}$$

In Equation 2, all coefficients are normalized for each case (k), making it possible to directly compare them across models (l denotes the l-th model in a panel of models and i/j denotes the i/j-th covariate out of n used in training the model).

Intuition

Our method seeks to algorithmically replicate a scenario similar to a tumor board, where a panel of diverse experts examines an individual patient's status, estimates their risk for a particular outcome, and decide together which intervention is most likely to reduce that patient's risk. Here, experts are supervised learning models. Expertise is quantified by test performance on the supervised task for which the models are trained. Diversity of opinion is quantified by standard classification reports. Each expert explores their own intervention strategies. This is accomplished by perturbing the value of a single variable for a given patient, while maintaining all of the other original variables for that patient. The data is perturbed by moving the patient's actual observed value to a clinically average value. The updated patient data is then passed through the trained model for inference and difference between the original risk is calculated.

Ultimately, this produces a coefficient providing the estimated direction and magnitude of a likely intervention for each patient, for each variable, for each model. Expert panel consensus on the best intervention and the impact thereof is determined by a user-defined policy that combines the metric deemed most appropriate to analytically quantify expertise, each expert's performance with respect to that metric, and variability with respect to expert opinion on the impact of each potential intervention. In our case, we simply consider the feature-wise sum amongst the models, which is then ranked to show the features most contributing to the outcome.

Dataset and Features

Data for this study is extracted from the MIMIC-III database. The MIMIC-III database contains de-identified records from patients admitted in an intensive care unit (ICU) at Beth Israel Deaconess Medical Center in Boston, Massachusetts in the period of June 2001 to October 2012. In this study, we use a cohort of 10415 non-surgical adult patients admitted at the hospital in ICU for the first time. Patients discharged from the ICU within 27 hours of admission are excluded.

We split our cohort of 10415 patients into a train and test set (using an 80/20 random split) and apply z-normalization for every feature. We train a multi-layer perceptron (MLP), a support vector machine classifier (SVM) and a logistic regression (LR) model. All three models are untuned, and represent equally weighted supervised experts.

In terms of features, we use laboratory values (19 features), vital signs (16 features), comorbidities (3 features) and demographics (2 features), all listed in Table 1. As lab values and vital signs are temporal values, we choose the latest available value from 24 hours prior to the ventilation to 3 hours prior (for vitals we also record the first available measurements). The target variable is a binary variable marking whether a patient received mechanical ventilation support within the next three hours.

After the training process, we apply the MAgEC method to the test set. MAgEC coefficients are generated, normalized and ranked for each case in the test set, without using the ensemble predictions.

Results and Discussion

As discussed earlier, we will evaluate the outputs behind MAgEC in two primary ways: how it performed in predicting MV (to assure that we can reasonably trust the output predictions for the test set), and how it compares to leading methods in predicting drivers for MV. Past these results, we will also discuss MAgEC itself; specifically how it responds to shifts in perturbation levels and whether the usage of non-homogenous models are truly exploring different parts of the problem space.

Mechanical Ventilation Prediction

Table 2 compares different metrics for the three models that make up our ensemble, along with the ensemble prediction itself, for the task of MV prediction three hours prior to the ventilation occurring. As this table is being shown purely to demonstrate that MAgEC is not compromising on model accuracy, we will compare ourselves to the current state of the art in MV prediction[8] (also shown in last column). We have tried to replicate our cohort selection process to match theirs, but we still end up with roughly twice as many MV patients as they have, so our results are not directly

Feature Type	Feature Extracted	Variable
Vital Sign	24 hourly measurements (averages) for heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, respiration rate, temperature, blood oxygen level and glucose. First and last hourly averages are extracted.	$\{first/last\}\_heartrate\_mean$, $\{first/last\}\_sysbp\_mean$, $\{first/last\}\_diasbp\_mean$, $\{first/last\}\_meanbp\_mean$, $\{first/last\}\_resprate\_mean$, $\{first/last\}\_tempc\_mean$, $\{first/last\}\_spo2\_mean$, $\{first/last\}\_glucose\_mean$
Lab Measurement	Daily (latest measurement) for anion gap, albumin, bicarbonate, bilirubin, creatinine, lactate, magnesium, phosphate, platelet count, potassium, partial thromboplastin time, prothrombin (international normalized ratio and time), sodium, blood urea nitrogen and white blood count.	$aniongap$, $albumin$, $bicarbonate$, $bilirubin$, $creatinine$, $chloride$, $glucose$, $hemoglobin$, $lactate$, $magnesium$, $phosphate$, $platelet$, $potassium$, ptt, inr, pt, $sodium$, bun, wbc
Demographics	Current value	age, $gender$
Co-morbidities	Current value (0/1)	$congestive\_heart\_failure$, $chronic\_pulmonary$, $pulmonary\_circulation$

Table 1: List of features extracted from each patient. Vital sign measurements are extracted every hour (average values within each 1h time period) and are joined with laboratory values, demographics and comorbidities. Missing values are imputed with averages for a given feature. 'First' vital signs refer to the first value found in the data-collection period, while 'last' refers to the last value found in the data collection period.

comparable.

However, we can still demonstrate somewhat similar results. Our AUC in particular has only a .04 deviation from theirs, but our specificity's and sensitivities are off by nearly .1. We suspect this is due to our dataset including more 'positive' cases than the other paper. We believe this is sufficient enough evidence to show that MAgEC is performing roughly as well as the next leading method, and that our proposed method to find drivers of MV does not hurt accuracy.

	LR	SVM	MLP	Ensemble	SOTA[8]
Sensitivity	0.67	0.69	0.64	0.61	0.78
Specificity	0.67	0.80	0.85	0.88	0.79
F1 score	0.45	0.56	0.58	0.59	-
ROC AUC	0.72	0.81	0.83	0.83	0.87
Confusion Matrix	1120 546 136 281	1344 322 129 288	1449 217 156 261	1478 188 167 250	-

Table 2: Individual and ensemble model metrics in predicting unexpected MV.

Comparison with SHAP

Figure 2 shows a comparison between aggregate predicted MV drivers of both MAgEC (left) and SHAP. We condition these importances on whether a patient was ventilated (label 1, top charts) or not ventilated (label 0, bottom charts). The left-side charts show MAgEC ranking, while the right-side charts show SHAP ranking.

All MAgEC top aggregate features are generated using 'full' perturbations (explained in the following subsection), and are based on the number of times a feature was ranked as the top-most important feature for any patient in the test set. The SHAP values are generated using TreeSHAP with default parameters, and aggregated based on the mean

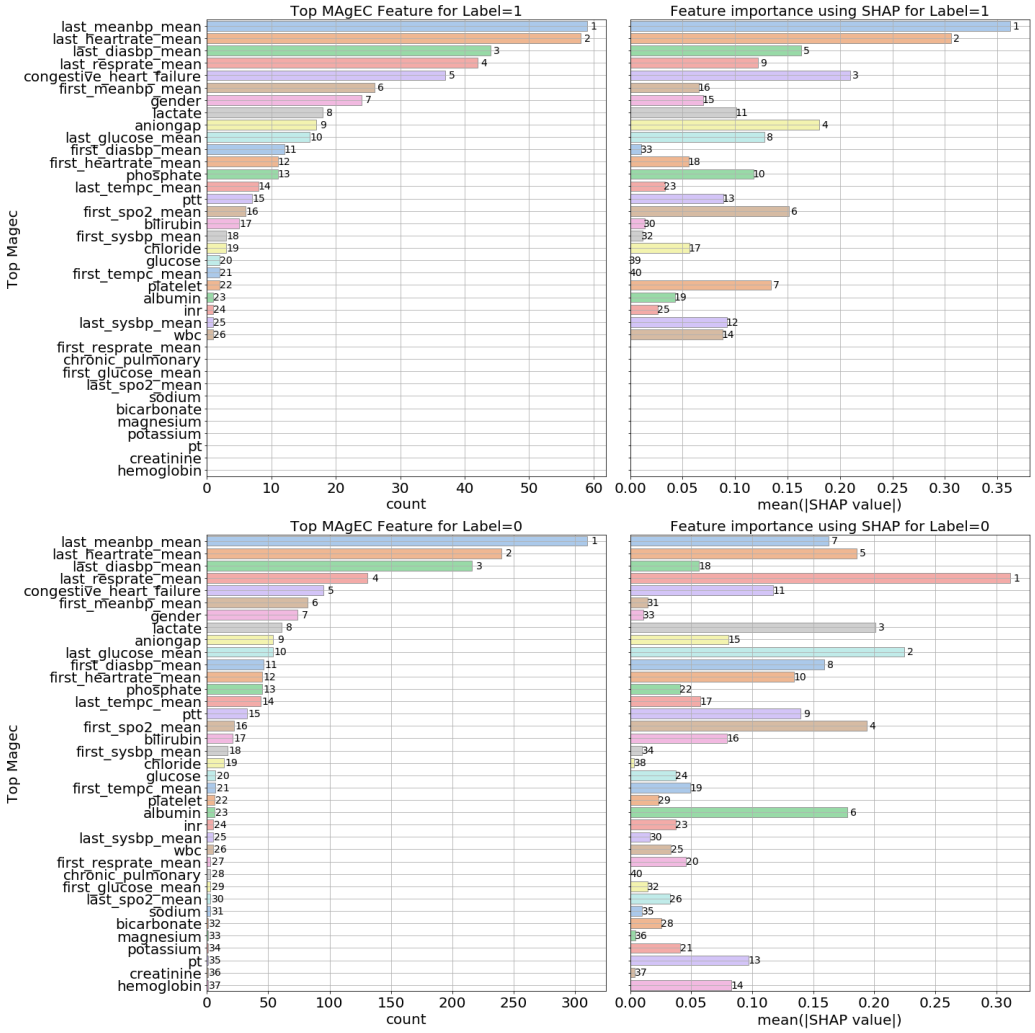


Figure 2: Feature Importance, as measured by the frequency of a given feature ranking first in a panel of 3 models (using an aggregate sum of MAgECs from all 3 models), is shown on the left. Corresponding SHAP values, using an xgboost model tree explainer, are shown on the right. Feature importance rankings are shown on the right of each bar. A significant overlap between SHAP values and MAgEC is observed (7/10 features overlap for label=1 and 5/10 features for label=0 in the top 10 MAgEC features). Feature importance varies with the class label for both MAgEC and SHAP.

absolute SHAP value. All charts are ordered according to the MAgEC count ranking.

Amongst both label 1 and 0, we can apply some basic clinical 'sanity checks' to check the validity of some of the rankings of both MAgEC and SHAP. Specifically, we find that the 'last' readings of vital signs, as in the last recorded reading of it, are generally seen as more important than the 'first' readings of it, which generally matches a clinicians expectation for the predictive value of a vital sign.

Furthermore, we generally find a good matching between MAgEC and SHAP for label 1 (ventilated), with several of the top-most features matching identically in ranking, and a similar set of features found to be of near-zero importance in each. There is a fair bit of disagreement, but a similar trend can be observed in each in regards to the general importance of features in driving ventilation. We also note that MAgEC identified phosphate, which has been clinically validated as a predictor for future need of MV[14], as an important feature in nearly 20 patients, while the SHAP values

indicate phosphate as nearly useless. Gender is also found to be among the top features, in both MAgEC and SHAP, which could indicate the presence of bias in receiving MV as observed in other studies[15].

On the other hand, we find far more disagreement in the plots conditioned on label 0 (unventilated). While there is some slight overlap in the general importance of some features versus others, MAgEC and SHAP largely disagree with each other. We argue that this is not necessarily a flaw of MAgEC, as, clinically, the reason why a clinical event never occurred is far more complex and nuanced than the reason why a clinical event did occur. In this sense, it is possible that neither SHAP nor MAgEC captured the true drivers for why some patients were not ventilated.

Stability of Predicted Drivers

We measure the stability of MAgEC-derived feature importances underneath different levels of perturbation in the MAgEC coefficient calculation. These results are shown in Figure 3. Full perturbations set a feature to the training cohort mean, .5 perturbations move a feature 50-percent towards the cohort mean, .3 to 30-percent of the cohort mean, and .1 to 10-percent towards the cohort mean.

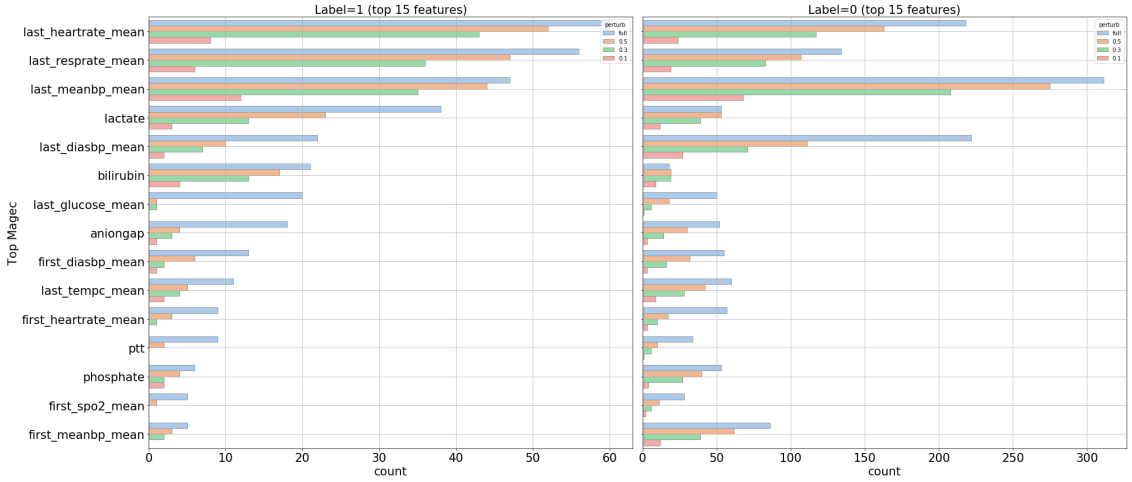


Figure 3: MAgEC Feature Importance of non-boolean features as a function of the perturbation strength (left plot for patients who received MV and right plot for patients who did not receive MV). For small perturbations (e.g. perturbing a feature by 10%), boolean features such as gender (not shown in the figure) become relatively more important than numerical features. MAgEC feature importance scores scale proportionally to the perturbation strength.

We generally find that the relative ranking of the features stay roughly the same at different levels of perturbations, which lends credence to MAgEC being relatively stable. However, we do notice that the distribution of importance flattens as the perturbations become smaller, with individual feature importances scaling proportionally to the perturbation strength. Due to this, for the purposes of finding drivers of ventilation, we primarily present our results using the full perturbation.

Inter-Model Agreement

A key assumption of the non-homogeneous model setup is that the usage of completely different models in our ensemble is more useful than just adjusting the random seed of multiple otherwise identical models. 'Usefulness' in this sense refers to whether each model is capable of being different in its assessment of feature importance, instead of producing nearly the same feature rankings. Understandably, this definition of 'useful' is dangerous, as even an ensemble of untrained, inaccurate models would be classified as 'useful', since they'd produce wildly different feature importance rankings. However, these results should be viewed in light of previous results, demonstrating both accuracy in predicting MV and a reasonable similarity to SHAP in deriving feature importance in ventilation.

To show that the models in our ensemble are generating varied rankings of drivers, we start with the normalized

MAgEC coefficients generated in Equation 2. Each model outputs an ordered list of MAgECs (in ascending order). To quantify the relative agreement amongst individual model rankings, we use rank-biased overlap[16], a method for comparing the similarity between two ranked lists. RBO is implicitly used to assess inter-model ranking variability. In our case, the ranked lists are the ordered normalized MAgEC coefficients of each individual model. These coefficients are then grouped into pairs, and used to calculate an RBO (Equation 3). We do not include the actual RBO equation here for brevity's sake, but can be found in the citations.

$$R_j^{(k)} = N(N-1)/2 \sum_{\substack{m=1,n=1 \\ m \neq n}}^{N} RBO(\tilde{m}_{jm}^{(k)}, \tilde{m}_{jn}^{(k)}) \tag{3}$$

The output of RBO is a value between 0 and 1, where 0 represents perfectly disjoint and 1 represents perfectly identical for any two given lists. In the case of N number of models, there would be a total of N*(N-1)/2 RBO values per patient. To quantify inter-model agreement for any given feature, we simply take all patients whose top feature is that given feature and average the RBO's across these patients.

For instance, if we have 4 models, each patient will have 6 RBO values. To quantify model agreement, lets suppose we have 10 patients, each with 3 features: A, B, and C. Each patient is run through MAgEC, which says that 6 of the patients top feature is A, 4 patients have a top feature of B, and none have a top feature of C. Thus, to calculate the model agreement on A, we simply take the subset of 6 patients with A being their top feature, and average their individual RBO's (of which where will be 24 in total). This identical flow can be extended to the model agreement on B. However, for C, due to no patients having it as a top feature, we cannot calculate a model agreement value for that feature.

In our case, we limit our number of models to three, and compare the average non-homogeneous model set used for MAgEC (logistic regression, SVM, and a neural network) mean RBO versus a homogenous model set (three neural networks, trained with different random seeds). Figure 4 shows the results of this. Overall, we find that the heterogeneous ensemble leads to slightly lower mean-RBO values compared to those of the homogeneous ensemble, which leads credence to our hypothesis that differing models leads to higher feature importance diversity.

Conclusion

MAgEC is able to identify patient level risk factors and prioritize among clinical drivers of mechanical ventilation by using ensemble models and simple consensus protocols. When compared to state of the art methods of feature importance (SHAP), MAgEC demonstrated comparable efficacy, feature importance diversity, and overall speed. We also found that the unique, model-agnostic nature of MAgEC to be not at all detrimental to its performance, and that the non-homogeneity that this allows leads to more diverse 'opinions' in feature importance within the ensemble.

Furthermore, the ability to identify reasonable targets of intervention, especially for a majority of high-risk patients, is a function not just of model performance but also of how a clinically normal value is specified. For instance, studies have shown that levels of A1C values in non-diabetic populations vary with age[17], or that normal BMI values may also vary with age and gender[18]. By providing the flexibility for the individual researcher to define 'normal' or the amount of movement towards 'normal' (e.g. make the patient exactly normal or move them some portion of a standard deviation towards normal), MAgEC allows for the selection of aggressive or conservative approaches which can be adapted to most circumstances.

The primary limitation of this work is that, as we were unable to replicate the patient cohort used in the current state-of-the-art comparision[8], it is difficult to compare our predictions in terms of accuracy (although this isn't the primary purpose of this work). Regarding future work, we'd like to explore the use of MAgEC in assisting clinicians in identifying potential interventions via highly ranked features.

This work is intended for research purposes only. It in no way proposes to replace panels of clinical experts in the situations that they are currently used (such as tumor boards) but instead to extend their functionality into places that they are not currently available. Clear prioritization of which patients are most likely to respond to intervention, and

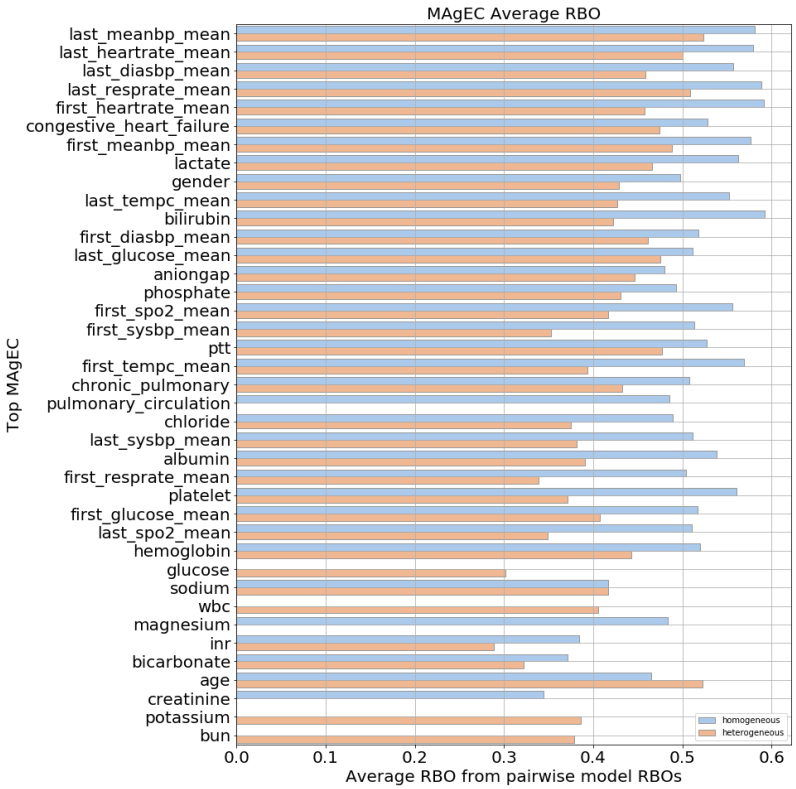


Figure 4: Average RBOs for top MAgECs (using 'full' perturbations). MAgECs are shown in decreasing feature imporance order (from top to bottom). RBO values are computed using p=0.9 (first 10 ranks have 86% of the weight in the metric). RBOs close to 1 would indicate no variability in the model outputs, such as when 2 models are almost identical, while values close to 0 would indicate very little overlap. A panel of 3 homogeneous models (MLP using 3 different initialization weights) and 3 heterogenous models (LR, SVM and MLP) is shown). Average inter-model agreement for the homogeneous panel is higher for all features (and larger than 0.5 for most of them) except for one feature ('age').

for them, which clinical variables should be targeted is likely to be most useful in situations where healthcare systems are under-staffed, overwhelmed, or have providers serving outside of their area of expertise.

The code for implementing MAgEC, alongside all experiments ran in this paper, is open-sourced on GitHub: https://github.com/gstef80/MAgEC.

Acknowledgments

We wish to thank Dave Prakash, Chinmay Belthangady, Axel Bernal, Milena Gianfrancesco, Adams Dudley and Atul Butte for fruitful discussions and comments in early versions of this work.

References

1. Bourel M, Crisci C, Martínez A. Consensus methods based on machine learning techniques for marine phytoplankton presence-absence prediction. Ecological Informatics. 2017;42.
2. Tuv E, Borisov, A., Runger, G, Torkkola K. Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. Journal of Machine Learning Research. 2009;10(45):1341-66.
3. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–77.
4. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 1135–44.
5. Ladha K, Vidal Melo MF, McLean DJ, et al. Intraoperative protective mechanical ventilation and risk of postoperative respiratory complications: hospital based registry study. BMJ. 2015;351:h3646. Published 2015 Jul 14. doi:10.1136/bmj.h3646
6. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
7. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics. 2015;24(1):44-65.
8. Ren O, Johnson AEW, Lehman EP, Komorowski M, Aboab J, Tang F, et al., editors. Predicting and Understanding Unexpected Respiratory Decompensation in Critical Care Using Sparse and Heterogeneous Clinical Data. 2018 IEEE International Conference on Healthcare Informatics (ICHI); 2018 4-7 June 2018.
9. Suresh H, Hunt N, Johnson A, Celi LA, Szolovits P, Ghassemi M. Clinical Intervention Prediction and Understanding with Deep Neural Networks. In: Finale D-V, Jim F, David K, Rajesh R, Byron W, Jenna W, editors. Proceedings of the 2nd Machine Learning for Healthcare Conference; Proceedings of Machine Learning Research: PMLR; 2017. p. 322–37.
10. Desautels T, Calvert J, Hoffman J, et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. JMIR Med Inform. 2016;4(3):e28.
11. Mohamadlou H, Lynn-Palevsky A, Barton C, et al. Prediction of Acute Kidney Injury With a Machine Learning Algorithm Using Electronic Health Record Data. Can J Kidney Health Dis. 2018;5:2054358118776326.
12. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform. 2018;22(5):1589-1604.
13. Song C, Zhang, S., Sadoughi, N., Xie, P., Xing, E. Generalized Zero-Shot Text Classification for ICD Coding. International Joint Conferences on Artificial Intelligence Organization. 2020(7):4018–24.
14. Talakoub R, Bahrami M, Honarmand A, Abbasi S, Gerami H. The Predicting Ability of Serum Phosphorus to Assess the Duration of Mechanical Ventilation in Critically Ill Patients. Adv Biomed Res. 2017;6:51.
15. Shen HN, Lu CL, Yang HH. Women receive more trials of noninvasive ventilation for acute respiratory failure than men: a nationwide population-based study. Crit Care. 2011;15(4):R174.
16. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans Inf Syst. 2010;28(4):Article 20.
17. Pani LN, Korenda L, Meigs JB, Driver C, Chamany S, Fox CS, et al. Effect of aging on A1C levels in individuals without diabetes: evidence from the Framingham Offspring Study and the National Health and Nutrition Examination Survey 2001-2004. Diabetes Care. 2008;31(10):1991-6.
18. Xu L, Au Yeung S, Schooling CM. Does the optimal BMI really vary by age and sex? International Journal of Epidemiology. 2015;45(1):285-6.

Analysis of Health Trajectories Leading to Adverse Opioid-Related Events

Aidan S. Gilson, BS[1], David Chartash, PhD[2], David Chang, BS MS[2], Kathryn Hawk, MD MHS[3], Gail D'Onofrio, MD MS[3,5], Adrian D. Haimovich, MD PhD[2,3], David A. Fiellin, MD[3,4,5], R. Andrew Taylor, MD MHS[2,3]

[1]Yale School of Medicine, New Haven, CT; [2]Center for Medical Informatics, Yale School of Medicine, New Haven, CT; [3]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT; [4]Yale School of Public Health, New Haven, CT; [5]Department of Medicine, Yale School of Medicine, New Haven, CT

ABSTRACT

Identifying patient risk factors leading to adverse opioid-related events (AOEs) may enable targeted risk-based interventions, uncover potential causal mechanisms, and enhance prognosis. In this article, we aim to discover patient diagnosis, procedure, and medication event trajectories associated with AOEs using large-scale data mining methods. The individual temporally preceding factors associated with the highest relative risk (RR) for AOEs were opioid withdrawal therapy agents, toxic encephalopathy, problems related to housing and economic circumstances, and unspecified viral hepatitis, with RR of 33.4, 26.1, 19.9, and 18.7, respectively. Patient cohorts with a socioeconomic or mental health code had a larger RR for over 75% of all identified trajectories compared to the average population. By analyzing health trajectories leading to AOEs, we discover novel, temporally-connected combinations of diagnoses and health service events that significantly increase risk of AOEs, including natural histories marked by socioeconomic and mental health diagnoses.

INTRODUCTION

Background and Significance

Adverse opioid events led to more than 45,000 deaths in 2018, a six-fold increase since 1999.[1] Understanding and combating the opioid epidemic requires identification of patient risk factors which may enable targeted risk-based interventions, uncover potential causal mechanisms, and enhance prognosis. While past work has identified a number of individual risk factors related to adverse opioid-related events (AOEs) including race, age, gender, veteran status, and the prescription of opioids, these studies analyzed only a handful of risk factors and did not consider the temporal or sequential trajectory of patient healthcare events.[2-7] Few large-scale temporal risk studies have been completed regardless of disease focus, and none have assessed opioid events as the terminal outcome.[8-12]

Trajectory based assessment of patient risk can uncover unique associations that may not be obvious or possible using conventional methods. This methodology identifies time-dependent event associations, or trajectories, made from temporally ordered pairs of healthcare codes that occur frequently in a population.[8] This is more meaningful then simply testing for co-occurrence of multiple coded events, as the temporal nature of the analysis necessitates the risk factor preceded the AOE. These trajectories are potentially more indicative of underlying causal pathways. For example, this approach allows testing for hypothesized natural histories like a pain event leading to opioid prescription and then an adverse opioid event. Examples of relationships found with this methodology include a correlative relationship between sleep apnea and diabetes, and disease trajectories that were identified to predict increased sepsis-related mortality.[9, 10]

Objective

In this work, we expand on previous efforts describing risk factors for adverse opioid-related events in three ways. First, we use an unbiased data mining approach across all prior diagnoses, medications, and procedures to uncover potential novel individual risk factors in a large patient population. Second, we identify temporally-restricted trajectories.[8] Third, we cluster patient-level trajectories to gain insight into more general trends present at the level of an organ system, disease group, or drug class. This allows us to identify general risk pathways related to pain, socioeconomic events, mental health diagnoses and opioid prescription. These novel risk trajectories can then be used to aid targeted clinical interventions and help begin to reduce the number of adverse events suffered from the opioid epidemic.

MATERIALS AND METHODS

Study Data

In this retrospective cohort study, we used health records from 1.6 million patients seen in a large New England Healthcare System in any capacity, inpatient, outpatient, or emergency department, between the years 2013 and 2019. The dataset includes 677,000 instances of Common Procedural Technology (CPT) codes, 11 million prescriptions for opiate medications in the form of RxNorm codes, and 110 million ICD-10 code diagnoses. Patients with complete demographic information including age, race, and sex were included. Following previously published methodology, we only included the first instance of a code, as it could appear multiple times in a patient's electronic health record (EHR).[8]

We define an AOE using ICD-10 terminology. Codes in the range T40.0 to T40.6, or in the F11 subchapter are considered adverse opioid-related events. Within this group we exclude all codes related to remission, or underdosing of opioids, as well as all codes in the T40.5 subchapter. This leaves codes related to opioid use or dependence disorders, as well as poisoning or adverse effects of opioids.

Trajectory Identification

The method used to identify diagnostic temporal pairs and extract trajectories is outlined below, and fully described in prior work.[8] During the identification of temporal pairs, all ICD-10 codes were truncated to three characters with the exception of codes in the F11 and T40 sub-chapters as these were used as the primary outcome.

Fisher Exact tests to identify co-occurrence ($p<0.05$) and binomial tests ($p<0.05$) to identify a temporal relation, both using Bonferroni correction, were used to identify initial pairs of codes of the form (C1→C2). Further reduction of these pairs was performed by comparing each patient that followed C1 and C2 sequentially within a cohort of 10,000 other patients who were of the same sex, race, and whose ages fell in the same decade of life. Inclusion of a pair was dependent on the relative risk of co-occurrence of C1 and C2 being greater than one and $p<0.05$. $P_{exposed}$ and P_i are the number of patients identified that follow C1→C2 from the control group and from the i^{th} comparison group. N is the number of comparison groups, in this case 10,000. RR and p are then given by,

$$RR = \frac{P_{exposed}}{\frac{1}{N}\sum_i P_i}, \qquad p < \frac{1}{N}\{i | P_i \geq P_{exposed}\}$$

This ensured that an ordered pair C1→C2 was included only if C1 indicated an increased risk of subsequent diagnosis of C2 independent of patient demographic, rather than the relationship being due to skewed presentation of the diagnoses along demographic lines.

Trajectories were then formed by combining two or more overlapping ordered pairs (C1 →C2 and C2→C3 become C1→C2→C3). In total, 785 trajectories were found that were present in at least 10 patients. The trajectories ranged from three to seven codes in length. All trajectories were then combined to form a directed acyclic graph (DAG). Each node was plotted according to a feature space embedding, described later, using t-SNE to reduce the dimensionality to a 2D coordinate system.[13] The complete workflow is shown in Figure 1.

Dynamic Time Warping

In order to identify more general trends, the trajectories were clustered into groups using an unsupervised algorithm called Dynamic Time Warping (DTW).[11] DTW clusters trajectories of varying lengths by "warping" the longer trajectory so that multiple nodes may correspond to a single node in the shorter trajectory. The distance between two trajectories of the form $t_1 = [C_1, C_2, C_3, \ldots, C_N]$ and $t_2 = [C_1, C_2, C_3, \ldots, C_M]$ is given by $D(n, m)$ where each element in D is defined as:

$$D(n,m) = Distance(n,m) + min(D(n-1,m), D(n,m-1), D(n-1,m-1))$$

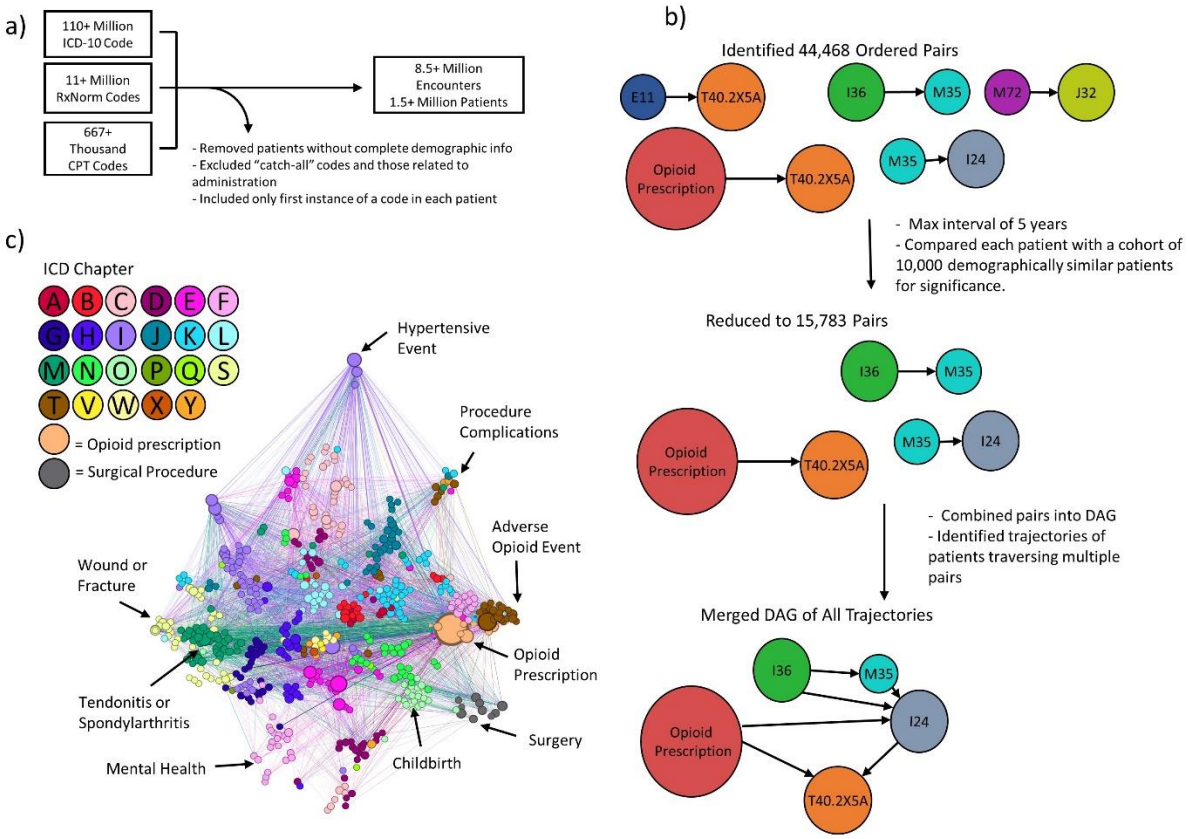


Figure 1: a) Initial data pruning. **b)** Workflow from temporal pairs to trajectories and directed acyclic graph (DAG). **c)** Complete DAG formed from all identified trajectories. Visualization is performed through dimensionality reduction of the 512-dimension

$C_1, C_2, C_3, \ldots, C_N$ represent temporally ordered codes in a patient's health record. $Distance(n, m)$ represents the distance between the n^{th} code in t_1 and the m^{th} code in t_2. In past work the distances between codes was defined according to the hierarchical nature of the ICD9 coding system.[11] As this method is not possible for CPT and RxNorm codes, all codes including ICD-10, RxNorm, and CPT codes were converted to Concept Unique Identifiers (CUI) using the UMLS Metahesaurus.[14] The codes were then embedded into a 512-dimensional feature space using the RotatE knowledge graph embedding model applied on the SNOMED-CT terms included in the UMLS Metathesaures.[15] Distances between codes were then calculated using several metrics in the embedding space. Euclidean, Chebyshev, and cosine distance metrics were used to validate clusters and show that the choice of distance metric had little effect on the resulting clusters. $D(n, m)$ can be calculated pairwise between all trajectories, resulting in a scalar distance between all possible pairs.

To form clusters of trajectories, the first trajectory is assigned to an initial arbitrary cluster. For each subsequent trajectory, the mean distance between the trajectory and every cluster is calculated by taking its average distance to trajectories in each cluster. If the minimum average distance found for any cluster is less than the global threshold, the new trajectory was added to that cluster. If no distance was less than the threshold, a new cluster was created for the trajectory. Threshold values were picked manually to balance excessive fragmentation and merging of clusters.

To calculate the relative risk for each trajectory and cluster, the patients who followed the trajectory up to the penultimate code before an opioid event were selected. The number of patients who went on to experience an opioid event were identified from this group. The rate of opioid events for this set of patients was then compared to the rate for the complete set of patients. For example, consider a population of 1,000 patients and a trajectory C1→C2→C3→C4, where 100 patients in the population experience C4. If 100 patients followed C1 through C3, and 50 of those patients went on to experience C4, the RR for C4 of the trajectory compared to the population is:

$$RR = \left(\frac{50}{100}\right) / \left(\frac{100}{1000}\right) = 5$$

All code is available at: https://github.com/Aidan-Gilson/Opioid-Trajectories.

RESULTS

After preprocessing, there were 8.5 million encounters and 1.5 million patients on which the analysis was performed. There were 15,783 temporally ordered pairs of the form C1→C2 generated from the patient cohort analysis (Fig 1a,b). Table 1 shows the 25 ICD-10 codes with the highest relative risks for any adverse opioid event. Conventional calculation of risk would likely overestimate the RR of the shown pairs, as adverse opioid-related events that occurred before the code of interest may be included. As such, conventional RR implies comorbidities in patients, while the temporal RR begins to extract underlying directionality. The individual temporally preceding factors associated with the highest relative risk were toxic encephalopathy, with a RR of 26.1, followed by septic arterial embolism, problems related to housing and economic circumstances, and unspecified viral hepatitis, with RR of 24.4, 19.9, and 18.7 respectively.

Table 1: Codes with the highest RR leading to an adverse opioid event

ICD Code	Name of diagnosis leading to opioid event	RR	95% CI	# of Patients
G92	Toxic encephalopathy	26.1	24.6-27.6	1011
I76	Septic arterial embolism	24.4	19.0-31.1	55
Z59	Problems related to housing and economic circumstances	19.9	19.0-20.7	2144
B19	Unspecified viral hepatitis	18.7	18.1-19.3	3512
Z56	Problems related to employment and unemployment	17.7	15.5-20.0	220
Y90	Evidence of alcohol involvement determined by blood alcohol level	17.5	15.9-19.1	415
Z81	Family history of mental and behavioral disorders	16.2	15.1-17.3	753
B17	Other acute viral hepatitis	15.8	13.6-18.1	172
B18	Chronic viral hepatitis	14.3	13.8-14.9	2499
X78	Intentional self-harm by sharp object	13.8	11.4-16.5	102
F60	Specific personality disorders	13	12.3-13.8	988
M96	Intraoperative and postprocedural complications and disorders of musculoskeletal system, not elsewhere classified	12.7	11.8-13.6	748
I33	Acute and subacute endocarditis	12.6	11.0-14.4	194
F39	Unspecified mood [affective] disorder	12.5	12.1-13.0	2770
X83	Intentional self-harm by other specified means	12.5	9.9-15.6	70
G06	Intracranial and intraspinal abscess and granuloma	11.1	9.3-13.2	122
R45	Symptoms and signs involving emotional state	11.1	10.7-11.4	4103
M90	Osteopathies in diseases classified elsewhere	10.5	8.0-13.5	53
Z21	Asymptomatic human immunodeficiency virus [HIV] infection status	10.2	9.4-10.9	630
K72	Hepatic failure, not elsewhere classified	10.2	9.4-10.9	711
Y95	Nosocomial condition	9.6	8.4-11.0	198
F31	Bipolar disorder	9.6	9.3-9.9	4142
K71	Toxic Liver Disease	9.1	6.9-11.9	51
M00	Pyogenic arthritis	8.8	7.6-10.1	186
J80	Acute respiratory distress syndrome	8.8	7.4-10.3	137

By combining sequential pairs, trajectories leading to a specific adverse opioid event were then identified. A subset of trajectories and their relative risks compared to the complete patient set are shown below (Table 2). Trajectories were selected from the 30 trajectories with the highest RR based on novelty.

Each code was assigned a unique 512-dimensional feature space embedding as described in the methods. The embeddings were used to visualize the DAG constructed from the combination of all trajectories. The complete DAG is shown in Figure 1c, with the nodes colored according to ICD-10 chapter heading. RxNorm, and CPT codes were colored manually with colors not used already.

The trajectories were then clustered using DTW. This allowed for clusters to be formed from trajectories that followed a similar qualitative path, even if no exact nodes were shared. For example, the two trajectories,

$$\text{Other and unspecified osteoarthritis} \rightarrow \text{Opioid analgesics} \rightarrow \text{Adverse effect of methadone initial encounter}$$

$$\text{Spondylosis} \rightarrow \text{Opioid analgesic, anesthetic adjunct agents} \rightarrow \text{Adverse effect of other opioids initial enoucnter}$$

would be clustered. Even though no two codes within each trajectory are the same they are qualitatively similar. The clusters were then visualized in graphing software and reviewed manually. A selection of clusters with RR>1 and large patient sample are show in Figure 2a.

Table 2: Diagnostic trajectories and RR of an opioid event

Path leading to opioid event	**RR**	**95% CI**
[Other and unspecified soft tissue disorders, not elsewhere classified] → [Opioid analgesic, anesthetic adjunct agents] → [Bacterial infection of unspecified site]	11.24	7.4-16.7
[Opioid dependence uncomplicated] → [Opioid analgesic, anesthetic adjunct agents]	9.4	7.3-11.9
[Other disorders of cartilage] → [Other deforming dorsopathies]	8.7	5.7-13.2
[Dorsalgia]→ [Opioid analgesic, anesthetic adjunct agents] → [Opioid antitussive-expectorant combination]	8.4	5.5-12.7
[Other psychoactive substance related disorders] → [Opioid analgesic, anesthetic adjunct agents]	7.8	6.1-9.8
[Essential (primary) hypertension] → [Opioid analgesic, anesthetic adjunct agents] → [Opioid antitussive-expectorant combination]	6.6	4.8-9.0
[Unspecified viral hepatitis] → [Opioid analgesic, anesthetic adjunct agents]	6.5	5.0-8.5
[Malignant neoplasm of prostate] → [Other functional intestinal disorders]	6.4	4.6-8.8
[Cocaine related disorders] → [Opioid analgesic, anesthetic adjunct agents]	6.2	4.4-8.7
[Disorders of lipoprotein metabolism and other lipidemias] → [Opioid analgesic, anesthetic adjunct agents] →[Bacterial infection of unspecified site]	6.1	4.0-9.1
[Type 2 diabetes mellitus] → [Other specified diabetes mellitus] → [Opioid analgesic, anesthetic adjunct agents]	6	3.9-9.0
[Type 2 diabetes mellitus] → [Other deforming dorsopathies]	5.5	4.0-7.4
[Essential (primary) hypertension] → [Other and unspecified arthropathy]	5.3	3.7-7.6
[Malignant neoplasm of breast] → [Secondary malignant neoplasm of other and unspecified sites]	4.9	3.6-6.4
[Type 2 diabetes mellitus] → [Other functional intestinal disorders]	4.4	3.9-4.9

To demonstrate the effect that differing natural histories may have on the risk incurred from a single diagnosis, multiple clusters all containing codes related to bursitis, arthritis, and tendinitis were compared, clusters are shown in Figure 2b. The RR for an individual diagnosed with a bursitis, arthritis, or tendinitis code, independent of other past medical history, was calculated at 4.3. The inclusion of select nodes can also raise or lower the risk incurred from bursitis, arthritis, and tendinitis. The cluster in Figure 2b which contains opioid prescriptions and various blood disorders has an increased RR of 7.2. However, when only including opioid prescription, as shown in another cluster,

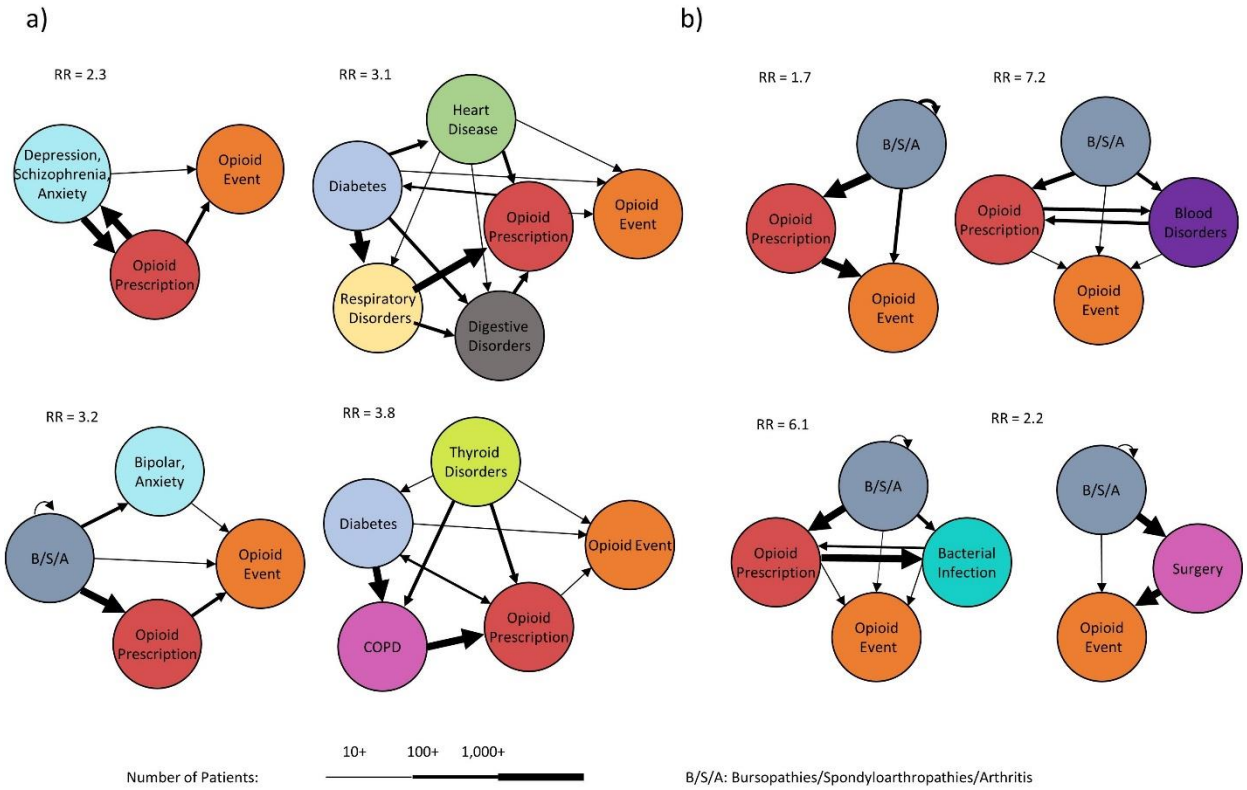


Figure 2: a) Example clusters identified through dynamic time warping (DTW). **b)** Clusters containing bursopathies, spondyloarthropathies and/or arthritis.

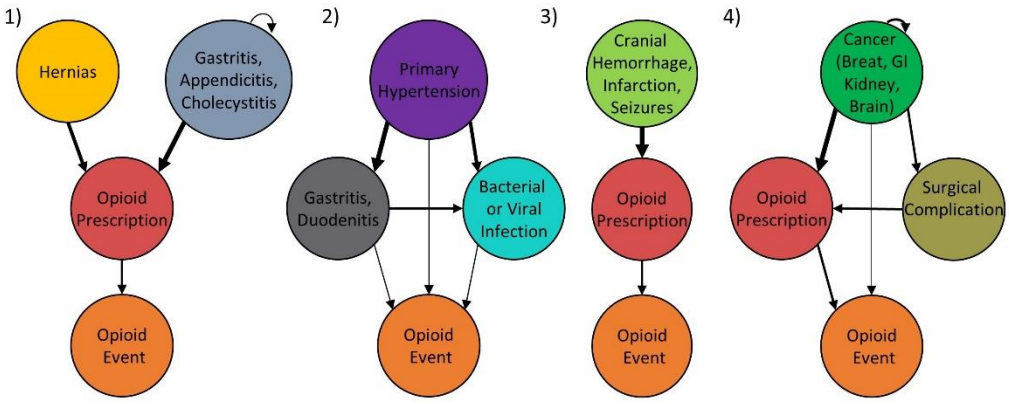
the patients risk falls to the same risk as the general cluster at 1.7. The final two clusters, containing opioid prescriptions and bacterial infection, and surgery respectively demonstrate two more examples of intermediate diagnoses affecting the RR of bursitis, arthritis, and tendinitis.

Figure 3 shows how the RR of trajectories and clusters changes when stratifying for only patients with a socioeconomic or mental health code in their history. Clusters selected all had baseline RR under two in order to demonstrate the dramatic effect that patient stratification based on mental health and socioeconomic status can have.

DISSCUSION

Through the analysis of 1.5 million patient records, we identified multiple novel temporally paired risk factors and 785 health care event trajectories that terminate in an AOE. Further clustering of these trajectories identified several trajectory-based risk factors for adverse opioid events that provide new insight into patients with AOEs and their preceding temporal health care events within electronic health records. It is worth noting that some of the top individual factors that were identified, toxic encephalopathy, septic arterial embolism, and viral hepatitis, are all possible sequelae of intravenous drug use (IVDU), and therefore their correlation with adverse opioid-related events is not unexpected. Their presence among the top risk factors provides justification for the validity of the methodology, by extracting expected risk factors even with an unbiased approach.

Two major groups of healthcare codes are present in the list of single risk factors, socioeconomic factors, and mental health events. For socioeconomic factors, problems related to housing and economic circumstances, and problems related to employment and unemployment were the factors with the 3rd and 5th highest RR for an adverse opioid event, at 19.9 and 17.7 respectively, as shown in Table 1. For mental health factors, family history of mental and behavioral disorders (16.2), intentional self-harm by sharp object (13.8), specific personality disorders (13.0),



	Cluster #1		Cluster #2		Cluster #3		Cluster #4	
	Value	95% CI	Value	95% CI	Value	95% CI	Value	95% CI
RR	1.1	1.0-1.3	1.7	1.5-2.0	1.1	1.0-1.4	1.9	1.6-2.3
RR w/ SES	7.9	6.0-10.5	4.2	2.7-6.7	7.6	5.2-11.0	6.5	3.4-12.4
RR w/ Mental Health	2.7	2.3-3.1	3.0	2.5-3.6	2.4	2.0-3.0	4.2	3.4-5.2
RR w/ SES & Mental Health	9.5	7.1-12.7	5.2	3.3-8.1	9.5	6.6-13.8	8.3	4.4-15.8

Figure 3: Examples of the effect of socioeconomic and mental health diagnosis can have on RR of trajectories and clusters.

Intentional self-harm by other specified means (12.8), unspecified mood [affective] disorder (12.5), and symptoms and signs involving emotional state (11.1), were all present in the 25 factors with the highest RR.

Most likely due to a lack of temporal specificity, socioeconomic factors were not present in any of the identified trajectories. To examine the effect of socioeconomic factors further, a cohort of patients who experience at least one socioeconomic event, defined as ICD-10 codes Z55-Z65, was compared to the average population. For the top trajectories in Table 2, patients with a socioeconomic event made up 9.5% of the patients who traversed the trajectory up to an opioid event, and 23.5% of the patients who went on to experience an adverse opioid event. For the clusters shown in Figure 3, patients with a socioeconomic event made up only 6.1% of patients traversing the cluster, but 16.7% of patients who experienced an adverse opioid event. In both cases patients with socioeconomic events were over twice as likely to experience an adverse opioid event after following the trajectory compared to random chance.

Mental health factors were present in multiple trajectories and clusters, two examples are shown in Figure 2a. Patients who experienced at least one mental health event made up an even greater percentage of the patients of the top trajectories who went on the experience an adverse opioid event. For trajectories shown in Table 2, patients with a mental health event made up 60.2% of the patients who traversed the trajectory, but 85.3% of the patients who experienced an adverse opioid event.

A general trend appearing in over 50% of trajectories, was an initial diagnosis or procedure which led to pain, followed by the prescription of opioids for pain management, and finally the adverse opioid event. For example, the first, third, and fourth trajectories outlined in Table 2 follow this pattern. Diagnoses likely to cause pain or discomfort, soft tissue disorders, cartilage disorders, and dorsalgia respectively, lead to opioid prescription and an adverse opioid event.

Event trajectories lacking a prescription heavily implicate a role of pain in the trajectory to an adverse opioid event. The lack of a prescription may also suggest non-prescription acquisition or a prescription outside the healthcare system. Many trajectories include a diagnosis that would cause significant pain or discomfort for the patient, strongly

suggesting the use of some drug management, opioid or otherwise. For example, the final trajectory shown in Table 2:

Type 2 Diabetes mellitus → other functional intestinal disorders → opioid event

may imply the use of non-prescribed opioids by a patient with an emergency department visit for gastric pain. Indeed, a separate, though less populated cluster contains opioid prescription subsequent to Type 2 Diabetes and intestinal disorders. This suggests a future hypothesis as to whether opioids are prescribed too often for dysmotility disorders, as opioids are known to reduce peristalsis, which can potentially exacerbate a patient's symptoms and predispose to the development of an opioid use disorder.[16]

This trend of pain, prescription, and non-medical use fits well with previously described literature. Of patients prescribed opioids for chronic pain, around 25% have non-medical use and as many as 12% develop and OUD.[17] Further analysis of the trajectories containing a prescription event can be performed to determine if a specific dose, or duration of prescription correlates with a higher likelihood of subsequent misuse or use disorder following a prescription. This work is important as it may serve as a baseline for the development of a predictive tool to guide clinicians in the management of pain for specific patient populations and can guide important discussions on the risks and benefit of prescribing controlled substances for the treatment of pain. Such tools have the potential to be available to clinicians "at the bedside" through improved prescription drug monitoring programs available in most states.[18]

Our findings also demonstrate how the RR of a diagnosis can be affected by the available comorbid diagnoses and procedures for a patient. As shown in Figure 2b, in the case of bursitis, arthritis, and tendinitis, which have a RR of 4.3 alone, differences in clusters can vary the RR from between standard risk to an almost ten-fold increase. This ten-fold increase occurs when bursitis, arthritis, and tendinitis are added to a cluster containing blood disorders and opioid prescription. On the other hand, the inclusion of an opioid prescription alone correlates with a reduced risk in an adverse opioid event when compared to bursitis, arthritis, and tendinitis alone.

Limitations

The study does have several limitations. The inclusion of prescription and procedural codes in the analysis was justified by the ubiquity of opioid prescription in trajectories. The code appeared in almost 50% of all trajectories, more than any other code. To simplify analysis, we grouped drugs by category withing opioid medication and did not consider the specific drug or dosage frequency or amount. However, even with the inclusion of prescription and procedural codes, we recognize that certain aspects of an individual's medical history, including highly relevant factors such as non-prescribed drug use, are not available in the medical record for incorporation into our analysis. In addition to prescription and procedural information, other information can be added. Test results would be the most obvious addition from medical data, and insurance information or socioeconomic data would begin to bridge the gap and address the inclusion of social determinants of health in the patient's clinical narrative.[19] This would allow for the stratification of patients based on risk categorization during analysis, which could provide an impetus to investigate the factors behind the increased risks identified in our trajectories as having a potential causal effect. Finally, as stated previously, although risk factors are temporally preceding adverse events, this method of analysis does not prove a causal relationship.

CONCLUSION

We examine the temporal sequencing of diagnoses, procedures and prescriptions as risk factors leading to an adverse opioid event. Using a data driven approach, we show how large-scale healthcare records can be leveraged to extract risk factors for future research, inform guidelines for practitioner prescribing of opioids and importantly highlights the incidences where further assessments and services are needed to address the patient's overall health.

References:

1. *Opioid data Analysis and Resources*. 2020 19/03/2020; Available from: www.cdc.gov/drugoverdose/data/analysis.html.
2. Osborne, V., M. Serdarevic, H. Crooke, et al., *Non-medical opioid use in youth: Gender differences in risk factors and prevalence.* Addictive Behaviors, 2017. **72**: p. 114-119.
3. Inacio, M.C.S., C. Hansen, N.L. Pratt, et al., *Risk factors for persistent and new chronic opioid use in patients undergoing total hip arthroplasty: a retrospective cohort study.* BMJ Open, 2016. **6**(4): p. e010664.
4. Park, T.W., L.A. Lin, A. Hosanagar, et al., *Understanding Risk Factors for Opioid Overdose in Clinical Populations to Inform Treatment and Policy.* Journal of Addiction Medicine, 2016. **10**(6): p. 369-381.
5. Sullivan, M.D., M.J. Edlund, L. Zhang, et al., *Association Between Mental Health Disorders, Problem Drug Use, and Regular Prescription Opioid Use.* Archives of Internal Medicine, 2006. **166**(19): p. 2087-2093.
6. Inturrisi, C.E., *Clinical Pharmacology of Opioids for Pain.* The Clinical Journal of Pain, 2002. **18**(4): p. S3-S13.
7. Bohnert, A.S.B., M. Valenstein, M.J. Bair, et al., *Association Between Opioid Prescribing Patterns and Opioid Overdose-Related Deaths.* JAMA, 2011. **305**(13): p. 1315-1321.
8. Jensen, A.B., P.L. Moseley, T.I. Oprea, et al., *Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients.* Nature Communications, 2014. **5**(1): p. 4022.
9. Beck, M.K., D. Westergaard, A.B. Jensen, et al., *Temporal order of disease pairs affects subsequent disease trajectories: the case of diabetes and sleep apnea*, in *Biocomputing 2017*. 2016, WORLD SCIENTIFIC. p. 380-389.
10. Cytoscape.Beck, M.K., A.B. Jensen, A.B. Nielsen, et al., *Diagnosis trajectories of prior multi-morbidity predict sepsis mortality.* Scientific Reports, 2016. **6**(1): p. 36624.
11. Giannoula, A., A. Gutierrez-Sacristan, A. Bravo, et al., *Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study.* Sci Rep, 2018. **8**(1): p. 4216.
12. Hanauer, D.A. and N. Ramakrishnan, *Modeling temporal relationships in large scale clinical associations.* 2013(1527-974X (Electronic)).
13. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.
14. Bodenreider, O., *The unified medical language system (UMLS): integrating biomedical terminology.* Nucleic acids research, 2004. **32**(suppl_1): p. D267-D270.
15. Chang, D., I. Balazevic, C. Allen, et al., *Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings.* arXiv preprint arXiv:2006.13774, 2020.
16. Khansari, M., M. Sohrabi, and F. Zamani, *The Useage of Opioids and their Adverse Effects in Gastrointestinal Practice: A Review.* 2013(2008-5230 (Print)).
17. Vowles, K.E., M.L. McEntee, P.S. Julnes, et al., *Rates of opioid misuse, abuse, and addiction in chronic pain: a systematic review and data synthesis.* Pain, 2015. **156**(4): p. 569-576.
18. Fiellin, L.E. and D.A. Fiellin, *Toward better stewardship: Gaining control over controlled substances.* Annals of Internal Medicine, 2018. **168**(12): p. 883-884.
19. Braveman, P. and L. Gottlieb, *The Social Determinants of Health: It's Time to Consider the Causes of the Causes.* Public Health Reports, 2014. **129**(1_suppl2): p. 19-31.

Hypothyroidism – A Causal Approach to Testing Assumptions against Empirical Results

Rosemary Glavin, MS[1], Robert O. Ness, PhD[2], Andrew Nguyen, PhD[1]
[1]University of San Francisco, San Francisco, CA; [2]Northeastern University, Boston, MA.

Abstract

There is a controversy in the diagnosis and treatment of hypothyroidism. We propose the disagreement is fueled by statistical paradoxes, and sampling biases that provide different perspectives depending upon the sample selection criteria. The statistical inconsistencies become more apparent when viewed using a causal lens. Foundational hypothyroid research does not reflect the current Levothyroxine treated population. Exploration of empirical data demonstrates an apparent breakdown of the T4 to T3 causal pathway in the treated population. This use case demonstrates the difficulty of translating controlled research into clinical practices for patients with multiple comorbid conditions. We make the case for redundancy in data collection, ongoing attempts to falsify current assumptions and the need for causal approaches to validate the results of controlled research in clinical settings, in order to avoid confirmation bias from statistically insufficient biometrics.

Introduction

There is currently a controversy in the diagnosis and treatment of hypothyroidism, that centers around quantitative versus qualitative assessment factors.[1–6] Hypothyroidism is a condition that has been recognized and treated for over a century[7]. Levothyroxine (LT4) has been the standard of care treatment for hypothyroidism since the 1970s[8] and has become one of the top 5 most dispensed prescription medications in the US and the UK[9]. There is clear patient dissatisfaction with the standard of care and professional differences of opinion on clinical guidelines[4,6,10]. There are significant sampling differences between research populations and treated populations. These sampling differences introduce statistical paradoxes at the population level, which support an incomplete theoretical model. The data collection process only tracks variables that the research consensus and theoretical model deem relevant. This inhibits the ability to assess impact of treatment using quantitative factors in some subpopulations.

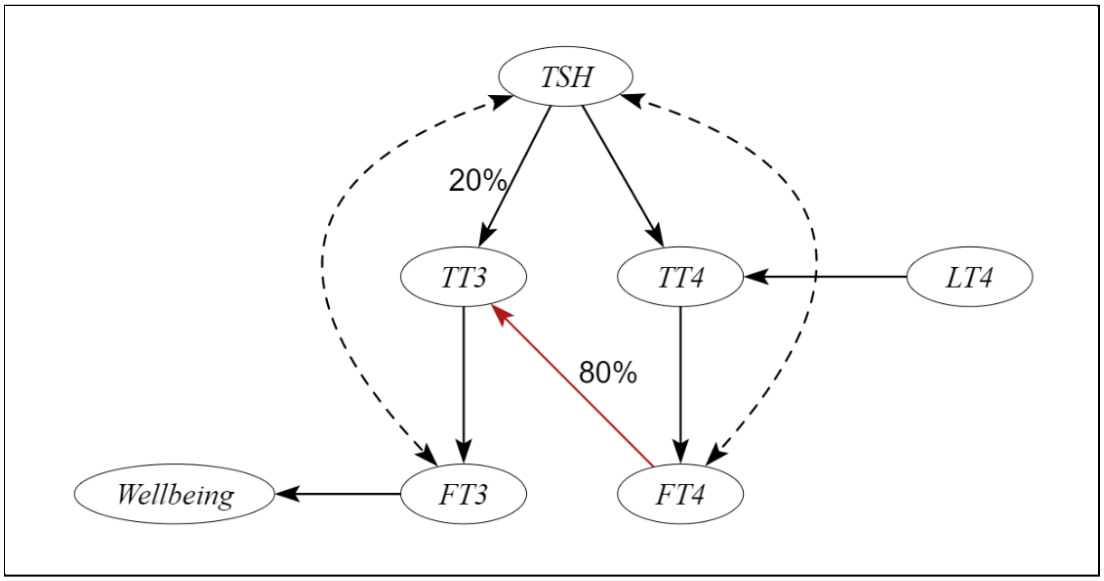


Figure 1. Simplified theoretical model of the thyroid system based upon the clinical guidelines.

In order to address this topic, consider a simplified model of hypothyroidism based on the clinical guidelines[1,2]. The thyroid produces two main hormones, Triiodothyronine (T3) and Thyroxine (T4). T3 is believed to be the bioactive end hormone, which regulates cellular metabolism throughout the body[11]. T4 is a prohormone – it is converted to T3 through a deiodination process in the peripheral tissues of the body[2,11,12]. Approximately 20% of circulating T3 comes directly from the thyroid gland and the remaining is converted from T4[13]. Peripheral conversion of T4 to T3 was confirmed in the 1970s[12], leading to changes in clinical practices. Treatment evolved from desiccated animal thyroid

tissue (NDT), containing both T3 and T4 to synthetic T4 monotherapy, partially because T3 is viewed as being more potent and thus riskier[1,2,8]. T3 and T4 hormones circulate both bound to protein and unbound or free and thus bioavailable. The thyroid is stimulated to produce T3 and T4 by a pituitary thyrotropin hormone, Thyroid Stimulating Hormone (TSH). T3 and T4 are involved in negative feedback loops with the hypothalamus and the pituitary to maintain homeostasis[1,2]. TSH is used as the primary biomarker to evaluate thyroid function.[1]

The theoretical model assumes T3 is the bioactive end hormone, yet T3 is not typically measured, because it is impacted by other complicating factors, including comorbid conditions[1,2]. However, the logic behind the intervention implicitly assumes that T3 will increase upon LT4 intervention. TSH and T4 are broadly used to diagnose and manage treatment[1]. However, there is research that shows significant disassociation between TSH and the other thyroid hormones, as well as system dynamics that are not well reflected in the current theoretical model[14–16]. Peterson et al.[16] reported 15–20% lower serum T3:T4 ratios in LT4 treatment, yet Braverman et al[12] and Jonklaas et al[17] report normal T3 levels were achieved with LT4 therapy. The assumption of adequate T4 to T3 conversion has implications for diagnosis, treatment and assessment of treatment effectiveness and therefore is a key issue in the controversy.

LT4 usage expanded as the theoretical model was extrapolated into predictions of benefit and harm, beyond the scope of the original research, often outside of the clinical guidelines which contain constraints to match the foundational, well-studied etiology of thyroid tissue damage[1,2,9,18–24]. Many patients take L-Thyroxine (LT4) for decades and once treatment is initiated, lifetime adherence is common[9]. Many of the patients taking LT4, have functioning thyroid glands, and comorbid conditions, often involving many daily medications. Many medications, as well as many health conditions impact thyroid hormone levels[2,25–28]. Usage of such medications have increased significantly in recent decades[29,30]. The evolution of medical intervention is fundamentally changing the LT4-treated population, while the clinical guidelines, written to implement best-practices based upon current assumptions, are (inadvertently) curating data to match the theoretical model and thereby obscuring limitations of the theoretical model when applied to a broader population.

In this research, the same large randomized dataset, the National Health and Nutrition Survey (NHANES)[31] referenced in the clinical guidelines for establishing the TSH reference range[2,32], is further analyzed using medication status to reflect the clinical presentation patterns of patients treated with LT4. The thyroid hormone profiles of heterogeneous subpopulations treated with LT4 (but often excluded from research) are presented and compared to LT4-only populations and unmedicated reference populations. This shows the gap between the empirical results and the theoretical model and demonstrates that the current clinical practices may instill a false sense that treatment is effective, while the underlying problem may be exacerbated.

Method

Continuous NHANES collects data from a randomized sample (including some up-sampling) of about 5000 people each year. Thyroid lab results were collected from NHANES participants for cycles years 2007-2008, 2009-2010 and 2011-2012. Thyroid lab results, age, sex, racial ethnic group (REG), and self-reported prescription medication usage within the previous 30 days were downloaded from NHANES. NHANES uses the Cerner Multum database to standardize and categorize medications. Drugs are classified into drug categories and subcategories. “Thyroid Agent” and “Antithyroid Agent” classifications exist. “Thyroid Agent” medications were further categorized into Triiodothyronine (T3) mediated, Thyroxine (T4) mediated, and natural desiccated thyroid (NDT) medications, using the specific drug names. The process used by NHANES to reconcile and categorize reported medications, as well as the questionnaire and laboratory protocols are described on the NHANES website.

The data were minimally processed to remove rows with empty fields, leaving 10465 observations with complete laboratory results for TSH, TT4, TT3, FT3 and FT4 and TPO. One of these observations contained a non-specific thyroid hormone, that could not be categorized as LT4, LT3, or NDT from the drug name. This observation was removed, leaving 10,464 observations, and 543 LT4 medicated observations. There were 3 LT4 medicated observations that also included LT3. These 3 observations were removed from the LT4 subpopulations used in comparisons.

Building upon the work of Peterson et al.[16] and Boucai et al.[32] where age, gender and racial ethic group (REG) were identified as independent predictors, subjects were stratified by reported medication status. Subpopulations of medicated and unmedicated were created based upon the NHANES field that holds the number of prescription medications reported. The medicated populations were further stratified into subjects not taking any thyroid medications (neither thyroid stimulating nor antithyroid medications), only LT4 medication, and taking any

medications (including thyroid hormones), those taking cardiovascular agents, antidiabetics, antihyperlipidemic agents, gastrointestinal agents, psychotherapeutics and sex hormones. In contrast to Peterson et al.[16] and Boucai et al.[32] no observations outside of the 4 mentioned above were removed.

White females were identified as the largest constituent population. Population parameters were created for the stratifications. One-sided t-tests, with 95% confidence intervals were used to compare mean of LT4–treated and not LT4-treated within the subpopulation. Additionally, 30 matched pair samples were drawn for each treatment (cardiovascular agents, antidiabetics, antihyperlipidemic agents, gastrointestinal agents, psychotherapeutics and analgesics) both with and without LT4 intervention. Levene's test was used to establish constant variance. ANOVA was used to compare population means of white females by medication status.

Graphical models were created, representing current assumptions about cause and effect between the thyroid hormones, based upon information taken from the clinical guidelines[1,2], the literature referenced therein[32] and literature dealing with associations between medical conditions and thyroid hormone levels[25,26]. Causal Inference theory[33] provides that these causal relationships imply a testable set of conditional independence relationships, between thyroid hormone measurements in the data. Statistical independence tests (using Fisher's Z transformation, distance covariance gamma, distance covariance permutation, HSIC gamma, HSIC permutation $p<0.05$) were applied to evaluate the degree to which the causal assumptions have empirical support.

R studio was used with the following libraries: Foreign, TidyVerse, bnlearn, ggplot, ggpubr, coreplot, pscl, gbm, pROC, caret, rtsne, qwrap2, fitdistribrplus, logspine.

The graphical models were created and tested in Causal Fusion, which is a software implementation of a causal inference engine[34].

Overall empirical population makeup and population parameters are combined with conditional independence testing as empirical evidence that contradict the foundational assumption that LT4 intervention through T4 conversion to T3 will normalize T3 levels to those of the control population mean[1,12,13,17].

Results

Statistical Independence Testing

Statistical independence testing took the relationship that TT3 is conditionally independent of LT4, given Age, FT4, TT4 and TSH as the null hypothesis. The test failed when tested against combined unmedicated white female subjects and those only treated with LT4 ($p=0$). As other medicated white female subjects were added to the population, this failure of conditional independence remained ($p< 0.05$). This finding suggests that there is a relationship between LT4 and TT3 not mediated through the assumed pathways even when other treatments are not involved.

LT4-Population Constituents

Approximately 5% (540/10464) of the sampled population report treatment with Levothyroxine (LT4). LT4 treated subjects generally have complex clinical presentations. Only 11% (60/540) of subjects taking LT4 medication are taking only LT4. This is significant when considering sampling bias and the confidence that can be attributed to thyroid research lacking heterogeneity. It is notable that patients with cardiovascular diseases and older patients are frequently excluded from thyroid research trials[32,35–37]. However, 60% (324/ 540) of LT4 treated patients take cardiovascular agents and 50.92% (275/ 540) are over 65 years old. Moreover, many of the LT4-treated population are treated for multiple comorbidities, with an average of 5 prescription medications. Exclusion of cardiovascular subjects also excludes subjects with other conditions, thus reducing the visibility of the treatment effect for these populations. This can introduce statistical inconsistencies.

Table 1. Common medications used by the 540 LT4 medicated subjects (out of the total sample of 10464)

Treatment	% of LT4 population	% of population	LT4 sample (/540)	Population sample (/10464)
LT4 medication	100	5.16	540	540
Only LT4 medication	11.11	0.57	60	60
Cardiovascular agents	60	26.74	324	2798
Anti-hyperlipidemic agents	40.56	16.1	219	1685
Gastrointestinal agents	25.56	10.28	138	1076
Psychotherapeutic agents	21.67	8.52	117	892
Analgesic agents	21.11	11.02	114	1153
Antidiabetic agents	19.26	9.12	104	954
Sex hormones	8.15	3.82	44	400

Hormone level profiles across LT4 treated populations reflect lower FT3 and TT3 than non-LT4 treated populations.

The standard of care[1] screens for primary hypothyroidism, presented clinically by high TSH (and low T4). The expected result of treatment with LT4 is suppressed TSH, relatively higher TT4 and FT4[1], with an implicit assumption that TT3 and FT3 should be relatively increased, ideally comparable to a healthy population. In order to minimize the opportunity of introducing a Simpson's paradox[33], white females (rather than the general population) were stratified by medication status and are presented in Table 2.

Row 4 of Table 2, (the LT4-only group, presumed to be diagnosed, primary hypothyroidism without comorbidity) shows the expected elevated TT4 and FT4, but elevated TSH, lower TT3 and lower FT3 than the unmedicated subpopulation. Row 3 of Table 2, shows TSH below 4.0 (a desirable range), but TT3 and FT3 remain relatively lower. Similar research, [38] and [39] report issues restoring T3 in athyreotic patients, but another[17] reports no significant change in T3 levels preoperatively and postoperatively for athyreotic patients. The mean FT3 in the Jonklaas et al.[17] study is higher than the mean results reported here (Table 2, rows 3 & 4) for LT4 treated patients. The Jonklaas et al. study[17] excludes participants with serious chronic disease and patients over 65 years. This difference between the research population and the broader treated population is a key issue being highlighted.

Boucai et al.[32] and Peterson et al.[16] have previously demonstrated that age is an important predictor of hormone levels. T-test comparison of TT3 in white females, and white females aged 60-70 years old with and without LT4 intervention, provide evidence that the mean TT3 is lower with LT4 intervention over and above the association of age and comorbid conditions with decreased T3 (Figures 3 & 4). ANOVA on matched pair random samples drawn from the white female population, blocked by common treatments in the population provide evidence of decreased mean T3 with LT4 intervention over and above the comorbid conditions (Figure 2). These results are consistent with significant failure of T4 to T3 conversion in the white female LT4-treated population.

Table 2. Within the white female population, mean FT3 and TT3 of LT4 treated populations are lower than non-LT4 treated populations.

	mean(sd) white females	Age years	TSH uIU/ml	FT3 pg/mL	FT4 ng/dL	TT3 ng/dL	TT4 ug/dL	Medications	Sample
1	Entire population	48.42 (20.92)	2.24 (3.80)	3.08 (0.41)	0.81 (0.18)	112.22 (25.64)	8.02 (1.65)	2.51 (2.96)	2241
2	Not medicated	37.23 (17.81)	2.19 (4.57)	3.19 (0.41)	0.78 (0.13)	115.54 (23.85)	7.70 (1.52)	0.00 (0.00)	692
3	Medicated with LT4 agents	62.95 (16.00)	2.66 (3.98)	2.83 (0.35)	0.98 (0.27)	95.61 (21.91)	9.21 (1.88)	5.01 (3.21)	285
4	LT4-only, exclude other medications	49.75 (20.41)	4.61 (8.64)	2.92 (0.36)	0.90 (0.23)	96.21 (21.00)	8.29 (2.03)	1.00 (0.00)	28
5	Medicated, exclude thyroid medications	51.21 (20.52)	2.18 (3.26)	3.08 (0.37)	0.78 (0.14)	113.76 (25.44)	7.95 (1.53)	3.30 (2.77)	1249

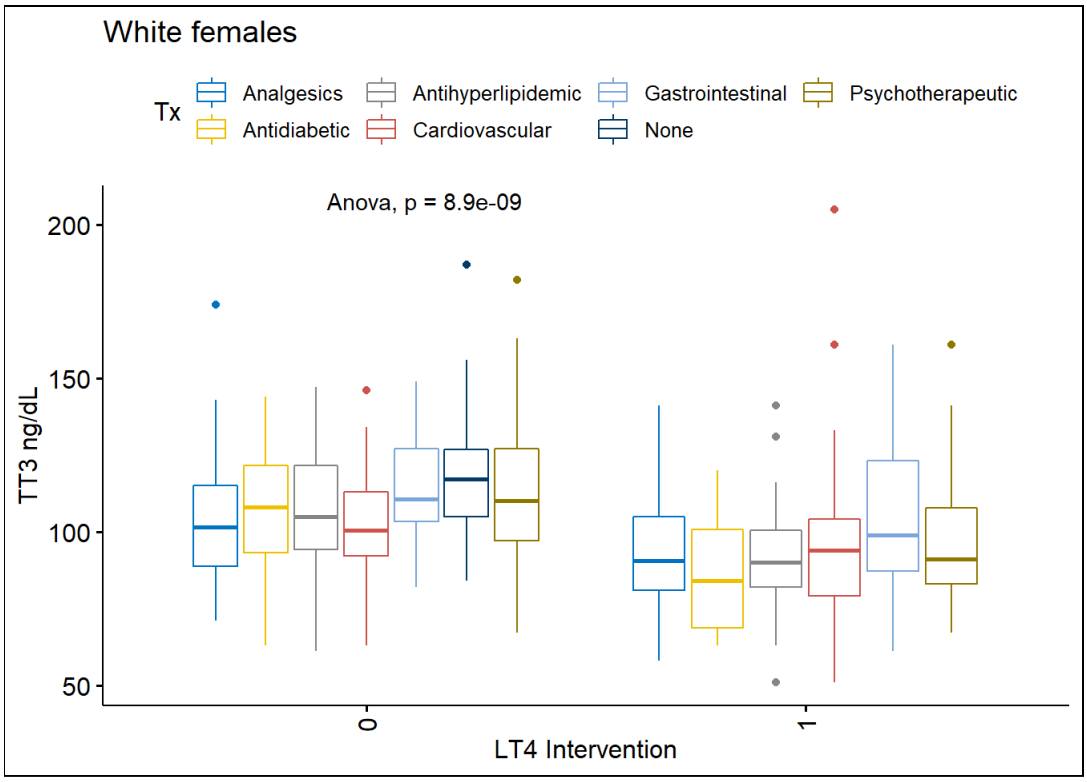


Figure 2. Mean TT3 levels are lower with LT4-intervention above within group variations of other treatments.

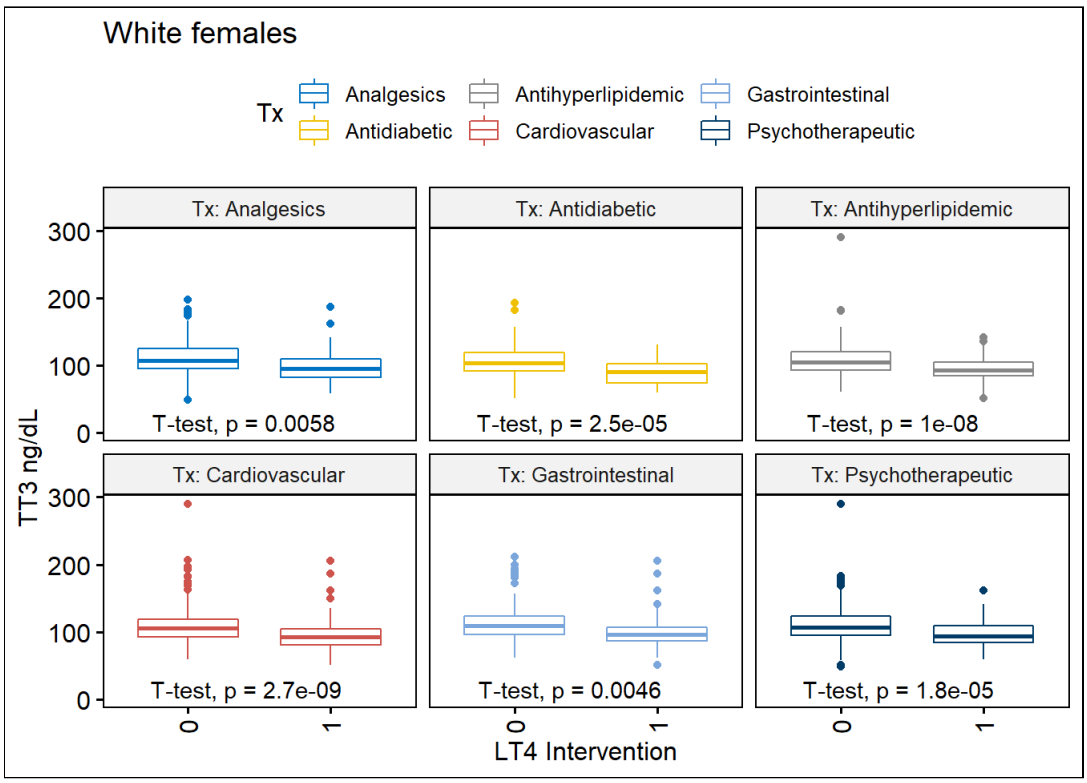


Figure 3. Mean TT3 levels are lower in LT4-treated white female population taking other medications.

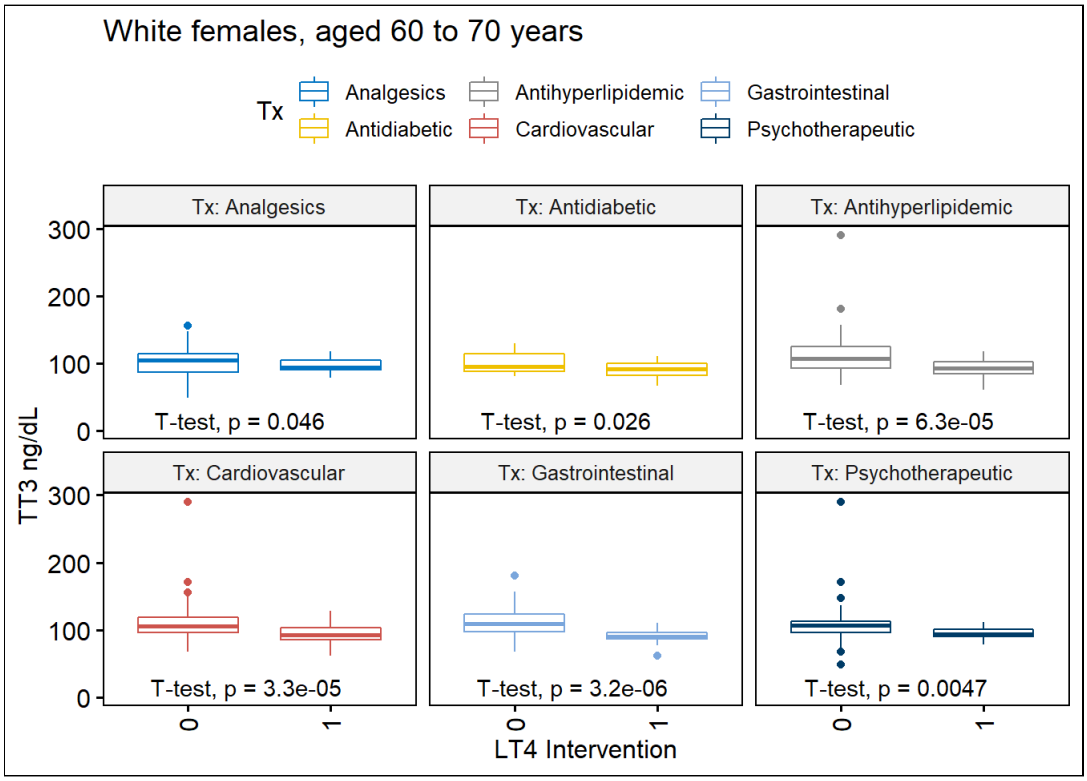


Figure 4. Mean TT3 levels are lower with LT4-intervention in white females aged 60 to 70 years.

Discussion

The widespread usage of LT4 intervention, makes an implicit assumption that the causal pathways of LT4 intervention in the treated population will match the causal pathways observed in the research populations. This assumption is not supported by the statistical evidence from empirical data. There is evidence that a breakdown in peripheral conversion of T4 to T3 is prevalent in some subpopulations. It is plausible that exogenous T4 suppresses TSH thus reducing serum T3 in subjects who did not initially present with elevated TSH. This may explain why these differences did not surface in the original research[8,12] nor in more recent research studying thyroid cancer[17]. Selection bias (through exclusion of cardiac patients and older patients) may also have minimized the impact of breakdown in T4 to T3 pathway in these studies. These studies serve very valuable purposes. However, the size and exclusion criteria of these studies must be considered as clinicians apply the results to a broader population.

LT4-treated white females have lower T3 than comparable populations (Figure 2). Current clinical practices of testing only TSH and T4 obscure this, both at a population and individual level. It is non-trivial to assess whether the hormonal profile presentation is true dysfunction or an adaptive mechanism and the difficulties of setting desirable ranges for FT3 and TT3 are significant[40]. However, establishing an individual's baseline thyroid hormone profile prior to intervention, allows patient and physician to validate the effectiveness of the intervention through a relative increase in TT3 and FT3. Once the intervention has been taken, it becomes very difficult to unravel cause and effect.

This research has limitations. The issue of tissue level deiodination[1] is simplified in the theoretical model (Figure 1) in an effort to make the material accessible. Preintervention and postintervention paired analysis as would be done in a Randomized Control Trial (RCT) is not possible. However, there are both practical and ethical limitations to what can be done in an RCT and expert opinion augments the evidence-based clinical decision process. The task force writing the clinical guidelines[1] made an exhaustive review of the literature and kept coming to the same point that the clinical significance of lower T3 was not known and the belief that there is broadly effective treatment available. The American Thyroid Association (ATA) has subsequently published findings that highlight patient dissatisfaction[4] and professional disagreement[10]. There is renewed interest in evaluating the clinical effectiveness of augmenting LT4 monotherapy with LT3[1,5]. It is possible that the same inconsistencies will exist unless measures are taken to make sure there are no sampling biases. Furthermore, the non-specific nature of thyroid symptoms coupled with complicated presentations of hypothyroid patients should elevate the importance of collecting all biometrics[1], yet the clinical guidelines suggest leaving it to expert opinion as to whether T3 should even be considered.

Selective exploration of data can create feedback loops that reinforce existing beliefs. Confirmation bias[41,42] has been well-studied, but in recent years a new data-driven polarization is beginning to surface online[43], whereby exposure to and curation of a subset of data endows an unjustified level of confidence in a position, resulting in polarization. The empirical data show a variety of thyroid hormone patterns (presumably reflecting different causal pathways), that observation of TSH and T4 do not capture. Practitioners do not see the gap between intervention and expected effects, which in turn creates more certainty. The complexity of the overall thyroid system, and the evolution of scientific knowledge about this system should temper confidence in existing research and current clinical practices, yet instead the clinical practices seem to reinforce existing beliefs.

It is difficult to translate controlled research into clinical guidelines for complex, dynamical patients which stress the research and ontological practices. The use case of hypothyroidism is especially elucidatory because of wide adoption of the LT4 intervention, the clinical complexity of hypothyroid subjects and the proposed extensive epidemiological applications of the theoretical model, well beyond the research focus of thyroid tissue damage, (and indeed well beyond the clinical guidelines put in place by clinical endocrinologists). The clinical decision process introduces additional legal and ethical dimensions, which are further complicated as patients more actively participate in the decision-making process. Bailao et al.[44] say nudging what is "beneficial" for the patient "can only be possible if we know with near certainty that the course of action we nudge is good for the patient". This issue of "near certainty" is important and is often lost or filtered out when we reduce scientific research built around 95% confidence intervals to binary conclusions.

Retrospective data science can help both validate and push the boundaries of current theories and ontologies, but only if paradigm independent data are collected. Complex clinical presentations with multiple comorbidities, involving long-term usage of multiple medications continue to become more common[30]. It is almost certain that clinical presentations will continue to have more heterogeneity than research studies. It is important that the limitations of the

original research are considered and balanced with empirical reports. Some amount of redundancy should be incorporated in data collection so that models can be falsified or verified on an ongoing basis[45,46].

The complex presentations of hypothyroid patients open many avenues of research opportunity involving causal pathways, confounding interactions, and unexpected system dynamics[40], as demonstrated by complications with amiodarone and the thyroid[47], the failure of LT4 intervention to prevent complications in pregnancy[18] and the association of psychological conditions with thyroid conditions[26,48]. These complexities are often filtered from research through exclusion criteria. While there is great value in such an approach to building knowledge, the output of this research needs to be integrated into clinical models with appropriate confidence levels, and not interpreted as immutable, binary conclusions. The non-specific nature of hypothyroid symptoms and discounting of qualitative subjective measures put patients at a disadvantage when they continue to report symptoms. Maximizing the biometric data available and ongoing efforts to understand where our models are limited or inconsistent, should help us avoid building clinical models that are not reflective of the patient population.

Conclusion

Measurement of TSH and T4 is not statistically sufficient to establish overall thyroid system function, given our knowledge of the T4 to T3 causal pathways. Collection of the overall thyroid profile preintervention and postintervention would provide a more complete mechanism to measure the effectiveness of the intervention and to identify potential points of failure, while providing additional data for retrospective analysis.

New models for how we approach complex, clinical presentations are needed. The existing tools, ontologies and mental models are further marginalizing the most vulnerable, chronically ill population with multiple comorbidities. We believe that Pearl's causal inference framework[33], further developed by Pearl and Bareinboim[34] to modify confidence to reflect the context associated with research, provide the technical tools and mental models to unravel some of these difficult issues.

Without data redundancy, paradigm independent data collection, and ongoing efforts to question assumptions and understand the limitations of formalized knowledge, there is a significant risk of creating feedback loops that reinforce existing beliefs which are contrary to empirical outcomes. Such feedback loops may be further amplified at scale with machine learning.

References

1. Jonklaas J, Bianco AC, Bauer AJ, Burman KD, Cappola AR, Celi FS, et al. Guidelines for the treatment of hypothyroidism: prepared by the american thyroid association task force on thyroid hormone replacement. Thyroid Off J Am Thyroid Assoc. 2014 Dec;24(12):1670–751.
2. Garber JR, Cobin RH, Gharib H, Hennessey JV, Klein I, Mechanick JI, et al. Clinical practice guidelines for hypothyroidism in adults: cosponsored by the American Association of Clinical Endocrinologists and the American Thyroid Association. Endocr Pract Off J Am Coll Endocrinol Am Assoc Clin Endocrinol. 2012 Dec;18(6):988–1028.
3. Medici M, Korevaar TIM, Visser WE, Visser TJ, Peeters RP. Thyroid function in pregnancy: what is normal? Clin Chem. 2015 May;61(5):704–13.
4. Peterson SJ, Cappola AR, Castro MR, Dayan CM, Farwell AP, Hennessey JV, et al. An online survey of hypothyroid patients demonstrates prominent dissatisfaction. Thyroid Off J Am Thyroid Assoc. 2018;28(6):707–21.
5. Leese GP. Nice guideline on thyroid disease: where does it take us with liothyronine? Thyroid Res. 2020;13:7.
6. Soldin OP. When thyroidologists agree to disagree: comments on the 2012 Endocrine Society pregnancy and thyroid disease clinical practice guideline. J Clin Endocrinol Metab. 2012 Aug;97(8):2632–5.
7. Murray GR. Note on the treatment of myxoedema by hypodermic injections of an extract of the thyroid gland of a sheep. Br Med J. 1891 Oct 10;2(1606):796–7.
8. Evered D, Young ET, Ormston BJ, Menzies R, Smith PA, Hall R. Treatment of hypothyroidism: a reappraisal of thyroxine therapy. Br Med J. 1973 Jul 21;3(5872):131–4.
9. Rodriguez-Gutierrez R, Maraka S, Ospina NS, Montori VM, Brito JP. Levothyroxine overuse: time for an about face? Lancet Diabetes Endocrinol. 2017;5(4):246–8.
10. Jonklaas J, Tefera E, Shara N. Physician choice of hypothyroidism therapy: Influence of patient characteristics. Thyroid Off J Am Thyroid Assoc. 2018;28(11):1416–24.
11. Brent GA. Mechanisms of thyroid hormone action. J Clin Invest. 2012 Sep;122(9):3035–43.

12. Braverman LE, Ingbar SH, Sterling K. Conversion of thyroxine (T4) to triiodothyronine (T3) in athyreotic human subjects. J Clin Invest. 1970 May;49(5):855–64.
13. Lum SM, Nicoloff JT, Spencer CA, Kaptein EM. Peripheral tissue mechanism for maintenance of serum triiodothyronine values in a thyroxine-deficient state in man. J Clin Invest. 1984 Feb;73(2):570–5.
14. Midgley JEM, Toft AD, Larisch R, Dietrich JW, Hoermann R. Time for a reassessment of the treatment of hypothyroidism. BMC Endocr Disord. 2019 Apr 18;19(1):37.
15. Hoermann R, Midgley JEM, Larisch R, Dietrich JW. Homeostatic control of the thyroid-pituitary axis: Perspectives for diagnosis and treatment. Front Endocrinol. 2015;6:177.
16. Peterson SJ, McAninch EA, Bianco AC. Is a normal TSH synonymous with “euthyroidism” in levothyroxine monotherapy? J Clin Endocrinol Metab. 2016;101(12):4964–73.
17. Jonklaas J, Davidson B, Bhagat S, Soldin SJ. Triiodothyronine levels in athyreotic individuals during levothyroxine therapy. JAMA. 2008 Feb 20;299(7):769–77.
18. Dhillon-Smith RK, Middleton LJ, Sunner KK, Cheed V, Baker K, Farrell-Carver S, et al. Levothyroxine in women with thyroid peroxidase antibodies before conception. N Engl J Med. 2019 04;380(14):1316–25.
19. Negro R. Levothyroxine before conception in women with thyroid antibodies: a step forward in the management of thyroid disease in pregnancy. Thyroid Res. 2019;12:5.
20. Ahn HS, Kim HJ, Welch HG. Korea’s thyroid-cancer “epidemic”--screening and overdiagnosis. N Engl J Med. 2014 Nov 6;371(19):1765–7.
21. Vanderpump MP, Tunbridge WM, French JM, Appleton D, Bates D, Clark F, et al. The incidence of thyroid disorders in the community: a twenty-year follow-up of the Whickham survey. Clin Endocrinol (Oxf). 1995 Jul;43(1):55–68.
22. Taylor PN, Albrecht D, Scholz A, Gutierrez-Buey G, Lazarus JH, Dayan CM, et al. Global epidemiology of hyperthyroidism and hypothyroidism. Nat Rev Endocrinol. 2018;14(5):301–16.
23. Villar HCCE, Saconato H, Valente O, Atallah AN. Thyroid hormone replacement for subclinical hypothyroidism. Cochrane Database Syst Rev. 2007 Jul 18;(3):CD003419.
24. De Groot L, Abalovich M, Alexander EK, Amino N, Barbour L, Cobin RH, et al. Management of thyroid dysfunction during pregnancy and postpartum: an Endocrine Society clinical practice guideline. J Clin Endocrinol Metab. 2012 Aug;97(8):2543–65.
25. Haugen BR. Drugs that suppress TSH or cause central hypothyroidism. Best Pract Res Clin Endocrinol Metab. 2009 Dec;23(6):793–800.
26. Dickerman AL, Barnhill JW. Abnormal thyroid function tests in psychiatric patients: a red herring? Am J Psychiatry. 2012 Feb;169(2):127–33.
27. Persani L. Clinical review: Central hypothyroidism: pathogenic, diagnostic, and therapeutic challenges. J Clin Endocrinol Metab. 2012 Sep;97(9):3068–78.
28. Vadiveloo T, Donnan PT, Murphy MJ, Leese GP. Age- and gender-specific TSH reference intervals in people with no obvious thyroid disease in Tayside, Scotland: the Thyroid Epidemiology, Audit, and Research Study (TEARS). J Clin Endocrinol Metab. 2013 Mar;98(3):1147–53.
29. Charlesworth CJ, Smit E, Lee DSH, Alramadhan F, Odden MC. Polypharmacy among adults aged 65 years and older in the United States: 1988-2010. J Gerontol A Biol Sci Med Sci. 2015 Aug;70(8):989–95.
30. Steinman MA. Polypharmacy-Time to get beyond numbers. JAMA Intern Med. 2016 Apr;176(4):482–3.
31. National Health and Nutrition Survey Data [Internet]. NHANES Data. 2020. Available from: https://wwwn.cdc.gov/nchs/nhanes/NhanesCitation.aspx
32. Boucai L, Hollowell JG, Surks MI. An approach for development of age-, gender-, and ethnicity-specific thyrotropin reference limits. Thyroid Off J Am Thyroid Assoc. 2011 Jan;21(1):5–11.
33. Pearl J. Causality: models, reasoning, and inference. Cambridge, U.K. ; New York: Cambridge University Press; 2000. 384 p.
34. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Natl Acad Sci [Internet]. 2016 Jul 5 [cited 2020 Aug 27];113(27):7345–52. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1510507113
35. Jonklaas J, Burman KD. Daily Administration of short-acting Liothyronine is associated with significant triiodothyronine excursions and fails to alter thyroid-responsive parameters. Thyroid Off J Am Thyroid Assoc. 2016;26(6):770–8.
36. González-Sagrado M, Martín-Gil FJ. Population-specific reference values for thyroid hormones on the Abbott ARCHITECT i2000 analyzer. Clin Chem Lab Med. 2004 May;42(5):540–2.
37. Jonklaas J, Nsouli-Maktabi H, Soldin SJ. Endogenous thyrotropin and triiodothyronine concentrations in individuals with thyroid cancer. Thyroid Off J Am Thyroid Assoc. 2008 Sep;18(9):943–52.

38. Woeber KA. Levothyroxine therapy and serum free thyroxine and free triiodothyronine concentrations. J Endocrinol Invest. 2002 Feb;25(2):106–9.
39. Gullo D, Latina A, Frasca F, Le Moli R, Pellegriti G, Vigneri R. Levothyroxine monotherapy cannot guarantee euthyroidism in all athyreotic patients. PloS One. 2011;6(8):e22552.
40. Dietrich JW, Landgrafe G, Fotiadou EH. TSH and thyrotropic agonists: Key actors in thyroid homeostasis. J Thyroid Res. 2012;2012:351864.
41. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. Nephron Clin Pract. 2010;115(2):c94-99.
42. Manchikanti L, Kaye AD, Boswell MV, Hirsch JA. Medical journal peer review: process and bias. Pain Physician. 2015 Feb;18(1):E1–14.
43. Del Vicario M, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W. Modeling confirmation bias and polarization. Sci Rep. 2017 11;7:40391.
44. Bailo L, Vergani L, Pravettoni G. Patient preferences as guidance for information framing in a medical shared decision-making approach: The bridge between nudging and patient preferences. Patient Prefer Adherence. 2019;13:2225–31.
45. Athey S, Chetty R, Imbens G, Kang H. Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. ArXiv160309326 Econ Stat [Internet]. 2020 Feb 29 [cited 2020 Aug 27]; Available from: http://arxiv.org/abs/1603.09326
46. VanderWeele TJ. Surrogate measures and consistent surrogates. Biometrics [Internet]. 2013 Sep [cited 2020 Aug 27];69(3):561–5. Available from: http://doi.wiley.com/10.1111/biom.12071
47. Hudzik B, Zubelewicz-Szkodzinska B. Amiodarone-related thyroid dysfunction. Intern Emerg Med. 2014 Dec;9(8):829–39.
48. Siegmann E-M, Müller HHO, Luecke C, Philipsen A, Kornhuber J, Grömer TW. Association of depression and anxiety disorders with autoimmune thyroiditis: A systematic review and meta-analysis. JAMA Psychiatry. 2018 01;75(6):577–84.

An Analysis of Two Sources of Cardiology Patient Data to Measure Medication Agreement

Rose C. Goueth, MS, Aaron M. Cohen, MD MS, Nicole G. Weiskopf, PhD
Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Abstract

Errors and incompleteness in electronic health record (EHR) medication lists can result in medical errors. To reduce errors in these medication lists, clinicians use patient self-reported data to reconcile EHR data. We assessed the agreement between patient self-reported medications and medications recorded in the EHR for six medication classes related to cardiovascular care and used logistic regression models to determine which patient-related factors were associated with the disagreement between these two information sources. From our 297 patients, we found self-reported medications had an overall above-average agreement with the EHR (κ = .727). We observed the highest agreement level for statins (κ = .831) and the lowest for other antihypertensives (κ = .465). Agreement was less likely for Hispanic and male patients. We also performed an in-depth error analysis of different types of disagreement beyond medication names, which revealed that the most frequent type of disagreement was mismatched dosages.

Introduction

The Kaiser Family Foundation estimates that around 3.8 million drugs were dispensed in 2018, and the CDC estimates that almost half of the United States population is on at least one prescription medication.[1, 2] With every medication that is prescribed and filled, there is a risk of an adverse drug event (ADE), which is defined as "any injuries resulting from medication use, including physical harm, mental harm, or loss of function." [3] ADEs may contribute to negative patient health outcomes, including emergency department visits, prolonged hospital stays, medication non-adherence and increases in hospital admissions overall.[4-7] ADEs also lead to increased costs and utilization of healthcare resources, which negatively impacts both patients and healthcare providers. Specifically, ADEs cause about 650,000 adults and children to visit the emergency department each year, with 27% of all adult emergency department visits leading to inpatient admittance.[8, 9] For hospitals, ADEs are costly in terms of time and money. The Institute of Medicine states the cost of ADEs for the healthcare industry is $3.5 billion per year.[10] Specifically, one study found on average, these events cost hospitals almost $3500 per patients and extend inpatient stays by three days.[11]

Previous research identifies some potential failure points in the healthcare delivery process that may lead to ADEs, including medication choice, prescription writing, formulation, medication dispensing, administration (by providers or patients), and therapy monitoring.[12, 13] Some of these errors are more likely to occur with incorrect or insufficient medication-related information in electronic health records (EHR). If a medication is not recorded on a patient's medication list, for example, a provider may be more likely to choose or write a prescription for another medication that results in overdosing or a drug-drug interaction. A medication list that is out-of-date may lead to the dispensing of drugs that should no longer be taken, as well as the self-administration by patients of medications that their providers don't want them to take. Medication lists issues also stem from the fragmented medical record system within the US, leading clinicians to depend on patient self-reported data for more accurate health histories.

Several approaches are used to mitigate the risk of ADEs stemming from inaccurate and incomplete medication lists. Direct patient-revision of medication lists during the pre-check-in process and pharmacist intervention have both been found to improve medication list completeness and accuracy.[14-16] The most popular approach is the process of medication reconciliation, "a formal process for creating the most complete and accurate list possible of a patient's current medications and comparing the list to those in the patient record or medication orders." [17] This process is a preventive measure for ADEs, allowing safe clinical decisions [18] and can be summarized as the collection, comparison, and usage of two lists: current medications according to the patient and prescribed medications.[19] Ideally, medication reconciliation results in a complete and accurate medication list, as well as an improvement in a patient's understanding of their medication regimen.

The widespread adoption of medication reconciliation and the other efforts described above have led to more accurate medication lists[19, 20] and higher patient literacy,[21] yet medication list data quality problems still exist.[22] Some factors that contribute to problems with the medication list have been identified. Infrequent medication counseling, for example, makes it less likely that a medication list will be up to date with respect to a patient's knowledge of their

current medications. Studies have also demonstrated that factors like hospital setting (i.e., primary care) and low patient comprehension result in decreased amounts of medication reconciliation completed inpatient encounters.[23] Conversely, decreasing medication complexity has been shown to improve reconciliation.[24]

Medication reconciliation has been explored in several contexts [23,25,26] with various data sources[19,20], but in most cases, medication reconciliation continues to be an in-person activity led and mediated by the provider. There has been promising work assessing the feasibility and usefulness of patients engaging more directly with the reconciliation process to improve the accuracy and completeness of medication lists, often utilizing patient portals and other forms of health information technology.[27-30] Further work is needed, though, to understand the medication-related and patient-related factors that are associated with the quality of patient self-report. In this study, we measured the agreement between patient-reported medication information and EHR medication list data and assessed what factors affect this agreement. We also conducted an in-depth error analysis to understand the types of disagreements that may occur and their frequency.

Methods

We collected patient-reported medication information using a REDCap survey and measured the agreement between these data and EHR medication list data. Specifically, we looked at six standard medication classes related to cardiovascular care: angiotensin-converting enzyme inhibitors (ACE inhibitors), angiotensin II receptor blockers (ARBs), antithrombotics, beta-blockers, statins, and included a sixth category: other antihypertensives (including diuretics, calcium channel blockers, etc.). Focus on cardiovascular diseases is a good contender for reconciliation analysis since these conditions affect almost half of American adults (48.0%),[31] with one in every four deaths being due to heart disease.[32]

We used Cohen's Kappa, F-measure, and descriptive statistics to evaluate and characterize agreement between the patient self-report of medication class and the EHR medication list. Further analyses with logistic regression models were employed to measure several patient-related factors' influence on the agreement level. Types of disagreements were further classified and assessed using descriptive statistics.

This study was conducted at Oregon Health & Science University (OHSU) in Portland, Oregon. OHSU is an academic medical center in Portland, Oregon, that includes two hospitals and multiple ambulatory care clinics. The primary care population at OHSU includes approximately 78,000 patients. Our data set consisted of two data sources: 1) patient self-report of several clinical concepts related to cardiovascular care and health via a REDCap survey and 2) medication list data and demographic data pulled from the EHR. This study was approved by OHSU's Institutional Review Board (#00017632).

Survey

Invitations to complete the REDCap[33] survey were sent via email to 1,700 eligible patients who met the following inclusion criteria: between 18 and 89 years of age, English as a preferred language, and had at least one outpatient visit with the OHSU Knight Cardiovascular Institute between 2/11/2017 and 2/12/2018.[18] All survey items were developed in close collaboration with cardiovascular clinicians. Within the survey, patients were asked whether they were currently taking medications in the six classes described above. Likert-like response items were used to allow patients to denote the confidence of their answer using a 5-point scale, ranging between "definitely not" and "definitely yes"). These responses were dichotomized: "maybe yes" or "definitely yes" were considered affirmative. Patients who

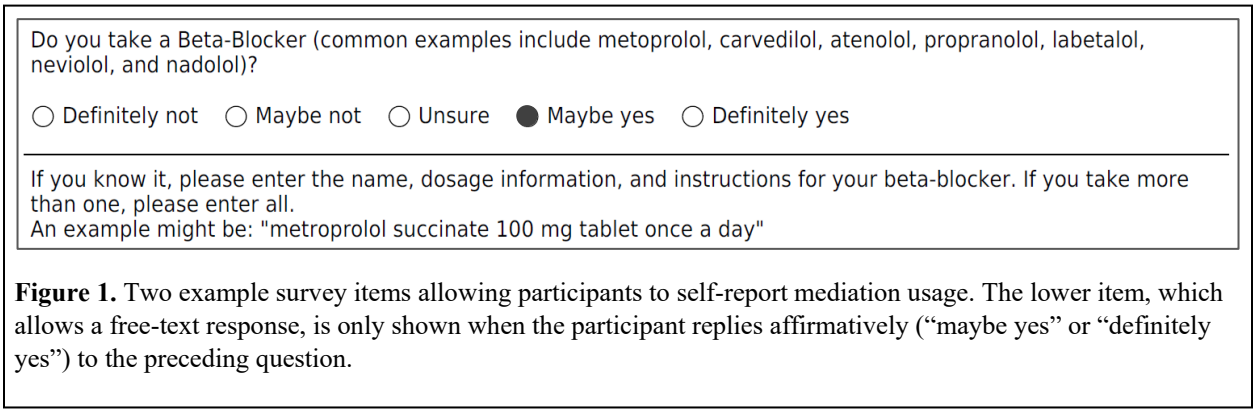


Figure 1. Two example survey items allowing participants to self-report mediation usage. The lower item, which allows a free-text response, is only shown when the participant replies affirmatively ("maybe yes" or "definitely yes") to the preceding question.

responded affirmatively to a medication question were then prompted to provide a free text response with more details about the mediation name and dosage. An example is shown in Figure 1.

EHR Data

Key demographic concepts (age, sex, race, and ethnicity) and active medication list data were queried from the OHSU Epic instance via the Integrated Care Coordination Information System (ICCIS), a population management system.[34] Medications in the six target classes were queried from the system using relevant value sets from the Value Set Authority Center.[35] The specific value sets used to extract medications within the six drug classes of interest were: Aspirin and Other Antiplatelets (2.16.840.1.113883.3.464.1003.196.12.1211), Low Intensity Statin Therapy (2.16.840.1.113762.1.4.1047.107), Moderate Intensity Statin Therapy (2.16.840.1.113762.1.4.1047.98), High Intensity Statin Therapy (2.16.840.1.113762.1.4.1047.97), ACE Inhibitor or ARB (2.16.840.1.113883.3.526.3.1139), Beta Blocker Therapy (2.16.840.1.113883.3.526.3.1174), Anti-Hypertensive Pharmacologic Therapy (2.16.840.1.113883.3.600.1476), and Diuretics (2.16.840.1.113883.3.666.5.829).

Analysis Methods

We used descriptive statistics and logistic regression to characterize medication agreement between patient self-report and the EHR medication list. For these analyses, we looked only at the structured responses about whether patients believed they were taking a medication class or not. Free text responses were then used to assess more granular agreement.

Agreement assessments: We employed several metrics to determine the agreement of self-reported medication class use with the EHR medication list. For assessing the overall agreement, we used Cohen's Kappa and F1 Scores. We also used Cochran's Q to determine whether the proportion of agreement between medication classes was equal. In these analyses, we only explored agreement in terms of the presence or absence of data with the EHR and patient survey. Likert responses were converted to a binary outcome to match the EHR data. Negative responses (Untrue, Maybe not, and Definitely not) were denoted with a zero (0) and affirmative responses (Maybe yes and Definitely yes) were denoted with a one (1). To determine which factors may contribute to differences in agreement, Cohen's Kappa was calculated for age, ethnicity, and sex. Age was converted into six age ranges for easier interpretation of the agreement results and to normalize the age distribution. Race was excluded due to its low sublevel counts.

Statistical modeling: To determine which factors contribute to the agreement of the two data sources, we employed logistic regression models, implemented using the R package stats.[36] Our outcome agreement was determined by the concordance of our two data sources. True positives and negatives became agreement (1), and false positives and negatives became disagreement (0). All regression models used three variables sex, age, and medCount. medCount is the total amount of medications found within the patient's medication list. We use this factor as a proxy measure for medical complexity. For our continuous variables, age, and medCount, we used the min-max scaling standardization.

True positive exploration: Because our agreement analysis only looks at the presence or absence of medication classes to calculate true positives, we conducted a further investigation of medication class true positives within our data with a free-text analysis of patient-reported medication details. We used a random sample of 80% of the EHR structured text for this analysis and matched these data to the patient survey free-text responses. Three main categories: active ingredient, drug name, and drug name and dosage, were used to discern where discrepancies within the true positives originated. We assumed patients would be less likely to know the exact dosage of their medications,[37] so dosage was only explored if the drug name was correctly identified.

Results

We received complete responses from 298 participants, for a response rate of 17.5%. One participant was omitted due to incomplete demographic data for a final sample of 297. Participants were 52.8% female, 94.6% white, 96.7% non-Hispanic, and had an average age of 60.4 years (range 19 – 87). At least one medication in the six classes of interest was self-reported by 80.9% of patients, and 85.9% had at least one present in the EHR, creating a 6.00% difference in the presence of information. The average number of medication classes reported by patients was 2.18 ± 0.171, and the average number, according to the EHR, was 2.67 ± 0.192.

Table 1. Demographics of 297 study participants. Data are presented as percentages unless otherwise stated.

Average Age (years/range)		60.8 (19 - 87)
Sex		
	Female	158 (52.8)
	Male	139 (46.5)
Race		
	Asian/Chinese/Japanese/Korean/Pacific Islander	4 (1.34)
	Black/African American	2 (0.67)
	White/Caucasian	283 (94.3)
	Unknown	9 (3.01)
Ethnicity		
	Hispanic	7 (2.34)
	Non-Hispanic	289 (96.3)
	Declined	2 (0.67)
Average Agreement between Data Sources		1.99 (±0.167)
Average Reported Medication Classes		2.18 (±0.171)
Average Medication Classes in EHR		2.67 (±0.192)

Table 2 shows the presence of medication classes in each data source. Antithrombotics were the most reported medication class for patients and the EHR and ARBs were the least reported in both sources.

Table 2. Percentages of patients on each medication class according to patient self-report and EHR, as well as the agreement between the two. Percentages are out of n=297.

Med Class	**Survey**	**EHR**	**Kappa (95% CI)**	**F1 Score**
ACE Inhibitors	22.5%	29.9%	0.724 (0.633, 0.815)	0.848
Antithrombotics	62.1%	60.1%	0.717 (0.632, 0.801)	0.903
ARBs	14.8%	20.8%	0.726 (0.621, 0.832)	0.774
Beta-blockers	37.9%	48.3%	0.735 (0.658, 0.813)	0.848
Other Antihypertensives	29.2%	50.7%	0.465 (0.364, 0.566)	0.664
Statins	50.3%	56.7%	0.831 (0.768, 0.895)	0.922

Agreement Trends

We assessed medication class agreement with two statistics, Cohen's Kappa and F1 Scores, which are summarized in Table 2. Figure 2 shows the huge proportion of true negatives within most medication classes, the absence of the medication class in the survey and EHR, so we also decided to calculate the F1 scores to see the relationship between agreement and discordance. Overall, the two sources had an agreement of 0.727 (95% CI: [0.695, 0.759]), an above-average score.

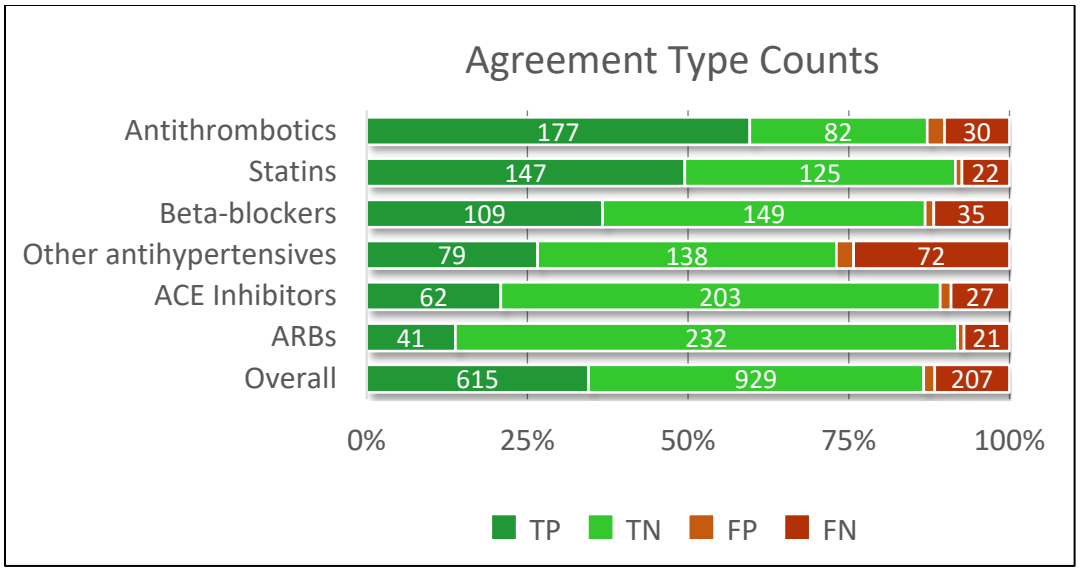


Figure 2. Counts of agreement types between the two data sources

Other antihypertensives had the lowest Kappa (0.465, 95% CI: [0.364, 0.566]) and F1 Scores (0.664), while statins had the highest Kappa (0.831, 95% CI: [0.768, 0.895]) and F1 scores (0.922). The difference in trends was not explained by the number of true negatives present, with ARBs having the highest amount.

Differences in agreement are seen with demographic factors, with age and sex having the strongest association. When we divided age into age ranges, we saw varied differences in Kappa scores. The other category had the lowest agreement for four out of six age ranges and statins had the highest agreement for three age ranges (40's, 50's and 60's). For sex, we saw females had a higher average agreement ($\kappa = 0.705$, 95% CI: [0.657, 0.753]) than males ($\kappa = 0.686$, 95% CI: [0.636, 0.735]). In ethnicity, we saw higher average agreement in non-Hispanics ($\kappa = 0.732$, 95% CI: [0.611, 0.854]) than Hispanics ($\kappa = 0.706$, 95% CI: [0.671, 0.740]). Kappa scores were not computed for race due to insufficient sample sizes.

Statistical models

We used logistic regression models to determine the influence of different patient-level factors on the agreement between self-report and the EHR medication list. A Cochran's Q test confirmed that agreement differed across medication classes, so we ran a separate model for each medication class. We ran the same model for each medication class with agreement as the dependent variable and sex, age, and medCount as the independent variables. Table 3 shows the results of the regression analyses. Sex was the most influential factor for antithrombotics and ACE-inhibitors, both being significant. Even though both medication classes see a significant effect, the size of the effect is different. Men on ACE-inhibitors are .433 times less likely to see agreement than women, while men on antithrombotics are 2.70 times more likely to see agreement than women. Age was the most influential factor for angiotensin receptor blockers and statins but was only significant for the former. For age, a coefficient less than one indicates a certain percent decrease in the outcome (agreement between the two data sources). For every one year increase in age, there is a 5.50% decrease in the odds that patients who reported being on an ARBs will agree with their EHR record. Medication count was the most influential factor for beta-blockers and other antihypertensives but was only significant for the latter. For each medication added to a patient's total medication list, patients who reported taking an other antihypertensive show a 5.90% decrease in the odds that their medication will match with the EHR.

Table 3. Medication class regression analysis results.

Med Class	Sex(M)	Age	medCount
ACEs	0.434*	0.982	0.966
Antithrombotics	2.70*	0.986	1.03
ARBs	0.984	0.945**	0.943
Betas	0.886	0.992	0.954
Other	0.584	0.989	0.941**
Statins	0.646	0.977	1.03

** Significant at 0.05 ** Significant at 0.001, results are exponentiated and unscaled*

True Positive Exploration

Structured responses to the medication class questions that were identified as true positives within our random sample were manually reviewed in order to validate patient responses. Table 4 shows the proportions of patients who correctly described the components of the medications reported. Active ingredient refers to the primary biologically active ingredient in a medication, and drug name refers to the full name of the medication (brand names and generic names were treated equivalently), which can include the main ingredient. If patients correctly identified the correct drug name, we also looked at the dosage. The total amount of matched medications per class denotes the amount of true positives patients correctly identified the active ingredient over half of the time (58.4%) and identified the proper medication name less than half of the time (48.7%). Drug name and dosage information was correctly identified 21.4% of the time. We also looked for spelling errors and found that most people spelled their medications correctly (90.6%).

Table 4. An in-depth analysis of errors found within the patient survey data*

Med Class	Active Ingredient	Drug Name	Drug Name and Dosage	N
ACE Inhibitors	53.1%	32.7%	18.4%	20
Antithrombotics	84.2%	84.2%	26.3%	110
ARBs	85.7%	78.6%	14.3%	15
Beta-blockers	35.7%	31.0%	21.4%	50
Other Antihypertensives	69.7%	54.1%	22.9%	43
Statins	50.7%	50.7%	23.2%	70
Total	58.4%	48.7%	21.4%	**308**

**Percent Correct Compared to Structured EHR data*

Discussion

This study explored the factors that influence the agreement between the EHR medication list and patient-reported medications for five standard medication classes within the cardiology specialty. The more medications were present in the EHR (85.9%) than were self-reported by patients (80.9%). Within medication classes, we only saw the equivalent presence of medications in antithrombotics. Kappa and F1 scores revealed varying degrees of agreement within demographic factors and medication classes. This variation is partially explained by errors examined within patient-reported data, mostly accounted for by dosage errors.

Overall, the data showed an above-average agreement with medication classes having no apparent pattern. Demographic variables also revealed similar results with sex, ethnicity, and age. Using age ranges for our Kappa analysis, we saw the other category having the lowest agreement three times, with the lowest agreement class for people in their 40s and 50s being ARBs. As a patient's age increased, we saw fluctuating decreases in agreement, where the lowest average Kappa score belonged to patients above 80 ($\kappa = 0.635$, 95% CI: [0.492, 0.778]).

Our regression analyses continued to show no clear pattern in our data but showed significance for some variables. The other category had the smallest change in agreement for every added medication count but was the only one to be significant. The small change could be explained by the within-category variation happening, meaning many of the medications found in this category could be uncommon treatments for cardiac diseases. ARBs had the coefficient with the largest effect (2.70), but the large coefficient can partially be explained by having the greatest amount of true negatives (n = 232).

The true positive validation partially explains the agreement variability in our study. More patients were likely to know the active ingredient of their medication (58.4%) than the drug name (48.7%), yet most of these medications were spelled correctly (90.6%). Patients on beta-blockers were least likely to identify the correct active ingredient (35.7%) or drug name (31.0%). Most errors stemmed from self-reported medications with correct names and incorrect dosages (21.4%).

Learning about medication reconciliation with live patient data was crucial to illustrate an accurate representation of the completeness of each data source. Patient surveys are an important data source that allows patients to reflect on their current health history and improves the accuracy of EHR records. This study showed patient surveys provide the opportunity for EHR records to close the 6.00% difference of information and improve medication lists to ensure their accuracy for long term patient-centered care. Repetition and efficiency of these surveys have the potential to increase the presence and accuracy of medications found in both sources above a .727 agreement rate, which is considered moderate agreement,[38] to reach a strong agreement (> .80) for this clinically essential data type. Improvement in this process may have huge health, cost, and time benefits for the patient, provider, and healthcare institution.

Limitations and Future Work

There were several limitations of this study, the largest being its generalizability. The study used a racially and ethnically homogeneous set of patients within a specialty (cardiology). Another limitation came from our small sample size. A larger sample would have allowed us to include and analyze factors that could contribute to agreement through more complex regressions and more robust machine learning algorithms. The temporality of medication list data causes another limitation as it cannot ensure complete accuracy of EHR data. To mitigate this limitation, we pulled EHR data to match the survey completion dates to represent the most current data. There was also a limitation in using one EHR record as it limits the medication list to a patient's memory at any one visit and ones prescribed at the institution. One EHR record does not reflect the totality of a patient's health history, so future work would seek multiple data sources to provide a comprehensive picture of a patient's current medications from numerous healthcare institutions' EHR data, claims data, and pharmacy records of dispensed prescriptions.

Conclusion

In this paper, we sought to identify factors that influence medication reconciliation between our two data sources. We analyzed clinical and demographic factors to observe variability in agreement. Preliminary analysis displayed differences, but no one pattern described the trends observed. Our regressions analysis found that most of the differences were not significant, but some significant findings (i.e., the other antihypertensives category) did point to the usefulness of patient-reported data. Our patient survey data provided the opportunity to validate the true positives within our dataset and explore how patient inaccuracies can affect agreement. In the future, we plan to assess additional sources of medication data to analyze new factors contributing to agreement, ultimately proving why we must improve medication reconciliation processes for more accurate health histories within the EHR.

References

1. Number of Retail Prescription Drugs Filled at Pharmacies by Payer 2019 [Available from: https://www.myendnoteweb.com/EndNoteWeb.html?func=downloadInstallers.]
2. Therapeutic Drug Use [updated January 19, 2017. Available from: https://www.cdc.gov/nchs/fastats/drug-use-therapeutic.htm.]
3. Adverse Drug Events Office of Disease Prevention and Health Promotion [Available from: https://health.gov/hcq/ade.asp.]
4. Barnsteiner JH. Medication Reconciliation. In: Hughes RG, editor. Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Advances in Patient Safety. Rockville (MD)2008.
5. Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events. Implications

for prevention. ADE Prevention Study Group. JAMA. 1995;274(1):29-34.
6. Leguelinel-Blache G, Dubois F, Bouvet S, Roux-Marson C, Arnaud F, Castelli C, et al. Improving Patient's Primary Medication Adherence: The Value of Pharmaceutical Counseling. Medicine (Baltimore). 2015;94(41):e1805.
7. Sentinel Event Alert 2006 [updated February 9, 2006. 35:[
8. Adverse Drug Events in Children [updated 2019-10-10T07:11:59Z/. Available from: https://www.cdc.gov/medicationsafety/parents_childrenadversedrugevents.html.]
9. Adverse Drug Events in Adults [updated October 11, 2017. Available from: https://www.cdc.gov/medicationsafety/adult_adversedrugevents.html.]
10. Aspden P, Wolcott JA, Bootman JL, Cronenwett LR. Preventing medication errors: National Acad. Press; 2007.
11. Hug BL, Keohane C, Seger DL, Yoon C, Bates DW. The costs of adverse drug events in community hospitals. Jt Comm J Qual Patient Saf. 2012;38(3):120-6.
12. Aronson JK. Medication errors: what they are, how they happen, and how to avoid them. QJM. 2009;102(8):513-21.
13. Aronson JK. Medication errors: definitions and classification. Br J Clin Pharmacol. 2009;67(6):599-604.
14. Balon J, Thomas SA. Comparison of hospital admission medication lists with primary care physician and outpatient pharmacy lists. J Nurs Scholarsh. 2011;43(3):292-300.
15. Lindquist LA, Gleason KM, McDaniel MR, Doeksen A, Liss D. Teaching medication reconciliation through simulation: a patient safety initiative for second year medical students. J Gen Intern Med. 2008;23(7):998-1001.
16. Staroselsky M, Volk LA, Tsurikova R, Newmark LP, Lippincott M, Litvak I, et al. An effort to improve electronic health record medication list accuracy between visits: patients' and physicians' response. Int J Med Inform. 2008;77(3):153-60.
17. Persell SD, Osborn CY, Richard R, Skripkauskas S, Wolf MS. Limited health literacy is a barrier to medication reconciliation in ambulatory care. J Gen Intern Med. 2007;22(11):1523-6.
18. Weiskopf NG, Cohen AM, Hannan J, Jarmon T, Dorr DA. Towards augmenting structured EHR data: a comparison of manual chart review and patient self-report. AMIA 2019 Annual Symposium Proceedings. 2019.
19. Nassaralla CL, Naessens JM, Chaudhry R, Hansen MA, Scheitel SM. Implementation of a medication reconciliation process in an ambulatory internal medicine clinic. Qual Saf Health Care. 2007;16(2):90-4.
20. Varkey P, Cunningham J, O'Meara J, Bonacci R, Desai N, Sheeler R. Multidisciplinary approach to inpatient medication reconciliation in an academic setting. American Journal of Health-System Pharmacy. 2007;64(8):850-4.
21. Salanitro AH, Osborn CY, Schnipper JL, Roumie CL, Labonville S, Johnson DC, et al. Effect of Patient- and Medication-Related Factors on Inpatient Medication Reconciliation Errors. Journal of General Internal Medicine. 2012;27(8):924-32.
22. Walsh KE, Marsolo KA, Davis C, Todd T, Martineau B, Arbaugh C, et al. Accuracy of the medication list in the electronic health record—implications for care, research, and improvement. Journal of the American Medical Informatics Association. 2018;25(7):909-12.
23. Kaboli PJ, McClimon BJ, Hoth AB, Barnett MJ. Assessing the accuracy of computerized medication histories. Am J Manag Care. 2004;10(11 Pt 2):872-7.
24. Patel CH, Zimmerman KM, Fonda JR, Linsky A. Medication Complexity, Medication Number, and Their Relationships to Medication Discrepancies. Ann Pharmacother. 2016;50(7):534-40.
25. Gardella JE, Cardwell TB, Nnadi M. Improving medication safety with accurate preadmission medication lists and postdischarge education. Jt Comm J Qual Patient Saf. 2012;38(10):452-8.
26. Keogh C, Kachalia A, Fiumara K, Goulart D, Coblyn J, Desai SP. Ambulatory Medication Reconciliation: Using a Collaborative Approach to Process Improvement at an Academic Medical Center. The Joint Commission Journal on Quality and Patient Safety. 2016;42(4):186-AP2.
27. Staroselsky M, Volk LA, Tsurikova R, Newmark LP, Lippincott M, Litvak I, et al. An effort to improve electronic health record medication list accuracy between visits: Patients' and physicians' response. International Journal of Medical Informatics. 2008;77(3):153-60.
28. Chae SY, Chae MH, Isaacson N, James TS. The Patient Medication List: Can We Get Patients More Involved in Their Medical Care? The Journal of the American Board of Family Medicine. 2009;22(6):677-85.
29. Heyworth L, Paquin AM, Clark J, Kamenker V, Stewart M, Martin T, et al. Engaging patients in medication reconciliation via a patient portal following hospital discharge. Journal of the American Medical Informatics Association. 2013;21(e1):e157-e62.

30. Prey JE, Polubriaginof F, Grossman LV, Masterson Creber R, Tsapepas D, Perotte R, et al. Engaging hospital patients in the medication reconciliation process using tablet computers. Journal of the American Medical Informatics Association : JAMIA. 2018;25(11):1460-9.
31. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics -- 2020 Update: A Report From the American Heart Association. Circulation. 2020;141(9):e139-e596.
32. Fryar CD, Chen T-C, Li X. Prevalence of Uncontrolled Risk Factors for Cardiovascular Disease: United States, 1999–2010. National Center for Health Statistics: NCHS Data Brief; 2012 08/2012.
33. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377-81.
34. Dorr DA, Wilcox A, Burns L, Brunker CP, Narus SP, Clayton PD. Implementing a multidisease chronic care model in primary care using people and technology. Dis Manag. 2006;9(1):1-15.
35. Value Set Authority Center United States National Library of Medicine [updated 07/29/2020. Available from: https://vsac.nlm.nih.gov/.]
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria. 2020. 2020.
37. Wolf MS, Davis TC, Shrank W, Rapp DN, Bass PF, Connor UM, et al. To err is human: Patient misinterpretations of prescription drug label instructions. Patient Education and Counseling. 2007;67(3):293-300.
38. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82.

Extracting Concepts for Precision Oncology from the Biomedical Literature

Nicholas Greenspan,[1]* Yuqi Si, MS,[2] Kirk Roberts, PhD[2]
[1]Department of Computer Science, Columbia University
New York City NY, USA
[2]School of Biomedical Informatics,
The University of Texas Health Science Center at Houston
Houston TX, USA

Abstract

This paper describes an initial dataset and automatic natural language processing (NLP) method for extracting concepts related to precision oncology from biomedical research articles. We extract five concept types: CANCER, MUTATION, POPULATION, TREATMENT, OUTCOME. *A corpus of 250 biomedical abstracts were annotated with these concepts following standard double-annotation procedures. We then experiment with BERT-based models for concept extraction. The best-performing model achieved a precision of 63.8%, a recall of 71.9%, and an F1 of 67.1. Finally, we propose additional directions for research for improving extraction performance and utilizing the NLP system in downstream precision oncology applications.*

1 Introduction

Precision medicine is a paradigm in which treatment decisions are based not just on a patient's disease status, but on a variety of other factors including specific genetic, environmental, and other factors.[1] The preeminent use case for precision medicine thus far has been cancer, i.e. precision oncology. Precision oncology is a rapidly-developing field[2], with a growing number of treatments, trials, and genomic markers. Since drugs can be targeted to relatively rare mutations, the number of studied treatments is greatly expanded[3,4] and these can be referred to by a variety of names (e.g., the name used in pre-clinical trials is often different than the final drug name). Since the gene mutations can be relatively rare, clinical trial structures have had to be altered to better fit the precision medicine paradigm.[5] And, critically, there are thousands of known genetic mutations from hundreds of cancer-related genes.[6] Sizable effort is thus required to curate all of these types of information to make them available in a usable form to both researchers and clinicians.

Our prior work has focused on this problem from an information retrieval (IR) perspective: how does one find patient-specific information (given a type of cancer, mutation, etc.) from the vast trove of precision medicine-related publications. IR systems were evaluated for this task in the TREC Precision Medicine tracks.[7,8,9] We also developed PRIMROSE[10], a search engine that implements many of the best aspects of precision oncology search. A consistent weakness in these IR approaches, however, was difficulty dealing with the complex semantics of precision oncology articles: identifying the exact treatments studied in an article, which types of cancer the treatment applies to, etc. This task is more consistent with a natural language processing (NLP) information extraction (IE) approach. Therefore, in this work we report the initial development of an NLP system for extracting five key elements of biomedical articles for precision oncology: the type(s) of cancer studied, the mutations that were targeted, the specific population it is limited to, the treatment evaluated, and any available outcome information summarized in the abstract. Because of the fast-moving nature of the field, we focus on biomedical abstracts instead of full-text articles. Not only are the abstracts publicly available well before the full text, but many of the latest-breaking developments in precision oncology are presented at talks in major oncology conferences and only the abstracts for these talks are provided.

To gauge the complexity of this NLP task, we collected a pilot corpus of 250 biomedical abstracts drawn from the TREC Precision Medicine dataset. The five concept types–CANCER, MUTATION, POPULATION, TREATMENT, and OUTCOME–were double-annotated and reconciled. Two models based on BERT[11], and specifically the BioBERT[12] model pre-trained on biomedical text, were evaluated: $\text{BioBERT}_{\text{BASE}}$ and $\text{BioBERT}_{\text{LARGE}}$. The difference between these models is the number of parameters, in terms of number of layers, hidden units, and attention heads.

*This project was undertaken during an undergraduate internship at UTHealth-SBMI.

The remainder of this paper is organized as follows. Section 2 discusses related work in NLP for cancer and precision medicine. Section 3 describes the methods, including data (§3.1), annotation (§3.2), and automatic concept extraction (§3.3). Section 4 details the results. Section 5 provides a discussion, including an error analysis, implications, and directions for future work. Finally, Section 6 concludes the paper.

2 Related Work

Biomedical Literature NLP for Cancer Cancer is one of the more frequently studied aspects of NLP for biomedical literature articles. Early works such as MedScan[13] employed rule-based systems to extract and interpret information from MEDLINE abstracts. Chun et al.[14] developed a corpus and extracted relations between prostate cancer and genes from abstracts using a maximum entropy classifier. Baker et al. developed a corpus for identifying the hallmarks of cancer from the biomedical literature and proposed a support vector machine (SVM) model[15] and later a convolutional neural network[16] to automatically classify abstracts. A different take on cancer NLP for the biomedical literature is the development of literature-based discovery (LBD) tools such as LION LBD[17] to identify implicit links within the network of literature articles. LION in particular focuses on the molecular biology of cancer. Beyond the biomedical literature, a tremendous amount of NLP research has been conducted for cancer on other data types. Most notable among these are electronic health records, for which several review articles exist that overview cancer NLP for clinical notes.[18,19,20]

Biomedical Literature NLP for Genomics A tremendous amount of NLP work has focused on extracting information related to genomics from the literature. Early work includes EDGAR[21] , which identified gene-drug relations from biomedical abstracts. Libbus et al.[22] identified genes from MEDLINE abstracts based on the Gene Ontology[23] for the purpose of linking literature-based data to structured knowledge sources. Work in pharmacogenomics has required extensive use of NLP to build resources such as the use of SemRep[24] or the construction of the pharmacogenomics knowledge base PharmGKB.[25,26,27] In turn, PharmGKB has been utilized as a knowledge base for many further NLP studies.[28,29,30] Similarly, the PGxCorpus[31] is a manually-annotated corpus for pharmacogenomics–similar in many ways to our goal here, but their work is not specific to cancer. Finally, more general biomedical literature NLP has included genomic components, particularly the CRAFT corpus.[32,33]

Biomedical Literature NLP for Precision Oncology There has indeed been some work specific to precision medicine for NLP within the space of the current work. For instance, Deng et al.[34] classifies abstracts with an SVM based on whether they focus on cancer penetrance. Bao et al.[35] extends this with a deep learning model. Instead of extracting the particular concepts, however, these works focus is simply to classify the entire abstract for use in downstream meta-analyses. Next, Hughes et al.[36] reviews how to utilize precision oncology NLP specific for breast cancer. Finally, the TREC Precision Medicine track[7,8,9] is an ongoing information retrieval shared task focusing on identifying articles relevant to precision oncology. This has inspired the creation of many search engines, including our own,[10] for clinical decision support in precision oncology. Of the many search engines to participate in the TREC Precision Medicine track, however, none has successfully integrated biomedical knowledge sources to greatly improve retrieval performance. We believe this is partly due to the fact that it is difficult to properly link the key aspects of precision oncology in an abstract to these powerful knowledge bases. Instead, most use of biomedical knowledge in such search engines is simply to expand synonyms (e.g., through query expansion) which gives at most small boosts to retrieval performance. Our goal in this paper, then, is to lay the groundwork for improvements in precision oncology search and knowledge acquisition by identifying the key elements to precision oncology in biomedical abstracts. This will allow for the downstream linking of these articles with existing biomedical knowledge bases for better semantic comprehension of the precision oncology scientific landscape.

3 Methods

The high-level study design for this paper follows the standard supervised NLP pipeline: data identification (Section 3.1), manual data annotation (Section 3.2), and automatic NLP extraction (Section 3.3). Since this is a pilot study, our primary goal has been to identify the key barriers to large-scale system development, which is discussed in more detail in the Discussion (Section 5).

3.1 Data

Since the latest developments of precision oncology research are only publicly available in abstracts, we focus only on abstract-based annotation and extraction. Compared to biomedical research in general, precision oncology is disproportionately less represented in PubMed Central given its funding structure (less open access, more embargoed journal articles) and heavy use of abstract presentations for presenting results–which means many of the latest developments that are so important to capture are not available as full text articles, but only abstracts. We focus on a set of abstracts known to be relevant to precision oncology by annotating only abstracts judged as relevant in the TREC 2017 Precision Medicine track[7]. A random selection of 250 abstracts was chosen from those judged relevant during the assessment process.

3.2 Annotation Process

The 250 abstracts were imported into Brat[37] and double-annotated with the following concept types:

1. CANCER. The type of cancer being studied in the article (e.g., "*breast cancer*", "*non-small cell lung cancer*", "*mantle cell lymphoma*", "*solid tumor*"). If the abstract mentions a type of cancer but it is clearly not the cancer investigated in the study, then it is additionally labeled as a Non-study cancer. If multiple types of cancer are included in the study, all are annotated.

2. MUTATION. The gene mutation being studied in the article, be it a gene with any mutation (e.g., "*KRAS*", "*FGFR2*", "*PIK3R1*"), a specific variant (e.g., "*BRAF V600E*", "*KRAS G13D*", "*NF2 K322*"), or some other form of genetic mutation (e.g., "*CDK4 Amplification*", "*PTEN Inactivating*", "*EML4-ALK Fusion transcript*"). Similar to cancer type, mutations mentioned in the abstract but not investigated in the study are marked as Non-study mutations.

3. POPULATION. The specific population in the study (e.g., "*Hunan Province in China*", "*never or light smokers*", "*adults (> 18 years)*", "*European patients*", "*no history of chemotherapy for metastatic disease*"). As shown by the examples, this can include age, sex, location, ethnicity, cancer status, etc. Populations mentioned in the abstract but not investigated in the study are marked as Non-study populations.

4. TREATMENT. The drug used in the study (e.g., "*sorafenib*", "*abemaciclib*", "*trastuzumab*"). If the drug was used as part of a combination, each individual component is annotated separately. If the drug was a comparator but not directly investigated in the study, then it is marked as a Non-study treatment (this is more common than Non-study cancers, mutations, and populations).

5. OUTCOME. The result of the study with regards to the success or failure of the treatment. Non-study outcomes are not annotated. The outcomes are generally a sentence or long phrase describing the overall outcome. E.g.,

    - *Main grade 3 or 4 toxicities were rash (11 [13%] of 84 patients given erlotinib vs none of 82 patients in the chemotherapy group), neutropenia (none vs 18 [22%]), anaemia (one [1%] vs three [4%]), and increased amino-transferase concentrations (two [2%] vs 0).*
    - *Treatment with crizotinib results in clinical benefit rate of 85%-90% and a median progression-free survival of 9-10 months for this molecular subset of patients.*
    - *Although nearly all patients with GIST treated with imatinib experienced adverse events, most events were mild or moderate in nature.*

Additionally, negated concepts were marked as such.

Two annotators (the first author and a biomedically-trained graduate student) labeled each abstract in batches of 25, reconciling after each batch. Instead of using highly-refined guidelines, the goal of this annotation process was more exploratory in nature. The concepts were defined as above, but no further. The goal was to identify the range of possible ways in which the information can be expressed, without too much regard for maximizing inter-rater agreement.

Number of abstracts	250
Average length of abstract (tokens)	278.1
Total concept annotations	4,722
CANCER	1,622
MUTATION	2,293
POPULATION	133
TREATMENT	544
OUTCOME	130
Percent Non-study annotations	1.2%
CANCER	0.9%
MUTATION	0.8%
POPULATION	0.8%
TREATMENT	4.0%
OUTCOME	0.0%
Average concept length (tokens)	3.3
CANCER	2.7
MUTATION	2.3
POPULATION	4.4
TREATMENT	3.0
OUTCOME	28.5

Table 1: Descriptive statistics of the annotated corpus.

Anecdotally, some concepts had more inconsistent agreement throughout the process (notably POPULATION and OUTCOME), while others had early disagreement that improved over time (such as how to handle acronyms with CANCER and MUTATION). These issues are ultimately reflected in the automatic extraction scores described in Section 4.

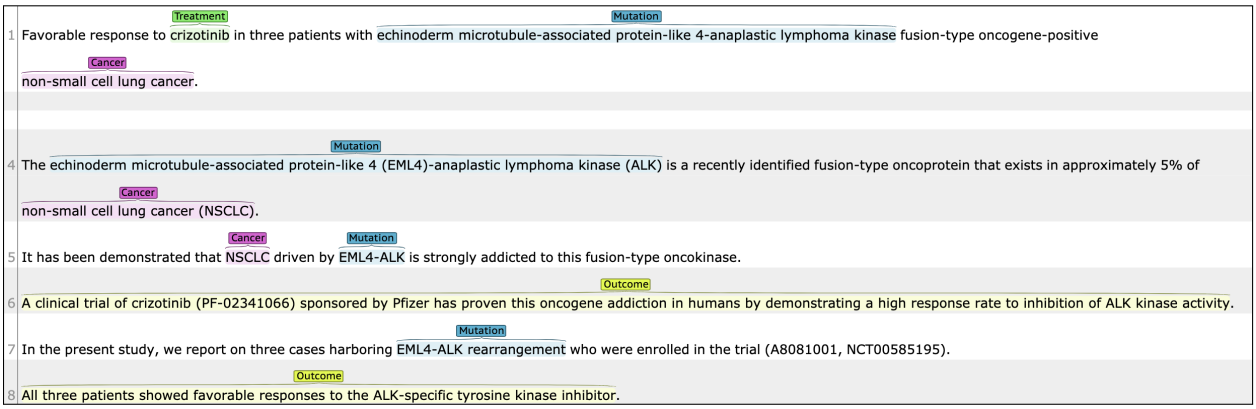
Descriptive statistics of the annotated corpus are provided in Table 1. Example annotations from the corpus are shown in Figure 1.

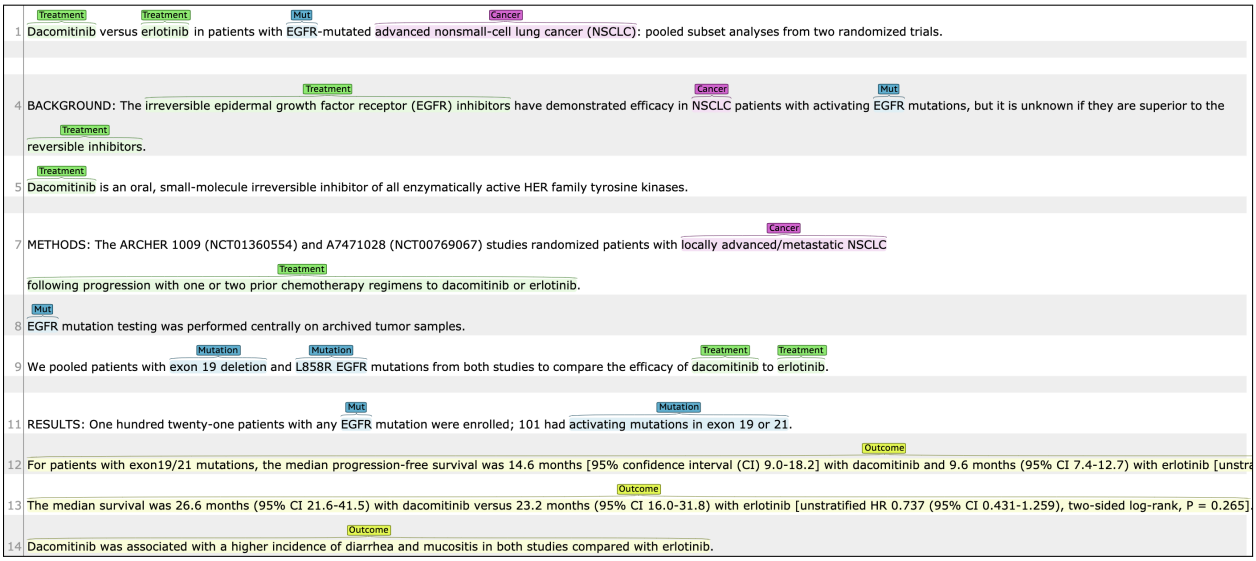
3.3 Automatic Extraction

The abstracts were tokenized and split into sentences using spaCy[38]. A BILOU scheme was used for sequence classification, where B is the first token of a sequence, I an inside token, L the last token, O a token outside any sequence, and U a single-token concept. So "*K - ras and PTEN mutations*" would be [B-MUTATION, I-MUTATION, L-MUTATION, O, U-MUTATION, O]. Non-study concepts were handled by adding a N- before the concept name (e.g., B-N-TREATMENT).

We follow the standard BERT framework for named entity recognition tasks. Two variants of BioBERT[12] were evaluated: $\text{BioBERT}_{\text{BASE}}$ v1.1 and $\text{BioBERT}_{\text{LARGE}}$ v1.1, which are versions of $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ respectively pre-trained on both 1 million PubMed abstracts (note that the BioBERT v1.0 models are pre-trained on 200k PubMed abstracts and 200k PubMed Central full-text articles, but BioBERT v1.1 is only pre-trained on abstracts, though a larger number). As such, BioBERT is an ideal starting point for a transformer-based language model to use for our task. $\text{BioBERT}_{\text{BASE}}$ has 12 layers, 768 hidden units per layer, and 12 attention heads per layer (a total of 110 million parameters); $\text{BioBERT}_{\text{LARGE}}$ has 24 layers, 1024 hidden units per layer, and 16 attention heads per layer (a total of 340 million parameters). Generally, the larger BERT variant offers some improved performance, but in many cases the performance delta is neglible and not worth the additional computational cost. As such, we experiment with both models to assess whether a larger BERT model would be beneficial in this task.

The data was split 70% for training the BioBERT models, 10% for validation (early stopping), and 20% for testing (results discussed below). The default BioBERT parameters were used other than a learning rate of 2×10^{-5}, maximum sequence length of 128, training batch size of 32, validation batch size of 8, and test batch size of 8.





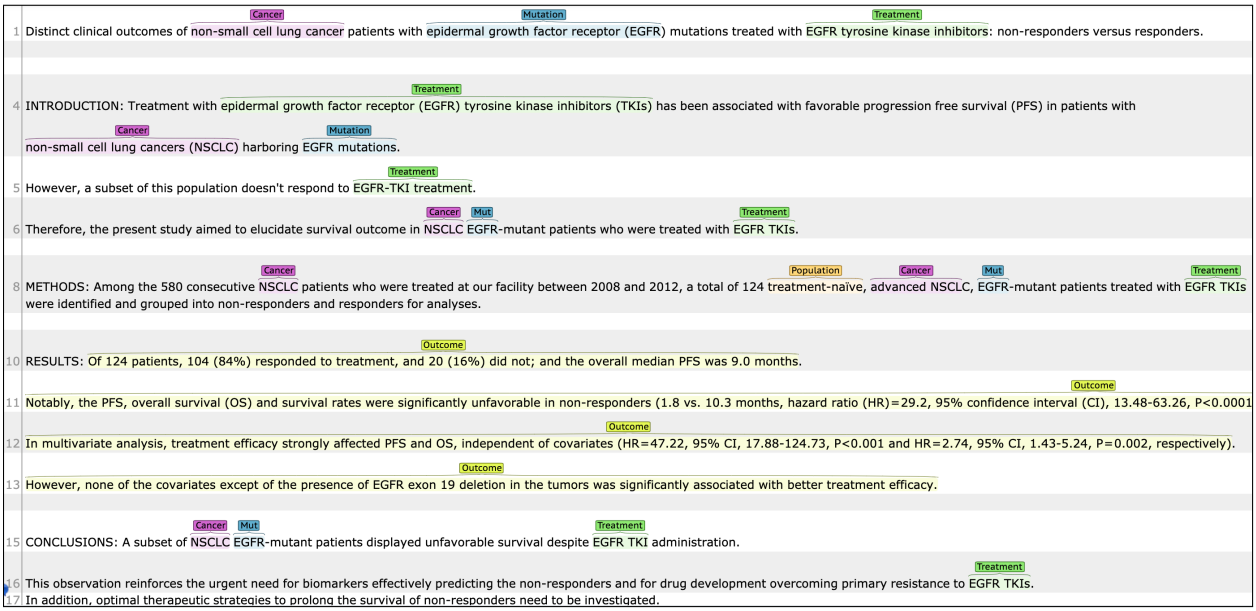


Figure 1: Example annotations

Annotation	Precision	Recall	F1
Overall	60.48	70.73	65.20
CANCER	69.31	78.65	73.68
MUTATION	59.35	69.13	63.87
POPULATION	41.82	42.59	42.20
TREATMENT	47.79	71.05	57.14
OUTCOME	0.0	0.0	0.0

Table 2: Results using $\text{BioBERT}_{\text{BASE}}$ model.

Annotation	Precision	Recall	F1
Overall	63.79	71.90	67.61
CANCER	70.54	80.06	75.00
MUTATION	61.51	68.78	64.94
POPULATION	56.25	50.00	52.94
TREATMENT	58.59	76.32	66.29
OUTCOME	0.0	0.0	0.0

Table 3: Results using $\text{BioBERT}_{\text{LARGE}}$ model.

4 Results

The results for the $\text{BioBERT}_{\text{BASE}}$ and $\text{BioBERT}_{\text{LARGE}}$ models are provided in Table 2 and Table 3. Not enough Non-study concepts are present in the test set to merit an evaluation here. We thus focus on boundary extraction and type classification without the Non-study attribute.

In almost every case, the $\text{BioBERT}_{\text{LARGE}}$ results outperform the $\text{BioBERT}_{\text{BASE}}$ results (the lone exception being MUTATION recall, while neither model successfully extracts any OUTCOME). The differences between $\text{BioBERT}_{\text{BASE}}$ and $\text{BioBERT}_{\text{LARGE}}$ are often several points, including substantial boosts for both POPULATION (+10.74 F1) and TREATMENT (+9.15 F1). Notably, the improvements from $\text{BioBERT}_{\text{BASE}}$ to $\text{BioBERT}_{\text{LARGE}}$ are roughly proportional to the number of available annotations for training, with the most common concept type (MUTATION) receiving the smallest boost. We suspect this is caused by the $\text{BioBERT}_{\text{LARGE}}$ model's superior transfer learning ability having a greater impact for concepts with fewer available manual annotations.

For both models, their performance across the different concept types was roughly proportional to the number of annotations for training. While there were more MUTATION annotations than CANCER annotations, there was a far greater variety of MUTATION mentions than CANCER mentions, which likely explains why CANCER outperforms MUTATION in both models by roughly 10 points of F1. TREATMENT is the next most common concept type, and while for $\text{BioBERT}_{\text{BASE}}$ this performs 6.73 points of F1 worse than MUTATION, for $\text{BioBERT}_{\text{LARGE}}$ TREATMENT actually outperforms MUTATION by 1.35 points of F1. Meanwhile, for both models POPULATION is the second-worst-performing concept type, while as mentioned neither model correctly identifies a single OUTCOME. The latter is almost certainly due to the combination of few annotations (130 in the entire corpus) and long, complex nature of each concept span (28.5 tokens). Clearly, OUTCOME extraction is not an ideal named entity recognition task and should be handled by a different type of extraction (e.g., sentence classification).

Finally, it is interesting that with the exception of POPULATION for $\text{BioBERT}_{\text{LARGE}}$, all concepts have higher recall than precision. This requires further investigation, but one possibility is that the BERT models are good at identifying instances very similar to those in the training data, but additionally predict spans with high biomedical similarity that are nonetheless not one of the annotated concepts.

5 Discussion

This work is an initial feasibility study on the extraction of key variables for precision oncology from biomedical literature abstracts. We focus on identifying the type of cancer, mutation, population information, treatment, and outcomes. A small corpus of 250 abstracts was manually annotated, then two BioBERT models were evaluated. While none of the five concept types performed up to the level one would hope, CANCER performed reasonably well (F1 of 75.00), while MUTATION and TREATMENT showed promise (F1 of 64.94 and 66.29, respectively). POPULATION performed below a level that is likely usable (F1 of 52.94), while OUTCOME was not successfully extracted at all. Here, we discuss the successes and shortcomings of this feasibility pilot and what should come next to address the key problems.

The most obvious need for improvement is the small size of the dataset. Our point of reference for appropriate dataset sizes is the NCBI Disease Corpus,[39,40] which has 793 abstracts, or roughly three times the size of what is presented here. BioBERT's performance on that corpus is an F1 of 89.71, which we can assume is a rough upper bound for automatic extraction if the corpus was scaled up. We will note, however, that even the CANCER, MUTATION, and TREATMENT concepts themselves are more diverse than what is in the NCBI Disease Corpus, and the lexical variation seen with even these concepts is likely greater (especially TREATMENT, see Figure 1), so this would be an ambitious upper bound. Ultimately, it seems clear that increasing the corpus size would be beneficial.

Regarding the lower-performance concepts, it is likely that POPULATION needs to be refined as a concept, which would allow it to incorporate pre-defined lexicons. In this study we intentionally did not define this concept narrowly in order to assess the range of populations mentioned in abstracts. Going forward, however, we can focus on the set of populations that are critically important to precision oncology. These usually differ from the normal medical notion of a population. Instead of demographics, in precision oncology the cancer and treatment history are primary populations of interest (e.g., "*treatment-naive*" in Figure 1 refers to patients who have not yet undergone chemotherapy). Regarding OUTCOME, this is clearly an item that is more appropriately tackled as a sentence classifier than via entity extraction. As can be seen in Figure 1, the OUTCOME sentences have fairly clear features not seen in the other sentences, so it is likely that a sentence classifier could identify these with relatively high efficacy.

The comparison of $\text{BioBERT}_{\text{BASE}}$ and $\text{BioBERT}_{\text{LARGE}}$ is instructive. At the current size of the corpus, the larger model provides more than sufficient benefit to justify its additional complexity. Perhaps in a larger corpus, the base model will close the gap. In other works (e.g., Ji et al.[41]), the larger model performed no better than the base model. These experiments, then, should be revisited with a larger corpus.

Another logical place for improvement is the use of knowledge resources. In this study, we hoped to assess the performance of BioBERT alone, but future work should incorporate existing knowledge resources such as the NCI Thesaurus[42] for cancer names and COSMIC[43] for gene mutations. Above, we stated the NCBI Disease Corpus performance is a good estimate of an upper bound, but the one advantage of focusing exclusively on precision oncology is that more detailed knowledge resources can be brought to bear: a more specific domain allows us to make domain-specific assumptions. This could be critical for improving performance, but there is one important note of caution which also justifies our initial reasoning to evaluate a resource-free approach. Since precision oncology moves quickly as a field, the lexicon of terms used in papers is oftentimes well ahead of knowledge resources. A new oncogene may be identified months or years before it is incorporated into the appropriate knowledge base. Over-reliance on these knowledge sources may increase the NLP performance on the annotated corpus while simultaneously reducing the model's ability to recognize the very emerging concepts we are most focused on identifying. Thus, these knowledge resources cannot be integrated naively, and care should be taken in this process.

A final avenue for improvement focuses on the machine learning aspects. This includes adjusting the tagging scheme– we used BILOU in this study, but given the variance in concept length (see Table 1) other tagging schemes may be more appropriate. Not every concept type need use the same tagging scheme, either. E.g., the shorter MUTATION concepts may utilize a more simple BIO scheme. Additionally, the only form of transfer learning we experimented with in this paper is the use of the BioBERT model itself, which effectively transfers a language model pre-trained on large amounts of biomedical text. After the language modeling, but prior to fine-tuning the model on this precision oncology corpus, other existing datasets may be utilized for transfer learning, such as the NCBI Disease Corpus[39,40]

and PGxCorpus[31]. This would effectively reduce the need to scale up the size of our own manual corpus, though we do not believe that even with transfer learning the current corpus size is sufficient.

Limitations The data evaluated in this study was taken from the TREC Precision Medicine track,[7] and specifically the subset of abstracts marked as relevant for one of the topics. As such, it is certainly not representative of the full array of biomedical literature. This decision was made for annotation convenience–these abstracts were known to be highly relevant to precision oncology. However, the real bias introduced here is the manual nature in which they were chosen. Identifying potentially relevant abstracts to annotate via keywords or machine learning would result in a corpus that is more appropriate, as these methods could be re-applied when using the precision oncology NLP model on new abstracts. A second limitation is the training of the annotators was intentionally kept minimal so as to encourage exploration of potential concepts. Also, only one of the two annotators was biomedically trained. We have discussed the need for additional manual annotation, but this will also need to come with additional training and more refined guidelines to ensure annotation quality.

6 Conclusion

This work presents a pilot study for NLP information extraction of terms related to precision oncology from biomedical literature abstracts. Five concept types were targeted: CANCER, MUTATION, POPULATION, TREATMENT, and OUTCOME. A small corpus of 250 abstracts was manually annotated and reconciled. Two BioBERT models were evaluated for automatic extraction, with the best results ranging in F1 of 75.0 (for CANCER) to a complete inability to extract OUTCOME information. We finally discussed a set of opportunities for future work to improve these results, including a larger corpus, use of existing biomedical knowledge resources, and additional transfer learning.

Acknowledgments

This work was supported by the the Patient-Centered Outcomes Research Institute (PCORI) under award ME-2018C1-10963. The underlying TREC Precision Medicine data was supported by the National Institute of Standards & Technology (NIST).

References

[1] Collins FS, Varmus H. A New Initiative on Precision Medicine. New England Journal of Medicine. 2015;372:793–795.

[2] Garraway LA, Verweij J, Ballman KV. Precision Oncology: An Overview. Journal of Clinical Oncology. 2013;31(15).

[3] Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE. PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Research. 2002;30(1):163–165.

[4] Barbarino JM, Whirl-Carrillo M, Altman RB, Klein TE. PharmGKB: A worldwide resource for pharmacogenomic information. WIREs Systems Biology and Medicine. 2018;10(4):e1417.

[5] Fountzilas E, Tsimberidou AM. Overview of precision oncology trials: challenges and opportunities. Expert Review of Clinical Pharmacology. 2018;11(8):797–804.

[6] Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, Chang MT, Chandarlapaty S, Traina TA, Paik PK, Ho AL, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology. 2017;1.

[7] Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar A, Pant S. Overview of the TREC 2017 Precision Medicine Track. In: Proceedings of the Twenty-Sixth Text Retrieval Conference; 2017. .

[8] Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar A. Overview of the TREC 2018 Precision Medicine Track. In: Proceedings of the Twenty-Seventh Text Retrieval Conference; 2018. .

[9] Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar A. Overview of the TREC 2019 Precision Medicine Track. In: Proceedings of the Twenty-Eighth Text Retrieval Conference; 2019. .

[10] Shenoi SJ, Ly V, Soni S, Roberts K. Developing a Search Engine for Precision Medicine. In: Proceedings of the AMIA Informatics Summit; 2020. p. 579–588.

[11] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018;abs/1810.04805. Available from: http://arxiv.org/abs/1810.04805.

[12] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240.

[13] Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics. 2003;19(13):1699–1706.

[14] Chun H, Tsuruoka Y, Kim J, Shiba R, Nagata N, Hishiki T, Tsujii J. Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. BMC Bioinformatics. 2006;7:S4.

[15] Baker S, Silins I, Guo Y, Ali I, Högberg J, Stenius U, Korhonen A. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics. 2016;32(3):432–440.

[16] Baker S, Korhonen A, Pyysalo S. Cancer Hallmark Text Classification Using Convolutional Neural Networks. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016); 2016. .

[17] Pyysalo S, Baker S, Ali I, Haselwimmer S, Shah T, Young A, Guo Y, Högberg J, Stenius U, Narita M, Korhonen A. LION LBD: a literature-based discovery system for cancer biology. Bioinformatics. 2019;35(9):1553–1561.

[18] Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. International Journal of Medical Informatics. 2014;83(9):605–623.

[19] Yim W, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncology. 2016;2(6):797–804.

[20] Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. Journal of Biomedical Informatics;100:103301.

[21] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: Extraction of Drugs, Genes And Relations from the Biomedical Literature. In: Pacific Symposium on Biocomputing; 2000. p. 517–528.

[22] Libbus B, Kilicoglu H, Rindflesch TC, Mork JG, Aronson AR. Using Natural Language Processing, LocusLink and the Gene Ontology to Compare OMIM to MEDLINE. In: HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases; 2004. p. 69–76.

[23] Consortium GO. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research. 2004;32(suppl_1):D258–D261.

[24] Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting Semantic Predications from MEDLINE Citations for Pharmacogenomics. In: Pacific Symposium on Biocomputing; 2007. .

[25] Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, Rubin DL, Shafa F, Stuart JM, Altman RB. Integrating genotype and phenotype information: an overview of the PharmGKB project. The Pharmacogenomics Journal. 2001;1:167–170.

[26] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics Knowledge for Personalized Medicine. Clinical Pharmacology & Therapeutics. 2012;92(4):414–417.

[27] Thorn CF, Klein TE, Altman RB. PharmGKB: The Pharmacogenomics Knowledge Base. In: Innocenti F, van Schaik R, editors. Pharmacogenomics. vol. 1015. Humana Press; 2013. .

[28] Pakhomov S, McInnes BT, Lamba J, Liu Y, Melton GB, Ghodke Y, Lamba NBV, Birnbaum AK. Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies. Journal of Biomedical Informatics. 2012;45(5):862–869.

[29] Buyko E, Beisswanger E, Hahn U. The Extraction of Pharmacogenetic and Pharmacogenomic Relations–A Case Study Using PharmGKB. In: Pacific Symposium on Biocomputing; 2012. p. 376–387.

[30] Ravikumar KE, Wagholikar KB, Liu H. Towards Pathway Curation Through Literature Mining-A Case Study Using PharmGKB. In: Pacific Symposium on Biocomputing; 2014. p. 352–363.

[31] Legrand J, Gogdemir R, Bousquet C, Dalleau K, Devignes M, Digan W, Lee C, Ndiaye N, Petitpain N, Ringot P, Smaïl-Tabbone M, Toussaint Y, Coulet A. PGxCorpus, a manually annotated corpus for pharmacogenomics. Scientific Data. 2020;7:3.

[32] Cohen KB, Verspoor K, Fort K, Funk C, Bada M, Palmer M, Hunter LE. The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain. In: Ide N, Pustejovsky J, editors. Handbook of Linguistic Annotation; 2017. p. 1379–1394.

[33] Baumgartner W, Bada M, Pyysalo S, Ciosici MR, Hailu N, Pielke-Lombardo H, Regan M, Hunter L. CRAFT Shared Tasks 2019 Overview — Integrated Structure, Semantics, and Coreference. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks; 2019. p. 174–184.

[34] Deng Z, Yin K, Bao Y, Armengol VD, Wang C, Tiwari A, Barzilay R, Parmigiani G, Braun D, Hughes KS. Validation of a Semiautomated Natural Language Processing–Based Procedure for Meta-Analysis of Cancer Susceptibility Gene Penetrance. JCO Clinical Cancer Informatics. 2019;3.

[35] Bao Y, Deng Z, Wang Y, Kim H, Armengo VD, Acevedo F, Ouardaoui N, Wang C, Parmigiani G, Barzilay R, Braun D, Hughes KS. Using Machine Learning and Natural Language Processing to Review and Classify the Medical Literature on Cancer Susceptibility Genes. JCO Clinical Cancer Informatics. 2019;3.

[36] Hughes KS, Zhou J, Bao Y, Singh P, Wang J, Yin K. Natural language processing to facilitate breast cancer research and management. The Breast Journal. 2020;26:92–99.

[37] Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstration Session at EACL 2012; 2012. p. 102–107.

[38] Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing; 2017.

[39] Doğan RI, Lu Z. An improved corpus of disease mentions in PubMed citations. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing;. p. 91–99.

[40] Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of Biomedical Informatics. 2014;47:1–10.

[41] Ji Z, Wei Q, Xu H. BERT-based Ranking for Biomedical Entity Normalization. In: Proceedings of the AMIA Joint Summits on Translational Science; 2020. p. 269–277.

[42] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics. 2007;40(1):30–43.

[43] Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research. 2011;39(Database issue):D945–D950.

Successes and Misses of Global Health Development: Detecting Temporal Concept Drift of Under-5 Mortality Prediction Models with Bias Scan

Ifrah Idrees[1,2,*], Skyler Speakman, PhD[1,*], William Ogallo, RPh, PhD[1], Victor Akinwande[1]
[1]IBM Research – Africa, Nairobi, Kenya; [2]Brown University, Providence, RI; [*]These authors contributed equally

Abstract *Under-5 Mortality rates have been decreasing across Africa for the past two decades. Contributing factors include policy changes, technology, and health investments. This study identifies sub-populations that have experienced more-than-expected change in mortality rates (either increasing or decreasing) during this time period. We train under-5 mortality predictive models on Demographic and Health Survey (DHS) datasets from the early 2000s and apply those models to data collected in more recent versions of the survey. This provides an estimate of the risk current families would have faced in the past. We then apply techniques from anomalous pattern detection to identify sub-populations that have the most divergence between their predicted and observed mortality rates; higher and lower. These detected groups are examples of successes and possible misses of the health progress observed in Africa over the course of decades. Identifying these groups through data-driven discovery may lead to a better understanding of health policies in developing countries.*

Introduction

Improving child mortality rates is an important Maternal, Neonatal and Child Health (MNCH) priority for global sustainable development. Despite a global decline in child mortality rates, many countries are not on track to achieving the global targets of ending preventable deaths among newborns and children under 5 years the year 2030[1]. We contribute to the body of work which shows the progress towards MNCH-specific targets remains uneven within and across countries, reflected in disparities in access to healthcare services and inequitable allocation of resources for MNCH priorities[2].

Some of the key barriers to understand and address MNCH challenges are the complicated interactions of various factors and interventions captured in data. These include socio-economic, health system capacity, and quality of individual care. These scenarios are further complicated when considering health outcome progress over long periods of time (10+ years). The quintessential questions addressed in this paper are, "Which type of women and families are driving the overall decrease in child mortality rates observed in the past decades? Are there groups of women and families that have been left behind?"

This present work leverages data-driven discovery to identify sub-populations of women (and their households) that experience larger-than-expected changes in Under-5 mortality rates between two points in time, spaced approximately 10-15 years apart. The advantage of such approaches is that health investigators do not need to first posit which sub-population to test and then follow up with confirmation analysis and significance testing procedures.[3–8] Rather, it allows the data to highlight the sub-population(s) that are most anomalous in their divergence between observed and expected rates of Under-5 mortality.

Expected mortality rates are obtained by training a machine learning algorithm (Boosted Decision Trees[9,10]) on data from the earlier time-point, $T0$, and then applying that model to data from a more recent time, $T1$. This is analogous to estimating the mortality rates of recent families *if their children had been born 10 to 15 years earlier.* Due to shifts in the data that occur over the course of decades (see Table 1 and Table 2), the predictive model is less accurate when predicting mortality at time $T1$ than $T0$. "Concept Drift" is expected between time steps and is a key component of this work. Rather than viewing these data shifts as traps to be avoided, we use them to identify sub-populations that do not match the country-wide trends in under-5 mortality during the same time frame. Detecting the sub-populations that undergo the largest amount of shift between their predicted and observed mortality rates is accomplished by Bias Scan.[11] See the Methods Section for more details.

Ayele and Zewotir applied a Cox proportional hazard model to Ethiopian Demographic and Health Survey (DHS) data (2000, 2005, 2011) to identify how risk factors for under-5 mortality change over time[12]. The datasets and temporal component are similar to this present work. However, our goal and methodology differ substantially. We

wish to identify sub-populations (i.e. a subset of feature-values) that remain static over time that experience a large change in their under-5 mortality rates between times $T0$ and $T1$. For example, small households in Ethiopia with a single adult living in them and had two births saw their under-5 mortality rates decrease from 47.2% in 2000 to 7.5% in 2016. This drastic change exceeded that of Ethiopia on average between the same time period. More examples of these anomalous sub-populations are provided in the Results section.

Table 1: Description of possible shifts in data.[13] and how they are addressed in this work.

Data Shift Type Notation	Description	MNCH Domain Example	Addressed	How
Covariate Shift $P(X)$	Distribution of features is different	Access to water has changed within the past two decades.	No	Future work may identify the emergence or disappearance of a sub-population over time.
Prior Probability Shift $P(Y)$	Distribution of labels is different	Under-5 Mortality Rates have decreased in past two decades.	Yes	Our null hypothesis is that **all** sub-populations of women (households) have undergone the **same** decrease in odds in the past decade(s).
Concept Drift $P(Y\|X)$	Distribution of labels given features is different	Households with the same education attainment having lower under-5 mortality rates now than before.	Yes	Our alternative hypothesis is that **some** sub-population of women (households) has experienced a **greater** change in odds. Bias Scan identifies this sub-population.
Confounding Shift $P(Y\|X, Z)$	Distribution of labels given a variable that influences both features and labels is different	Changing employment status may influence both wealth index and under-5 mortality rates.	No	Future work may identify anomalous sub-populations by both their change in features and labels.

Data and Data Shifts

Critical to this work is recognizing and exploiting data shifts over time. Varshney summarizes four relevant data shifts that can cause problems for machine learning models.[13] These shifts are (re)listed in Table 1 along with MNCH domain examples and how these shifts are used in this work. Arguably the most common type of data shift is the covariate shift which changes the distribution of the features. This is in contrast to the prior probability shift which changes the distribution of the outcomes (or labels). The data shift of primary concern for this paper is the *Concept Drift* which is a change in the outcome *for a given set of features.*

We highlight that the unit of analysis for this study is the mother (and her household). We only consider survey respondents who have given birth within the 5 years leading up to the survey date. Births and deaths records within the survey are used to create the binary label for under-5 mortality experienced by the mother (and her household).

Table 2 lists information about the five countries explored in this work including the timesteps $T0$ and $T1$, number of mothers (households) in each survey, the under-5 mortality rates, and the discriminating power of a machine learning model to distinguish between women/households who experience under-5 mortality and those who do not (area under the receiver-operator curve, AUC). We highlight the decrease in under-5 mortality rates experienced by each country over time. This change in odds per country $q' = \frac{\text{odds}(T1)}{\text{odds}(T0)}$ is an important part of methodology (see Figure 1). Second, we note that the AUC decreases for $T1$ because it is using the model trained on $T0$ data to predict $T1$ data. This drop in accuracy is expected under data shifts such as temporally-induced Concept Drift. We apply Bias Scan[11] to exploit this drift and identify which sub-populations in $T1$ differ the most between their expected and

Table 2: Country-level statistics for years of surveys ($T0$, $T1$), sample size, under-5 mortality rates, and predictive model accuracy.

Country	Years	Size	Under-5 Mortality(%)	AUC
Burkina Faso	2003	7367	15.9	0.869
	2010	10364	11.7	0.855
Ethiopia	2000	7245	16.1	0.838
	2016	7193	8.1	0.798
Kenya	2003	3972	11.0	0.931
	2014	14949	5.5	0.895
Nigeria	2003	3775	19.6	0.870
	2018	21792	13.1	0.839
Tanzania	2004	5658	11.6	0.875
	2015	7050	6.8	0.868

observed mortality rates.

Table 3 provides a subset of the features and their values extracted from both timesteps of the DHS data at each country. These features were used both in the training of the predictive classifier and to create the search space for Bias Scan to efficiently identify anomalous sub-populations. In some scenarios, some features were removed to correctly address additions/removals of survey questions over the decades.

Methods: Training, Calibrating, and Shifting Models

In order to capture the relationship between the features X and the mortality label Y, we trained a predictive classifier using off-the-shelf software[10] and methods.[9] The advantages to these methods over more standardized regression techniques are well captured in Ogallo's approach to this similar problem[14]. In summary, boosted decision trees are able to better capture non-linearities in the relationship between the features and the outcome as well as interactions between features. This more expressive form of $P(Y|X)$ typically results in higher discriminating power. The number of trees and depth of individual trees were chosen through cross-validation that optimized the AUC of the held-out cross-validation set.

The second part of Step 1 as shown in Figure 1 is calibrating the model so that the predicted probabilities accurately reflect the true proportion of observed outcomes in time $T0$. This was done using Platt Scaling[15] option with held out training data.

An example of the calibrated $T0$ model from Ethiopia is shown in Figure 2 as a calibration plot. When the model is used to predict $T0$ data, there is strong agreement between the predicted probabilities and the observed fraction of positive cases. However, when predicting $T1$ data we observe that the model is systematically producing a higher predicted probability for the proportion of true outcomes. This is due to the data shift in Ethiopia's mortality rates between $T0$ and $T1$. The rate decreased from 16.1% to 8.1% over the course of 16 years.

This insight leads us to Step 3 in the workflow diagram (Figure 1). In order for Bias Scan to better identify the temporal Concept Drift in the data, we must separately attempt to address the prior probability shift. On average, the odds of under-5 mortality decreased by a factor of 0.46 in Ethiopia between $T0$ and $T1$. Therefore, a more accurate version of the predicted probability on $T1$ data is to also shift these predictions by the same change in the odds. This shift is done to *all* records in $T1$ predicted probabilities before running Bias Scan.

Methods: Bias Scan

Bias Scan[11] efficiently identifies subsets of the data where a predictive classifier is systematically over (or under) estimating the probability of of an observed outcome. Bias Scan exploits mathematical properties of the scoring

Table 3: Features and their values used for training and scanning. Due to variations in survey questions, not all countries have all the features.

Feature	Feature Values
Marital status	Married/Living with Partner, Divorced/Separated, Never in Union/Widowed
Respondent employed	Not working, Ag-self/Ag-employee, Clerical/Sales, Household/Services/Skilled/Unskilled/Other, Professional, Missing/Unknown
Source of drinking water	Piped, Well, Borehole, Other, Missing/Unknown
Ethnicity	Varies by country
Living with partner	Living with her, Staying elsewhere
Region	Varies by Country
Respondent Education Level	Primary, Secondary, Higher, No education
Visited health facility in the past 12 months	Yes, No
Gender of Head of Household	Male, Female
Respondent worked in the past 12 months	Yes, No
Relationship Structure	Three+ related adults, Two adults - opposite sex Two adults - same sex/Unrelated, One adult, No adults
Literacy	No/Blind/Missing/Unknown, Yes
Number of Children Under 5 who slept under No Mosquito Net	All children, No net in household , No, Some children
Wealth Index	Poorest, Poorer, Middle, Richer, Richest
Household Size	Missing/Unknown, 1 to 3, 4 to 5, 6 to 8 9 and above
Number Of Births	Missing/Unknown,One, Two, Three+
Respondent Age	Missing, Below 20, 20 to 29, 30 to 39, 40 to 49, 50 and above
BMI	Missing/Unknown,Underweight,Normal, Overweight, Obese
Ever Terminated Pregnancy	Yes, No

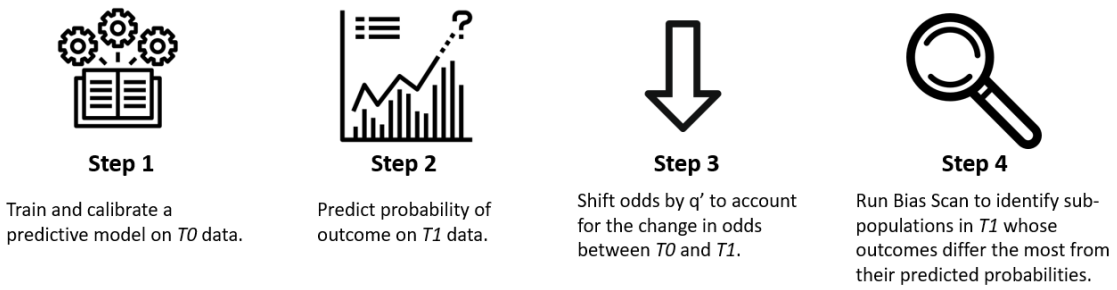


Figure 1: Workflow Diagram for detecting Concept Drift between timesteps $T0$ and $T1$. Step 3 uses q' which is the odds ratio of observed mortality between $T1$ and $T0$.

function[16,17] which makes it computationally feasible to "scan" over the exponentially-many possible subsets of records in a data set. Scanning for anomalous subsets, rather than investigating subgroups of apriori interest, is the critical component of data-driven discovery.

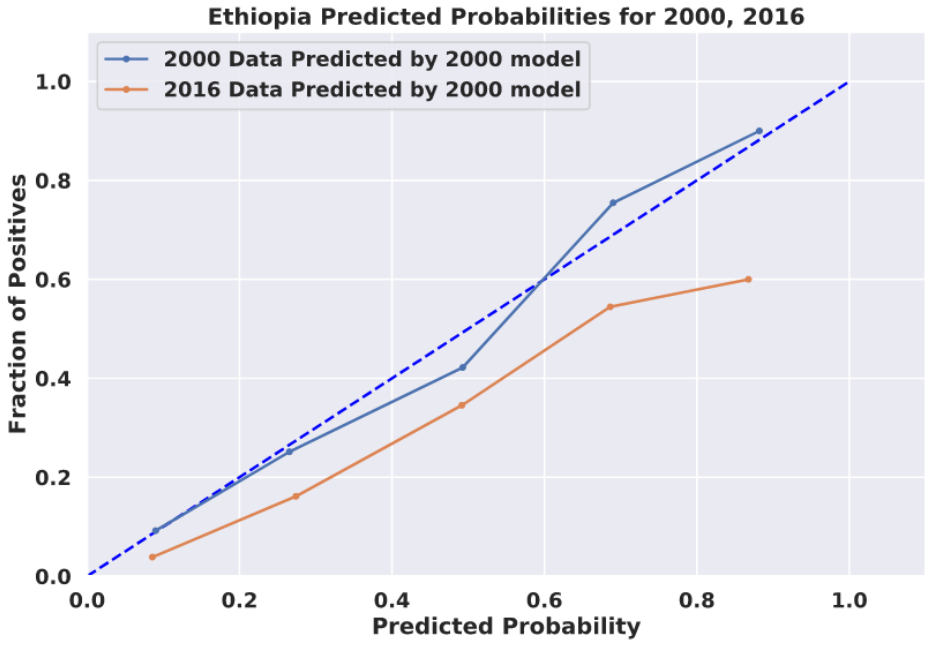


Figure 2: Calibration Plot showing the data distribution shift in Ethiopia between $T0$ and $T1$. The $T1$ predictions are pessimistic, assigning higher probability of mortality to women/households that did not experience Under-5 mortality. Bias Scan is used to identify the sub-population where this divergence shows the most evidence of concept drift.

Bias Scan may be viewed through lens of hypothesis framing. The null hypothesis is that *all* sub-populations of $T1$ households are accurately predicted by the model trained on $T0$ data. The alternative hypothesis assumes some multiplicative bias, q, in the odds for *some* sub-population, S.

$$H_0 : odds(y_i) = \frac{\hat{p_i}}{1-\hat{p_i}} \forall i \in D$$

$$H_1 : odds(y_i) = q\frac{\hat{p_i}}{1-\hat{p_i}}, \text{ where } q > 1 \, \forall i \in S \text{ and } q = 1 \, \forall i \notin S$$

These hypotheses form a log-likelihood ratio based on the Bernoulli distribution and form the "bias score" of some sub-population, S. This score appropriately balances the size of the sub-population along with deviation between the predicted probabilities and the proportion of positive outcomes in the subgroup. Any sub-population that has a large number of records with systematically higher predicted probabilities $\hat{p_i}$ as compared to the observed outcomes y_i in the same subgroup, will have a high bias score.

$$\begin{aligned} score_{bias}(S) &= \max_q \log \prod_{i \in S} \frac{\text{Bernoulli}(\frac{q\hat{p_i}}{1-\hat{p_i}+.q\hat{p_i}})}{\text{Bernoulli}(\hat{p_i})} \\ &= \max_q \log(q) \sum_{i \in S} y_i \sum_{i \in S} \log(1-\hat{p_i}+q\hat{p_i}) \end{aligned} \tag{1}$$

Bias Scan uses an iterative ascent procedure to efficiently optimize this objective function over all possible sub-populations. When optimizing over a feature with k possible values, Bias Scan does not consider each of the $2^k - 1$

possible subsets, as that would be computationally infeasible. Rather, previous work[16] has shown that at most $2k$ subsets must be considered while still guaranteeing that the subset with the highest bias score will be identified. This reduction from exponential to linearly-many subsets to consider is what makes Bias Scan computationally efficient for large data sets. Each ascent is guaranteed to converge to a local optimum and multiple random restarts are used to, ideally, converge to a global maximum. Thirty (30) random restarts were used for this piece of work. Additionally, Bias Scan has a tuning parameter that penalizes complex sub-populations that may span too many features, inhibiting interpretation. We used a penalty value between 2.5 and 4.0 in our experiments. These values resulted in interpretable subsets spanning 1-3 features. See the Results section for more details.

We conclude this section by noting two simple extensions to Bias Scan we used in this current piece. This is the first application of Bias Scan to explicitly search over temporally-induced concept drift. Previous uses of Bias Scan focused more on identifying faults within the predictive classifier. Here, we assume the classifier is operating correctly and it is the *data* that is shifting between $T0$ and $T1$. It is these shifts that we extract insights from rather than attempting to highlight weaknesses or bias in the predictive model. Second, this is the first application where prior probability shift is addressed separately from the scanning process. (See Step 3 in Figure 1). Without this global shift of $T1$ predictions, the scanning results would typically identify "All" sub-populations as anomalous. However, that result was due to the prior probability shift and not the concept drift. By shifting the predicted probabilities of $T1$, we are addressing the *country-wide* change in outcomes between $T0$ and $T1$. Therefore, any further deviations between the shifted probabilities and the observed outcomes at $T1$ can be more readily attributed to concept drift of a particular sub-population. Tables 4 and 5 in the Results Section provide the average predictions for the identified sub-population before and after this shift (last two columns). This is to reinforce our null hypothesis that all sub-populations have experienced the same decrease in the odds of mortality across the country.

Results

Bias Scan[11] was applied in two "directions". In the negative direction, it identifies sub-populations that have *lower* observed rates of mortality in $T1$ than expected according to the model trained on $T0$ data. These sub-populations are showing gains that outpace the country on average and could be argued as successes. The positive direction detects sub-populations with observed mortality rates *higher* than expected and highlight possible misses of development efforts to emphasize going forward. These sub-populations are listed in Tables 4 and 5, respectively.

We begin by looking at the successess in Table 4. We categorize these results as "drivers" of the change vs "exceptional cases" of the change. Burkina Faso and Nigeria have sub-populations that could be argued as driving the country-wide decrease in under-5 mortality (due to their large size). For example, in Nigeria women in the South South or South West regions saw their under-5 mortality rates decrease from 14.8% to 7.4%. Although this sub-population was already doing better than average in 2003 (19.6%) the decreases made over the next 15 years were substantial. The last two columns of this table show that the model was accurate in predicting approximately 14.0% mortality rate for the sub-population. If that sub-population experienced the same change in the odds as the rest of the country between $T0$ and $T1$, then the mean would be 10.3%. However, in reality the observed rate was 7.4% at time $T1$. This divergence (and large size of the group) flag it as an anomalous sub-population.

The successes in Kenya and Ethiopia are examples of an exceptional case in the decrease in under-5 mortality. Among educated women in Kenya between the ages of 40 and 49, the mortality rate decreased from 16.7% to 0.0%. Furthermore, in Ethiopia, small households with one adult and two births saw meteoric drops in rates from 47.2% to 7.5%. However, we note that these groups cover a small sub-population.

Finally we note the possible counter-intuitive result in Tanzania. How is it that households *without* a malaria net had their rates *decrease* from 8.9% to 0.2%? A household without a net in 2015 likely does not live in a Malaria area, whereas household without a net in 2004 (before their widespread distribution) simply may not have had a net despite living in a Malaria area.

We now look at the groups which are showing delayed improvements in mortality rates (compared to the average change) from the countries under consideration. In Tanzania, the age of the mother (under 20 years) highlights a sub-population that continues to have high under-5 mortality rates. For the women in this subgroup, the under-5 mortality rate has not decreased at the rate of the rest of the country. This result highlights the importance of

Table 4: Anomalous sub-populations with lower-than-expected under-5 mortality rates for each country.

Country Years Mortality	Sub-Population	Size		Mortality %		Mean Model Predictions %	
		$T0$	$T1$	$T0$	$T1$	$T1$ **Raw**	$T1$ **Shifted**
Burkina Faso 2003, 2010 15.9%, 11.7%	Ethnicity = Mossi, Worked in past year = Yes, Visited Health Facility in past year = Yes	1501	3390	15.1	8.5	13.7	10.8
Kenya 2003, 2014 11.0%, 5.5%	Respondent Education Level = Higher, Respondent's Age = 40 to 49 years	6	58	16.7	0.0	32.0	25.1
Ethiopia 2000, 2016 16.1%, 8.1%	Relationship structure = One Adult, Number of Births = 2, Household size = 1 to 3	53	67	47.2	7.5	53.5	36.7
Nigeria 2003, 2018 19.6%, 13.1%	Region = South South or South West	809	4737	14.8	7.4	14.0	10.3
Tanzania 2004, 2015 11.6%, 6.8%	Number of children who slept under mosquito net = No net in household, Number of Births = 1	1660	914	8.9	0.2	3.0	1.7

Table 5: Anomalous sub-populations with higher-than-expected under-5 mortality rates for each country.

Country Years, Mortality	Sub-Population	Size		Mortality %		Mean Model Predictions %	
		$T0$	$T1$	$T0$	$T1$	$T1$ **Raw**	$T1$ **Shifted**
Tanzania 2004, 2015 11.6%, 6.8%	Respondent's Age = Below 20	395	550	7.6	7.6	7.3	4.6
Burkina Faso 2003, 2010 15.9%, 11.7%	Visited Health Facility in past year = No Relationship structure = 3 or more adults	2957	1476	16.0	16.0	14.1	11.4
Kenya 2003, 2014 11.0%, 5.5%	Number of Births = 2, Gender of Head of Household = Male, Household size = 1-3	60	126	56.7	57.1	29.9	18.3
Ethiopia 2000, 2016 16.1%, 8.1%	Null	7245	7193	16.1	8.1	15.2	8.6
Nigeria 2003, 2018 19.6%, 13.1%	Relationship structure = Two adults, Number of Births = 2, Household size = 1-3	68	328	73.5	76.8	56.5	46.1

incorporating the prior probability shift before scanning for bias. Without that shift, this group of women would have appeared normal because the model's raw predictions (mean) was similar to that of the observed mortality rate at time $T1$. However, maintaining the same under-5 mortality rates while the rest of the country decreases should be reported as anomalous.

Second, we look at Ethiopia's promising result. There were no (large) sub-populations detected that systematically had higher rates of under-5 mortality. This failure to reject the null hypothesis suggests strong performance in Ethiopia that the overall reduction in mortality appears to be more inclusive than other countries. At lower penalty complexities a low-scoring subset did emerge, however.

Burkina Faso has an interesting result due to the feature of attending a health facility. Those who had not attended a health facility had an under-5 mortality rate that stayed constant at 16.0%. This is in comparison to those who did attend a facility in the past year and that group saw their mortality rates decrease from 15.1% to 8.5%. There are other features involved such as ethnicity and employment status but a plausible hypothesis is that quality of care at these facilities increased between the years 2003 and 2010 (at least for Mossi ethnicity).

We conclude the results section by highlighting the repeated presence of a few features in the groups where changes in the mortality rate were less than the national averages. These are the number of births and household size. We believe their presence is explained by "label leaking" in the predictive model. This occurs in the training stage of a predictive model when combination of features indirectly provide information about the class label that otherwise would not be known. For example, how can a household that has 2 adults and 2 births in the past 5 years be of size 1-3? This is possible if one of those children has died (otherwise the household would be larger). Label leaking results in an artificially strong predictive model, but for the wrong reasons. Future work is needed to determine if Bias Scan can be used as a more generalized detection of the "label leaking" effect. However, Bias Scan correctly identifies these groups as anomalous because the sub-population's under-5 mortality rate has not changed over the course of a decade of improvements.

Discussion, Strengths and Limitations

There are 10's of trillions of possible sub-populations that span the feature values listed in Table 3. Traditionally, this search space of possible hypotheses to investigate (i.e. Has the under-5 mortality rate of large households in Nigeria's South South region decreased in the past decade?) is narrowed down by domain experts who then manually inspect a small number of sub-populations of interest. In this work we show that data-driven discovery can be used to identify anomalous sub-populations that are experiencing a changing under-5 mortality rate more extreme than the country, as a whole. Critically, Bias scan efficiently explores this large space and is able to identify sub-populations that maximize a log likelihood ratio statistic based on the Bernoulli distribution. These are the sub-populations with observed mortality rates in $T1$ that differ from their expected mortality rates as predicted by a machine learning model trained on $T0$ data.

This is not meant to diminish the role of domain experts, by any means. An excellent example of the role of domain experts that is not explored in this paper is narrowing the search space of Bias Scan. For simplicity, we allowed Bias Scan to search over the exact same feature-values as those used to train (and predict) the under-5 mortality rates. This may result in some awkward sub-populations (See Tables 4 and 5). However, if an investigator wanted to know more about the effect of regional healthcare initiatives, they could remove some demographic features from the search space such as relationship structure of the adults in the household. This does not mean the investigator is specifying the sub-population, but rather the potential space for Bias Scan to efficiently explore. We believe this type of interaction captures the best parts of data-driven and hypothesis-driven research.

The analysis in this paper does fall short in a couple of areas. The primary limitation is our approach's indifference to covariate and confounding shifts. For example, it is possible that another sub-population experienced a larger change in the odds than the one we detected. However, that sub-population also decreased in size between $T0$ and $T1$ due to a covariate shift. This small size at $T0$ decreases the sub-population's bias score and therefore goes undetected. Future work in this space may incorporate penalties to the bias score when the sub-population has drastically different sizes between the two time steps. Alternatively, we could attempt to bypass the model training and predicting step entirely and form expectations of mortality directly from the mortality rates at $T0$. Second, we

are limited by domain expertise to explore the causal factors that may explain *why* these sub-populations experienced such relative large changes in mortality odds. Perhaps there was increased regional healthcare capacity during those years or government-backed initiatives to actively address under-5 mortality rates? The sub-populations identified in this work may lead to hypothesis generation in follow-up studies designed to identify a causal impact of an intervention put in place between $T0$ and $T1$ on a per-country basis. However, that is outside the scope of the current work.

Conclusion

The world's under-5 mortality rate is decreasing but not everyone is experiencing the gains. To that end, this piece of work demonstrated how temporal data distribution changes, such as concept drift and prior probability shift, can be used to identify sub-populations of women and their households that are benefiting the most (and the least) from these global trends. Data-driven discovery is at the core of this analysis. We do not investigate apriori sub-populations of interest for statistically significant changes in mortality rates. Such an approach puts too much onus on the investigator to correctly hypothesize the sub-population from domain knowledge. Instead, we train predictive classifiers to estimate what the mortality rate would have been if the global trends in reducing under-5 mortality were experienced uniformly across all sub-populations within a country. We then apply Bias Scan, a technique from anomalous pattern detection, to identify sub-populations of records that have mortality rates that differ the most from this expectation. The resulting groups represent the successes and delayed results (or possible missteps) of recent changes in global health developments.

Some of the identified sub-populations, such as the southern regions of Nigeria, can be considered "drivers" of the overall decrease in mortality rates. Other sub-populations, like single mothers in Ethiopia with 2 children, highlight exceptional cases where mortality rates decreased from 47% to less than 8% in 16 years. Future work is needed to appropriately assign *causal* connections to the observed changes in under-5 mortality rates identified in this work.

Bias Scan is a powerful tool for health informatics, more generally. As predictive models become more common-place in healthcare applications it will be more important for researchers to understand and acknowledge data distribution changes and how it impacts the performance of the model across different sub-populations of patients.

Acknowledgements

This work is funded by Bill & Melinda Gates Foundation, investment ID 52720. We would like to particularly thank Claire Mershon and Dr. Nosa Orobaton for their MNCH-specific insights on this work.

Dr. Daniel B. Neill also played an important role in sharing code updates for Bias Scan.

References

1. United Nations Development Program. The sustainable development goals. 3: Good health and well-being. 2019.
2. Aluísio JD Barros and Cesar G Victora. Measuring coverage in mnch: determining and interpreting inequalities in coverage of maternal, newborn, and child health interventions. *PLoS medicine*, 10(5), 2013.
3. Grace Kaguthi et al. Predictors of post neonatal mortality in western kenya: a cohort study. *The Pan African medical journal*, 31, 2018.
4. Hayelom Gebrekirstos Mengesha et al. Survival of neonates and predictors of their mortality in tigray region, northern ethiopia: prospective cohort study. *BMC pregnancy and childbirth*, 16(1):202, 2016.
5. Yared Mekonnen et al. Neonatal mortality in ethiopia: trends and determinants. *BMC public health*, 13(1), 2013.
6. Ifeoma D Ozodiegwu et al. Country-level analysis of the association between maternal obesity and neonatal mortality in 34 sub-saharan african countries. *Annals of Global Health*, 85(1), 2019.
7. Sanni Yaya et al. Decomposing the rural-urban gap in the factors of under-five mortality in sub-saharan africa? evidence from 35 countries. *BMC public health*, 19(1):616, 2019.
8. Zufan Bitew Dessie et al. Maternal characteristics and nutritional status among 6–59 months of children in ethiopia: further analysis of demographic and health survey. *BMC pediatrics*, 19(1):83, 2019.

9. Jerome H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, February 2002.
10. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
11. Zhe Zhang and Daniel B. Neill. Identifying significant predictive bias in classifiers, 2016.
12. Dawit G Ayele and Temesgen T Zewotir. Comparison of under-five mortality for 2000, 2005 and 2011 surveys in ethiopia. *BMC public health*, 16(1):930, 2016.
13. Kush R. Varshney. Trustworthy machine learning and artificial intelligence. *XRDS*, 25(3):26–29, April 2019.
14. William Ogallo, Skyler Speakman, Victor Abayomi Akinwande, Kush Varshney, Aisha Walcott-Bryant, Charity Wayua, Komminist Weldemariam, Claure-helene Mershon, and Nosa Orobaton. Identifying factors associated with neonatal mortality in sub-saharan africa using machine learning. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2020 (to appear).
15. John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.
16. Skyler Speakman, Sriram Somanchi, Edward McFowland III, and Daniel B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 25(2):382–404, 2016.
17. Daniel B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

Auto-mapping Clinical Documents to ICD-10 using SNOMED-CT

Natthanaphop Isaradech[1], Piyapong Khumrin, MD, PhD[1,2]
[1]Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand; [2]Biomedical Informatics Center, Department of Family Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

Abstract *Excessive paperwork is a considerable issue that leads to additional burdens for health-care professionals. In Thai health-care systems, physicians manually review medical records to select an appropriate principle diagnosis and other co-morbidities and convert them into ICD-10s to claim financial support from the government. Accordingly, 160,000 ICD-10 codes and 46,000 in-patient discharge summaries are documented by physicians at Maharaj Nakorn Chiang Mai hospital each year. As a result, to decrease physicians' burden of manual paper-work, we created a new approach to automatically analyse discharge summary notes and map the diagnoses to ICD-10s. We combined SNOMED-CT and natural language processing techniques within the approach through 3 steps: cleaning data; extracting keywords from discharge summary notes; and matching keywords to ICD-10. In this paper, we present that mapping clinical documents by using approximate matching and SNOMED-CT shows potential to be used for automating the ICD-10 mapping process.*

Introduction

ICD-10 stands for the 10th revision of the international classification of diseases and related health and medical problems documented by the World Health Organisation (WHO)[1]. The ICD-10 contains international identification numbers for terms such as signs and symptoms, disease names, procedures and abnormal findings, which are used for global health information standard classifications for mortality, morbidity statistics, clinical care and research, to analyse diseases and study disease patterns, as well as to manage health-care, monitor outcomes, and allocate resources[2–4]. In Thai health-care systems, physicians review medical records to select an appropriate principle diagnosis and other co-morbidities, and convert them into ICD-10s to claim financial support from the government. Errors occurring in the process, for instance, missing ICD-10 inputs or incorrect interpretation affect the correctness and completeness of the records, resulting in challenges such as: intensive labour necessary for rechecking the ICD-10 inputs; incorrect health epidemiological reports; and insufficient reimbursement for managing a hospital[2,5].

Paper-work overload is a major problem and leads to a heavy burden for health-care professionals[6]. On average, each year there are 160,000 ICD-10 codes documented by physicians and 46,000 discharge summaries generated at inpatient departments at the Maharaj Nakhon Chiang Mai hospital. Due to the massive workload, there is an increased risk of human error during the processing of ICD-10 documentation.

To address the challenges involved in human ICD-10 processing, computer technologies could help to assist or automate the ICD-10 mapping process. In 2008, a study was conducted into reverse mapping of ICD-10-CA (Canadian version of the 10th revision of the International Statistical Classification of Diseases) to SNOMED-CT by applying an exact and partial mapping algorithm[7]. However, this method had limitations for mapping abstract terms. Because of the limitations of this study, the algorithms were only recommended to semi-automate the ICD-10 mapping process and support humans to complete the documentation. Consequently, in 2009, natural language processing techniques were applied to increase the efficacy of the ICD-10 mapping process, producing more precise results using acronym processing and noun phrase extraction[8].

To address the ICD-10 mapping problem, we developed a new approach by combining SNOMED-CT and natural language processing techniques to automatically extract electronic discharge summary notes recorded by physicians in Maharaj Nakorn Chiang Mai hospital into ICD-10. Our goal is that this approach could help to decrease the burden of manual ICD-10 mapping processes for healthcare professionals, provide more complete documents with reduced errors, and further help the hospital administrators to fully and accurately claim financial support from the government.

Materials and Methods

Samples

We randomly collected 560 discharge summaries at Maharaj Nakhon Chiang Mai hospital from 2006 to 2016 to test our approach. The data were retrospectively retrieved under an ethical approval by the Research Ethics Committee of Faculty of Medicine, Chiang Mai University (Study code: PHY-2562-06152). All methods were carried out with exemption criteria (waiver of informed consent) in accordance with the Faculty of Medicine, Chiang Mai University regulations.

SNOMED-CT ICD-10 dictionary

We used the SNOMED-CT dictionary (version 45_THIS_SNOMED_CT_US_201903_Full_Excel[9]) provided by the Thai Health Information Standards Development Centre (THIS)[10]. The dictionary included the SNOMED-CT terms of abnormalities, disorders, findings, and procedures that map to ICD-10. The dictionary provides a complete map (representing relationship and hierarchy by concept id) between medical terms and ICD-10. SNOMED-CT or Systemic Nomenclature of Medicine Clinical Terms is an organised collection of clinical terms arranged in a comprehensive manner, created by the International Health Terminology Standards Development Organisation (IHTSDO) in 2007[11]. The library consists of core general terminology for electronic hospital documentation. It covers a wide variety of medical categories such as clinical signs and symptoms, findings, diagnoses, organisms, procedures, pharmaceuticals, medical devices, and so forth. A collaboration between the IHTSDO and the WHO in 2012 linked SNOMED-CT nomenclature to ICD-10 terminology using concept ids[12]. Based on this relationship, we applied these concepts to link medical terms in the discharge summary notes to ICD-10.

Map discharge summaries to ICD-10

The discharge summaries were processsed to map to ICD-10 with the following 3 steps:

1. Cleaning data
2. Extracting keywords from discharge summary notes
3. Matching keywords to ICD-10

Because the discharge summary notes and ICD-10 description terms of SNOMED-CT were in a free-form text, we were not able to directly map the contents to ICD-10. The contents were required to be tokenised and pre-processed to remove irrelevant words and characters. Regular expression was used to convert text to lowercase and remove non-English alphabet items, non-digits, and stop words. The remaining contents were treated as keywords. The keywords contained patient information, findings, abnormalities and some keywords or phrases related or relevant to ICD-10s. Mapping keywords from the notes directly to ICD-10 is less likely to achieve matches because ICD-10 terms were not the sole terms recorded in the notes. Therefore, we need to find a bridge between keywords and ICD-10. The SNOMED-CT dictionary links medical terms and ICD-10 by concept id which allowed us to match the keywords in the notes to concept id and then link the id to related ICD-10s. Because of the variation of the keywords in the notes, Levenshtein Distance was utilised to map terms between the keywords and SNOMED-CT terms. Levenshtein Distance theory was introduced by Vladimir Levenshtein in 1965[13]. The theory is used to measure the distance between two words (text similarity) by finding the minimum number of operations needed to change a word sequence into another word using insertions, deletions, or substitutions. For example, the Levenshtein distance between “Oedema” and “Edema” is 1 (1 deletion of “O”) or the Levenshtein distance between “Mycoplasma infection” and “Mycoplasma pneumonia” is 7 (3 substitutions of “i”, “t”, “i”, 2 insertions of “f” and “c”, and 2 deletions of “i” and “a”). The greater the distance, the more difference between two words.

We applied the Levenshtein Distance for measuring text similarity using a Fuzzy wuzzy package[14] to compute the standard Levenshtein distance and calculate similarity ratios (presented as percentages) between a collection of words of two sentences. We used the token_set_ratio function that performed a set operation to pairwise intersect words

and then paired these with other non-intersected words to calculate the ratio, then selected the highest ratio. The set operation allows flexibility of comparing unsorted words and sentences with different sizes. The higher the similarity ratio, the more similarity between two sequences. The performance of token_set_ratio function was compared with the other two baseline algorithms: 1) character set ratio and 2) Jaccard similarity. Character set ratio is calculated by the number of characters that intersects between two strings divided by the size of the union of all unique characters. Jaccard similarity[15] is calculated similar to the character set ratio but it is applied on the token level. All tokens of the discharge summary notes were compared with SNOMED-CT tokens. In case of one ICD-10 was linked to multiple SNOMED-CT terms, the term with the highest similarity was selected. Finally, we ranked ICD-10s by the similarity values and reset index to a percentage scale, ranked from 100 to 0.

We evaluated the algorithm performance by calculating the minimum, maximum, median, average and standard deviation of the ranks and compared between the three algorithms. The higher number of the rank of actual ICD-10s represented the higher quality of the ICD-10 mapping. We are interested to see how the actual IC0-10s were correctly detected (rated in a high rank) and the actual ICD-10s that were missed by the algorithms (rated in a low rank). In addition, we explored false predicted ICD-10s which were ranked higher than the actual ICD-10s. True and false positive were defined by the following criteria. True positive was the actual ICD-10s which were ranked higher than 50%. False positive was the false predicted ICD-10s which were ranked higher than actual ICD-10s. False negative was the actual ICD-10s which were ranked lower than 50%. True negative was not evaluated in this study. Finally, we manually evaluated and discussed the pattern of true positive, false positive, and false negative.

Results

We demonstrated how the raw discharge summary note was processed until mapping relevant ICD-10s was achieved. First, the raw notes were cleaned by using regular expressions to remove non-English alphabet items, non-digits, and stop words. Figure 1 shows text before and after the discharge summary was cleaned. SNOMED-CT terms were also cleaned by the same process (data not shown).

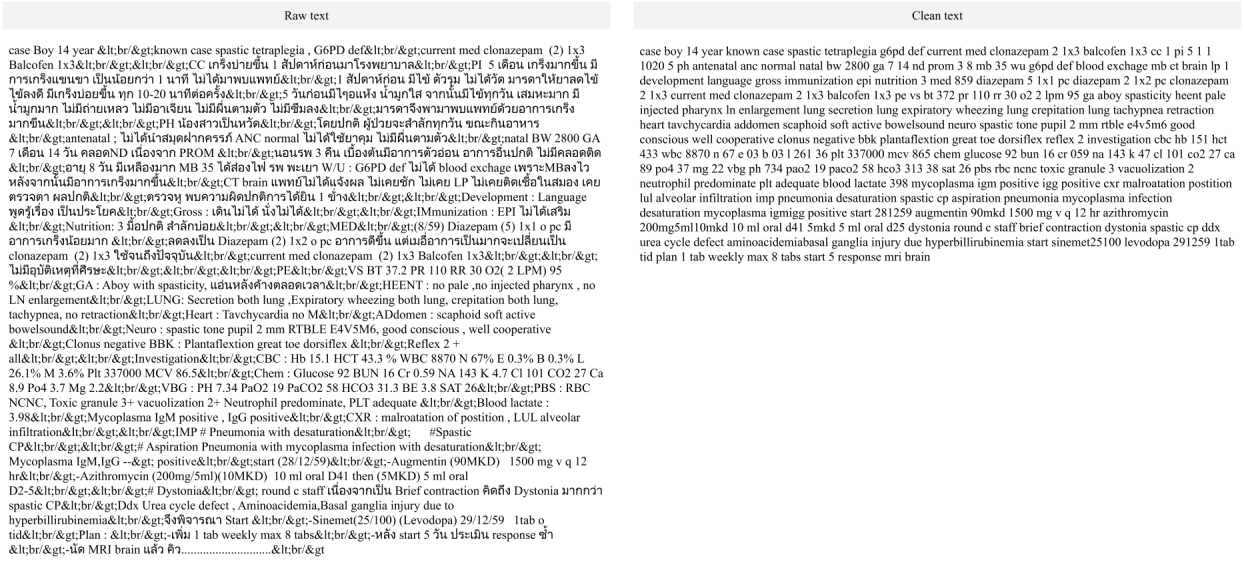


Figure 1: Text comparison between pre (raw text on the left hand side) and post (clean text on the right hand side) data pre-processing process

We applied Levenstein Distance using fuzzy_wuzzy package to compare keywords (from the clean discharge summary notes as presented in Figure 1) and SNOMED-CT terms in the SNOMED-CT dictionary to calculate pair-wise similarity ratios. We compared all keywords, ranked the similarity ratios, and linked them to ICD-10s using concept id. Table 1 shows the list of the top thirteen predicted ICD-10 mapping results. This case was diagnosed with pneumonia

due to Mycoplasma pneumoniae (J157) and Spastic tetraplegia (G824) which were both matched at the fourth, seventh and thirteenth orders, respectively.

Table 1: SNOMED-CT terms mapped to ICD-10

Concept id	SNOMED-CT terms	ICD-10s	ICD-10 terms
25797006	blood aspiration	P242	Neonatal aspiration of blood
56018004	wheezing	R062	Wheezing
233604007	pneumonia	J189	Pneumonia, unspecified
186464008	mycoplasma infection	A493	Mycoplasma infection, unspecified
125591009	injury pharynx	S198	Other specified injuries of neck
238159008	desaturation blood	P84	Other problems with newborn
46970008	mycoplasma pneumonia	J157	Pneumonia due to Mycoplasma pneumoniae
65520001	ph	E748	Other specified disorders of carbohydrate metabolism
299989006	infection toe	L089	Local infection of the skin and subcutaneous tissue, unspecified
722435003	dystonia 16	G248	Other dystonia
128601007	lung infection	J189	Pneumonia, unspecified
282776008	injury toe	S999	Unspecified injury of ankle and foot
192965001	spastic tetraplegia	G824	Spastic tetraplegia

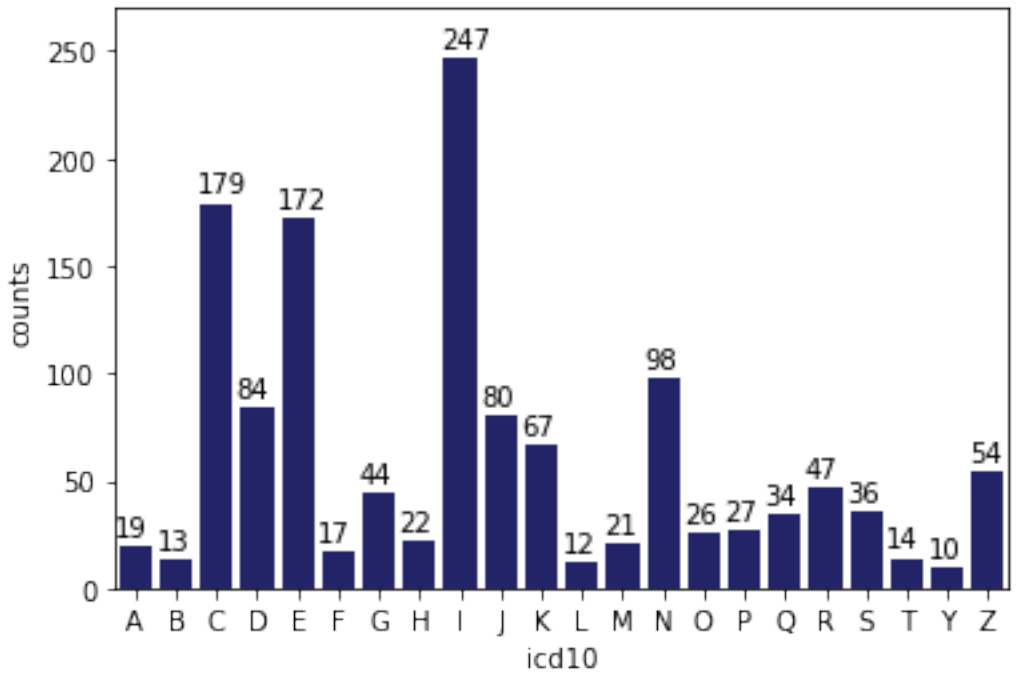


Figure 2: Frequency of ICD-10 terms in each category

Overall, there were 498 actual unique ICD-10s recorded by humans in the 560 discharge summaries. Figure 2 shows the frequency of ICD-10s grouped by ICD-10 categories, sorted from A to Z. The three most common ICD-10 categories are: Diseases of the circulatory system (I), Neoplasms (C), and Endocrine, nutritional and metabolic diseases

(E).

Table 2 shows the distribution of the rank of the actual ICD-10s predicted by character set ratio, Jaccard similarity, and set token ratio. The performance of set token ratio is the best compared with the other algorithms. Fifty percent of the correct predicted ICD-10s were ranked within the top two percent of the distribution.

Table 2: Statistical analysis of the prediction of actual ICD-10s

Algorithms	Median	Average	Max	Min	Std
Character set ratio	93.28	87.57	99.99	15.97	14.08
Jaccard similarity	97.05	86.41	100.00	1.12	23.57
Set token ratio	98.72	92.31	100.00	9.33	13.92

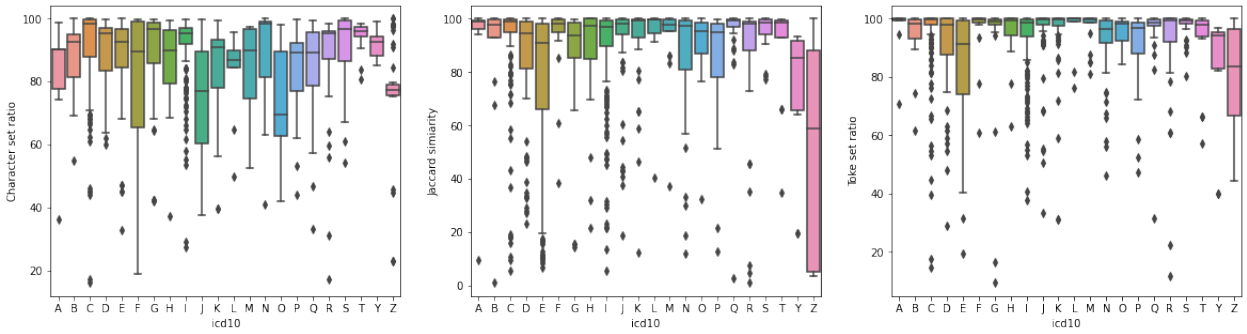


Figure 3: Distribution of predicted ICD-10 rank grouped by ICD-10 categories

Figure 3 shows the distribution of predicted ICD-10 rank grouped by ICD-10 categories. The distribution of prediction performance of set token ratio shows the best performance across all categories compared with the other algorithms. The performance of Jaccard similarity generally has better performance than character set ratio except in Z category. The poor performance of category Z in Jaccard similarity causes low general performance (as presented in the average in Table 2).

Figure 4 shows the rank distribution of actual ICD-10s in each category predicted by set token ratio (x-axis). 485 ICD-10s (97.39%) with the average rank above 50% represent true positive and 13 ICD-10s (2.61%) with the average rank below 50% represent false negative (annotated with ICD-10s in the figure). The false negative instances are the actual ICD-10s that the algorithm did not recognise ICD-10 related terms in the discharge summary notes.

Discussion

Mapping clinical documents to SNOMED-CT or ICD-10 is a challenging task and we have explored how to improve the process by utilising computerised techniques in order to reduce human work load. Our study showed that combining the fuzzy wuzzy approximate matching algorithm and using SNOMED-CT allowed flexibility to map clinical documents to ICD-10 which helped to discover more missing relevant ICD-10s that were not detected by humans in the document. SNOMED-CT terms cover major medical terms that commonly appear in clinical notes and the dictionary helped us to link the terms to related ICD-10s. By applying the approximate matching technique, the target words or phrases did not need to be in the perfect correct order with the same length or spelling in order to be matched. For example, Mycoplasma pneumonia was detected although this was not an exact phrase written in the notes. The flexibility of the automated system has a main advantage in that it has a higher sensitivity than a physician to match related ICD-10 from the discharge summary notes. Typically, a physician only selects a few important ICD-10s because of time limitations while our process is able to fully explore all relevant ICD-10s. This is very helpful for finance reimbursement because it could produce a more complete ICD-10 coding to report to the government, especially, for example, in a case presenting with many complications such as ectopic pregnancy.

In the example shown in Figure 5, a physician using manual coding determined one actual ICD-10 “Other ectopic pregnancy” for this case, while the algorithm detected the correct ICD-10 of “tubal pregnancy” and additionally

Figure 4: Rank distribution of actual ICD-10s in each category predicted by the set token ratio

identified other missing relevant ICD-10s from the discharge summary including pallor, abnormal bleeding, present pain, ovary tender, and history of doxycycline allergy. However, the advantage of flexibility also creates a major disadvantage of recruiting non-related ICD-10s, thus an optimal system needs to be able to distinguish sensible and irrelevant responses. In this case, irrelevant or incorrect ICD-10s such as murmur, single pregnancy, edema, small uterus, normal pregnancy, heart irregular, small ovary, conjunctiva closed, and lung cyst were also matched. We manually analysed the mapping results from all notes and especially 13 false positives. We found that the mapping process worked well when there were SNOMED-CT terms present (but not necessarily in the same order or with an exact match) in the discharge summary. However, there are some major multiple flaws that could be improved in future studies. First, the algorithms were unable to comprehend abbreviations that were not internationally used and not pre-documented in the mapped SNOMED-CT. For example, referring to "ICD-10: O820, Delivery by elective caesarean section", our results show that the algorithm incorrectly matched this discharge summary to an ICD-10 code when the physician noted C/S in the discharge summary instead of spelling out "caesarean section". Second, the

case female 37 years old known case rt tubal pregnancy g2 p0010 ga 6 wklmp cc bhcg day 4 pi known case rt tubal pregnancy g2 p0010 ga 6 wklmp present 201259 pain score 410 upt positive anemic symptom tvs rt tubal pregnancy 09111 cm subserous myoma 2124 2224 anterior wall free fluid bhcg 16922 tx mtx single dose bhcg 28468 gyne hx lmp 101159 pmp 11059 duration 3 day interval 3 padday irregular cycle last pap 2 yrs neg contraception ob hx g2p0010 last 10 years old past hx ud medication drug allergy penicillin doxycycline previous operation appendectomy 20 years old dx ruptured appendix vs stable heent mild pale conjunctiva heart regular murmur lung clear equal breath sound abd soft tenderness rebound tender negative mass ext edema neuro motor sensory intact pv 281159 iub normal vg normal mucosa cx os closed bleeding adx free mass tender investigation cbc hbhct 126359 wbc 6620 n648l299 plt 277000 bhcg 28468 miuml tvs 2 small intramural myoma 15 cm anterior wall seen intramural anechoic cintent 04713 cm pseudosac rt corpus luteal cyst 151212 cm rt tube swelling inhomogenous content suspected blood clot gestrational sac seen seperate rt ovary minimal free fluid 08 cm rt tubal pregnancy sp mtx suspected fail mtx 1st dose tvs 2 small intramural myoma 15 cm anterior wall seen intramural anechoic cintent 04713 cm pseudosac rt corpus luteal cyst 151212 cm rt tube swelling inhomogenous content suspected blood clot gestrational sac seen seperate rt ovary minimal free fluid 08 cm seen sac rt adexa fu 1160 bhcg day 7 26552 67 us bedside empty uterus free fluid mx repeat dose mtx 75 mg im dc fu b hcg 4160 opd 3

Figure 5: Ectopic pregnancy case

algorithm did not weigh the importance of term sequences. If a term that was placed in the beginning of a paragraph could meaningfully match with terms from later in the document, the algorithm combined those terms to form ICD-10 match. For example, Figure 6 shows a document with an incorrect ICD-10 mapping of IUD contraception instead of malignant neoplasm of the endocervix.

case female 43 yr p2002 last 20 yr active si contraception ud htn dlp amlodipine cc khown case hsil sp leep ii pi 1 yr pta nv checkup pap smear 24859 asch colpo dwe 34 612 hsil leep c ecc endocervix positive hsil colpo 251059 unsat seen lesion leep endocervix positive dysplastic epithelium hsil cannot excluded colpo 61259 unsat twe 39 fine punctation 39 imp persistent hsil advice tah accept ph ud ht dlp amlodipine10 11 opc obgyn hx p2002 last 20 yr lmp 251259 3 days pmp 281159 pe vs bt 366 pr 80 bpm rr 08min bp 12076 mmhg adb soft tender palpable mass pv 241259 iub wnl vg normal mucosa cx normal withish discharge ut ns adexa mass pv 281259 iub wnl vg minamal bloody discharge cx postleep inactive bleeding per os ut ns free av adexa mass imp hsil sp leep marginal cannot excluded hsil management set tah bilat salphingectomy 291259 finding normal size uterus smooth endometrium normal cervix normal tubes ovaries postop complication 1160

Figure 6: Malignant neoplasm of endocervix case

The algorithm combined "Contraception" with "iub" and then converted "b" in "iub" to "D" according to Levenstein Distance and considered that these two words were similar. The correct ICD-10, malignant neoplasm of endocervix (C530), was not detected in a top ranking position in this case because of the abbreviation problem. HSIL stands for High grade squamous intraepithelial lesion (R87), a type of malignant neoplasm of the endocervix which was not fully named in the document. Regarding the abbreviation problem, we did try to apply Allie: a database of abbreviation[16] to convert abbreviations in the discharge summary notes to full text. We found that by applying the database to directly map abbreviations to full text created a lot of incorrect mappings. For example, "PE" was commonly noted in the document as "Physical Examination" but "PE" also mean "Pulmonary Embolism". Physical examination has less meaning to ICD-10 mapping in opposite to pulmonary embolism. Therefore, the automatic direct conversion trended to convert PE to pulmonary embolism which was mostly incorrect. Human doctor is able to distinguish between those two terms by considering surrounding context. Therefore, we might need to add this factor to achieve better conversion performance in the future work.

A third challenge was that negative findings are commonly reported in a physician's discharge summary notes. Typical negative findings such as no palpable mass, no pale conjunctiva, no jaundice, no hepatosplenomagaly were generally found in the notes; the algorithms were not sophisticated enough to comprehend the negated phrases. Negative findings of relating to murmur, single pregnancy, edema, small uterus, normal pregnancy, heart irregular, small ovary, conjunctiva closed, and lung cyst in the ectopic pregnancy case were all detected as false positives. The same pattern of false positive was also found in the malignant neoplasm of endocervix case with "Bleeding" identified while the document clearly noted "inactive bleeding".

Lastly, the algorithms were not able to distinguish past, present and future when noted in the discharge summary. For example, a patient presented at a hospital with dyspnea on exertion and had a history note of acute myocardial infarction and ruptured appendicitis 10 years ago. Although the events of appendicitis and myocardial infarction were already completed in the past, the algorithm still comprehended both diseases as present problems which will lead to ICD-10 mismatching.

In conclusion, the results of this study show that mapping clinical documents by using approximate matching and SNOMED-CT has the potential to be used to screen keywords to map relevant ICD-10s, although there are some clear challenges still to be addressed. Prior research showed that using advanced machine learning methods on mapping ICD-10s from clinical text have shown the significant improvement[17]. In the future work, we aim to apply the machine learning methods and compare the results with the method in this paper.

Conclusion

Using approximate matching and the SNOMED-CT dictionary to map ICD-10 from discharge summary notes shows potential for use case for improving the intensive documentation workload in a health-care system. Although there are still some limitations of this approach, we may consider continuing doing further research to improve this approach, including pre-processing common abbreviation terms, utilising term frequency-inverse document frequency or TF-IDF to prioritise the importance of keywords, weighing the importance of word sequencing to minimise the effects of negative findings, and applying advanced machine learning methods.

Acknowledgement

This work was supported by the Student Affair, and IT Department, Faculty of Medicine, Chang Mai University. We would like to thank the Tilley family for assistance in editing and reviewing the manuscript.

References

[1] WHO — International Classification of Diseases (ICD) Information Sheet.

[2] Jan Horsky, Elizabeth A. Drucker, and Harley Z. Ramelson. Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2017:912–920, 2017.

[3] Fu Wen Liang, Liang Yi Wang, Lin Yi Liu, Chung Yi Li, and Tsung Hsueh Lu. Physician code creep after the initiation of outpatient volume control program and implications for appropriate ICD-10-CM coding. *BMC Health Services Research*, 20(1):1–7, feb 2020.

[4] Rosy Tsopra, Daniel Peckham, Paul Beirne, Kirsty Rodger, Matthew Callister, Helen White, Jean Philippe Jais, Dipansu Ghosh, Paul Whitaker, Ian J. Clifton, and Jeremy C. Wyatt. The impact of three discharge coding methods on the accuracy of diagnostic coding and hospital reimbursement for inpatient medical care. *International Journal of Medical Informatics*, 115:35–42, jul 2018.

[5] Rikinkumar S. Patel, Ramya Bachu, Archana Adikey, Meryem Malik, and Mansi Shah. Factors related to physician burnout and its consequences: A review, oct 2018.

[6] Shelagh McRae and Robert Hamilton. The burden of paperwork. *Canadian family physician Medecin de famille canadien*, 52(5):586, 588, may 2006.

[7] Dennis Lee and Francis Lau. Exploratory reverse mapping of icd-10-ca to snomed ct. volume 410, 01 2008.

[8] Holger Stenzhorn, Edson José Pacheco, Percy Nohama, and Stefan Schulz. Automatic mapping of clinical documentation to SNOMED CT. In *Studies in Health Technology and Informatics*, volume 150, pages 228–232. IOS Press, 2009.

[9] Thai health information standards development center (this) snomed ct download.

[10] Thai health information standards development center (this).

[11] SNOMED - Home — SNOMED International.

[12] Kathy Giannangelo and Jane Millar. Mapping SNOMED CT to ICD-10. In *Studies in Health Technology and Informatics*, volume 180, pages 83–87. IOS Press, 2012.

[13] V. I. Levenshtein and V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *SPhD*, 10:707, 1966.

[14] GitHub - seatgeek/fuzzywuzzy: Fuzzy String Matching in Python.

[15] Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912.

[16] Y. Yamamoto, A. Yamaguchi, H. Bono, and T. Takagi. Allie: a database and a search service of abbreviations and long forms. *Database: The Journal of Biological Databases and Curation*, 2011, 2011.

[17] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Trajectory Inspection: A Method for Iterative Clinician-Driven Design of Reinforcement Learning Studies

Christina X. Ji, MEng[1*], Michael Oberst, MS[1*], Sanjat Kanjilal, MD, MPH[2,3], David Sontag, PhD[1]
[1]MIT CSAIL, IMES, Cambridge, MA; [2]Harvard Medical School, Boston, MA; [3]Harvard Pilgrim Healthcare Institute, Boston, MA; * Equal Contribution

Abstract *Reinforcement learning (RL) has the potential to significantly improve clinical decision making. However, treatment policies learned via RL from observational data are sensitive to subtle choices in study design. We highlight a simple approach, trajectory inspection, to bring clinicians into an iterative design process for model-based RL studies. We identify where the model recommends unexpectedly aggressive treatments or expects surprisingly positive outcomes from its recommendations. Then, we examine clinical trajectories simulated with the learned model and policy alongside the actual hospital course. Applying this approach to recent work on RL for sepsis management, we uncover a model bias towards discharge, a preference for high vasopressor doses that may be linked to small sample sizes, and clinically implausible expectations of discharge without weaning off vasopressors. We hope that iterations of detecting and addressing the issues unearthed by our method will result in RL policies that inspire more confidence in deployment.*

Introduction

Reinforcement learning (RL) has emerged as a popular tool for trying to learn effective policies for managing sequential decisions in patient treatments, including applications in sepsis treatment[1–3], ventilation weaning[4], acute hypotension treatment[5,6], and HIV[7]. RL has the potential to provide significant support in clinical decision making. However, in using purely retrospective data, these approaches inherit the usual challenges of learning and evaluating dynamic treatment regimes from observational data, a well-studied topic in epidemiology and bio-statistics[8,9]. For instance, evaluation methods typically assume that all confounding variables are included in the model, and that either existing clinical practice or the dynamics of patient health can be accurately modelled. Just as in any observational study, if the assumptions do not hold, the analysis can lead to misleading results that could adversely affect clinical practice if adopted. Small effective sample sizes and inadequate specification of the clinical outcome or available actions can further exacerbate the problem[10,11]. Research has been ongoing to overcome these challenges[5,6,12]. Because these problems are nuanced, model-checking and interpretability techniques are needed to detect these issues and allow clinicians to actively engage in an iterative process of study design. We use the term "study design" to emphasize that this goes beyond choosing a statistical model and encompasses all the design choices involved in translating a clinical decision problem into a RL problem that produces clinically sensible results. By iteratively detecting and fixing these problems until the study is clinically sound, the RL algorithm will be more robust for deployment.

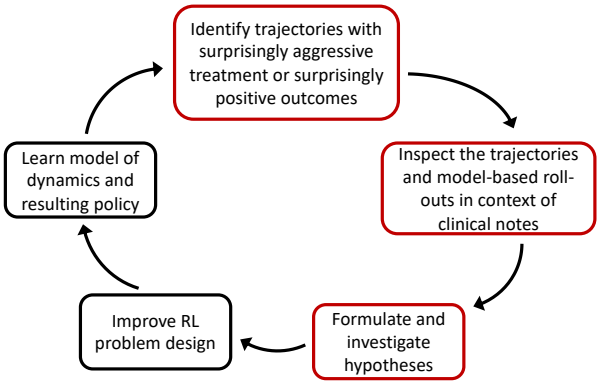


Figure 1: Our proposed method (red boxes) can be integrated into this workflow for improving model-based RL.

In this work, we highlight a simple approach for examining learned models and policies in model-based RL, a common technique used in recent work to improve clinical decision-making in the management of sepsis[1,3]. Our workflow is shown in Figure 1 and proceeds as follows: (i) Select patient cases using one of two heuristics: First, we identify patient states where the learned policy suggests far more aggressive treatment than the observed standard of care, and second, we find patients where the learned policy is predicted to dramatically out-perform the current standard. We hypothesize that both cases may be due to flaws in the study design. (ii) Contrast the observed trajectories of these patients with the predicted trajectories under the learned model and policy, and flag suspicious model behavior by also

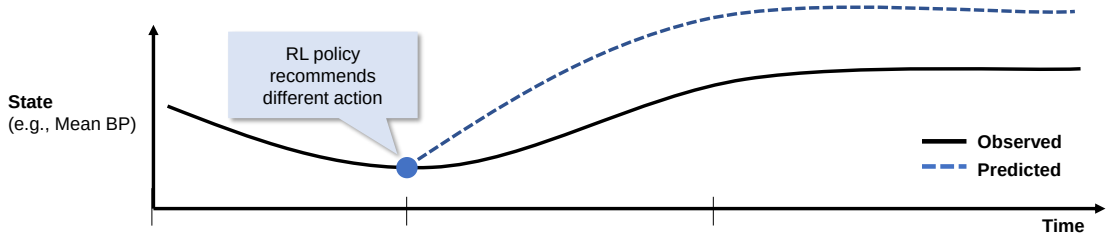


Figure 2: Conceptual illustration of a patient trajectory (in black) and a model-based roll-out (in blue) which tracks what the RL model predicts would occur as a result of the actions considered 'optimal' by the learned policy.

comparing predictions against the actual clinical course of patients documented in the medical record. (iii) Investigate questions and hypotheses raised by suspicious model behavior. The code for reproducing our work is located at `https://github.com/clinicalml/trajectory-inspection`.

We demonstrate the utility of this approach by applying it to recent work published in *Nature Medicine*[1] that we replicate using MIMIC-III, a freely available ICU dataset[13]. We generalize our anecdotal observations through aggregate analysis and discover that the data pre-processing heuristic used in recent works on RL for sepsis management[1,2] can lead to biases in estimating the model. We also observe that high recommended vasopressor doses may be linked to small sample sizes, and that the model has clinically implausible expectations, such as patient discharge from the ICU within 4 hours of receiving large doses of vasopressors.

Background

Reinforcement learning methods: The approach that we replicate[1] learns to manage sepsis using RL[14,15]. The framework is a *Markov decision process (MDP)*. This model specifies a set of possible states and actions called the *state space* and *action space*, respectively. The state contains patient features, such as demographics, vital signs, and prior treatment. The action corresponds to a treatment decision. At each time step, an action is taken based on the current state, and the patient transitions to a new state. This transition model captures how the patient responds to treatment and evolves over time. The *Markov property* states that this transition depends only on the most recent state and action and is otherwise independent of history. The *reward function* specifies how each transition maps to a positive or negative outcome. Each sequence of states and actions and the reward forms a *trajectory* and represents the course of the patient stay. A *policy* specifies which action to take from each state. This can be a *behavior policy* observed in the training samples or a more optimal policy. The goal is to learn a policy that maximizes the reward.

Model-based trajectories: In *model-based* RL, a transition model and reward function are estimated, and the policy is learned using those estimates.[a] We can generate *model-based trajectories* (sometimes referred to as "roll-outs") given an initial state, a transition model, and a policy. These trajectories are predictions, made by the model, for how a patient would progress under the policy. Concretely, at each time step, an action is drawn from the distribution specified by the policy for the current state. Then, the next state is drawn from the distribution specified by the transition model for that state-action pair. This process repeats until an absorbing state (with a reward) or a maximum length of 20 steps is reached. Because the transitions and sometimes the policy are probabilistic, multiple different model-based roll-outs can be generated from the same initial state. A conceptual illustration is shown in Figure 2.

Notable challenges with off-policy evaluation: Just as with any observational study, confounding and other biases can lead to unreliable estimates of the value of a policy. As a conceptual illustration, imagine that all patients are either "healthy" or "sick", that doctors tend to aggressively treat sick patients, and that sick patients have higher mortality rates than healthy patients, even with aggressive treatment. In this setting, if we do not adjust for comorbidities and severity of the presenting illness (by including it in the state space of our model), then our evaluation might wrongly conclude that "never treating" patients is a good policy, due to the association between treatment and mortality.

Indeed, this phenomenon has been observed in the literature on applying RL to sepsis management. In some cases, it has been noted that a "zero-drug" policy has a high estimated value relative to current practice, using the same evalua-

[a] Alternatively, *model-free* RL still relies on modelling existing practice, but no explicit transition model is learned.

tion methodologies that are used to justify the value of RL policies[1,16,17]. In other instances, the RL policy associates the more intensive treatments observed for high acuity patients with high mortality and instead recommends less intensive treatments that are rarely observed[10]. Possible unmeasured confounding is not the only hurdle to performing valid inference. Other challenges include small effective sample sizes and inappropriate reward definitions[11].

Related work: We differentiate our approach from recent work in counterfactual off-policy evaluation[18]. While that work also involves simulating from a model and has a similar heuristic for selecting trajectories, it requires additional causal modelling assumptions, and we view our approach as simpler to apply. More importantly, we demonstrate our approach on a real dataset and model, while that work only considered a synthetic dataset.

Methods: Setup

Main outcome measurement: Our primary objective is necessarily qualitative, demonstrating that our approach can help researchers better understand the challenges in designing RL studies and give them a concrete tool to integrate clinician input into improvement of study design. That said, we do provide quantitative metrics related to our trajectory selection heuristics, e.g., what it means for a recommended action to be surprisingly aggressive, or a model-based roll-out to have a surprisingly positive outcome (see Figure 3). We also compute several diagnostic metrics during our investigation of specific questions raised by trajectory inspection.

Design/patient cohort: Our work is a secondary use of electronic health record data, specifically the MIMIC-III dataset[13]. As we are replicating Komorowski et al (2018)[1], we use their code[b] for defining the sepsis cohort: adults fulfilling the sepsis-3 criteria[19,20], including antibiotic prescription, lab work for signs of infection, and a sequential organ failure assessment (SOFA) score ≥ 2. Patients with missing fluids or mortality data are excluded. Patients who satisfy all of the following conditions are also excluded as they may have been placed on comfort measures: (a) died within 24 hours of the end of the data collection window, (b) received vasopressors at any point, and (c) whose vasopressors were stopped at the end of the data collection. Our cohort consists of 20,090 ICU admissions.

Setup of Komorowski et al (2018)[1] The state space includes demographics (e.g., age), vital signs (e.g., blood pressures, sequential organ failure assessment), lab values (e.g., white blood cell count, creatinine), ventilation parameters (e.g., FiO2), and treatment information (e.g., fluid balance and output aggregated over 4-hour intervals). The policy takes in these state features and gives a recommendation on vasopressors and fluids. We follow the publicly available code provided by the authors and use the same version of the MIMIC-III dataset[13] in our replication. In this section, we review some details of the set-up, focusing on modelling decisions. Other details can be found in the original paper.

Trajectories are discretized into 4-hour intervals and limited to a maximum of 20 time steps, representing up to 28 hours before and 52 hours after onset of sepsis[c]. Starting from a set of continuous and discrete variables[d], the state space is discretized using k-means clustering with 750 clusters, and two absorbing states are added for 90-day survival and 90-day mortality (including in-hospital mortality). Model actions are limited to providing blood pressure support through intravenous fluids or adjusting vasopressors. The action space contains 25 discrete choices, corresponding to either zero or one of four quartiles of total fluid input and maximum vasopressor dosage over each 4-hour time step. Rewards are defined as ± 100 depending on 90-day mortality and are recorded in the retrospective trajectories as a transition into an absorbing state described above. After replicating this work, we have a transition model, a behavior policy, and a target policy, learned from observational data.

Methods: Selecting trajectories for inspection

As previously described, our approach relies on inspecting clinical trajectories alongside their model-based roll-outs. We propose two heuristics for selecting trajectories: (i) *Surprisingly aggressive treatments*: We find states where the RL-recommended action is rarely observed and more aggressive than the observed practice of clinicians. We then start the roll-outs where those states occur in trajectories. (ii) *Surprisingly positive outcomes*: We select clinical trajectories with initial states that have the largest differences between predicted and actual outcomes.

[b] `https://github.com/matthieukomorowski/AI_Clinician`

[c] The paper reports using 24 hours before and 48 hours after. Our implementation follows their publicly released code, which uses 28 and 52.

[d] Our replication follows the publicly available code, which uses 47 (excluding Elixhauser Index) state variables instead of the stated 48.

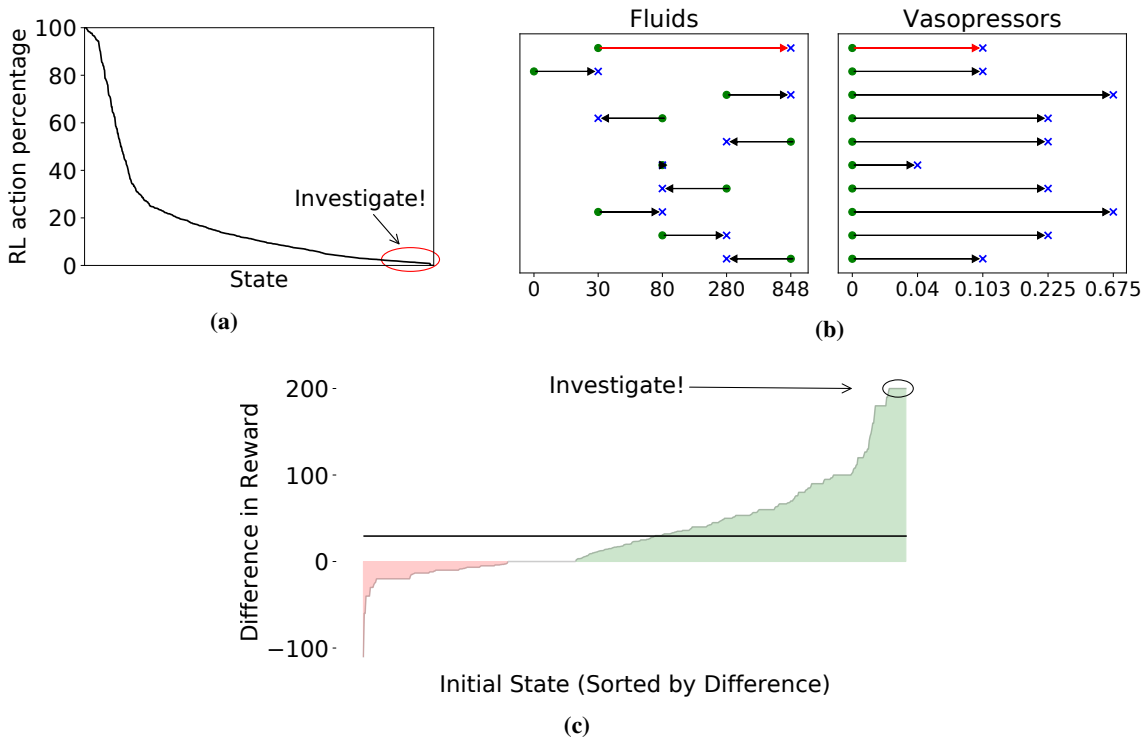


Figure 3: Visualization of our heuristics for trajectory selection. (a) Observed frequency of the RL action for each state in the training data. States are sorted by this quantity on the X-axis. (b) Differences between the RL action (blue cross) and the action most commonly observed in the training data (green dot) from the 10 states whose RL action was rarest. The X-axis is total milliliters over the past 4 hours for fluids and micrograms per kilogram of body weight per minute for vasopressors. Each row is a state, with the one we inspect in red. (c) Differences in predicted outcomes. X-axis sorts states by the average difference between model-based and observed rewards for trajectories with that initial state. The black line denotes the overall average difference weighted by initial state.

Selecting states with surprisingly aggressive treatments: To select trajectories, we identify instances where the learned policy recommends aggressive treatment even though conservative management is observed. Concretely, across all states, we first look for actions recommended by the learned policy that occur at most 1% of the time in the training data at that state, as visualized in Figure 3a. This yields 15 states, and we plot the differences between the most common clinician action and the RL action of the top 10 in Figure 3b. We analyze a state where the patient receives significantly higher amounts of fluids and vasopressors under the RL policy as compared to standard clinical practice. The clinical parameters that characterize this state are mostly within normal limits except for a mildly elevated white blood cell count. The most common action performed by the clinicians for patients in this state was to provide 30cc of fluids, which is suggestive of a very low infusion necessary to maintain the patency of the IV line, and no vasopressors. In contrast, the RL policy recommends 848cc of fluids and 0.13 micrograms/kg/min of vasopressors. Of the 632 times the state is observed in the training data, the common action noted above is taken 161 times, while the 'optimal' action according to the learned RL policy is chosen only 6 times by clinicians. For these trajectories, we start the model-based roll-outs at the time step where the selected state and common action occurred.

Selecting initial states with surprisingly positive outcomes: To select trajectories under this heuristic we identify patients where the model expects that the new policy will most out-perform the observed current practice. We hypothesize that by focusing on extremes, we will enrich for instances that highlight problems with the RL model. Concretely, we create model-based roll-outs for each actual trajectory in the test set, starting from the observed initial state. We then select trajectories with substantially higher model-based reward than the actual reward. In particular, for each state, we examine all trajectories that start in that state and take the mean difference between (a) the reward of 5

Transferred in from outside hospital after undergoing cardiac catheterization that revealed coronary artery disease. She [...] was brought to the operating room for **coronary artery bypass graft surgery**. [...] She was **transferred to the intensive care unit for post operative management**. In the first twenty four hours she was weaned from sedation, awoke neurologically intact, and was **extubated without complications. She was started on betablockers and gently diuresed toward preoperative weight. On post operative day one she was transferred to the floor.** Chest tubes and pacing wires were discontinued without complication. [...] By the time of discharge on POD 5 the patient was ambulating freely, the wound was healing and pain was controlled with oral analgesics. **The patient was discharged to home in good condition**

(a)

Pt had Septic physiology on admission [...] and the most likely source was felt to be postobstructive pneumonia and parapneumonic effusion [...] Pt was intubated [...] [and] R-sided chest tube initially drained a large volume of mucinous fluid [...] noncontrast chest CT showed a RLL rounded opacity [...] CT chest with contrast showed the RLL opacity was a large mass with significant surrounding lymphadenopathy and marked pleural tumor, with possible invasion into the chest wall [...] **pt was ultimately diagnosed with stage IIIA lung cancer.** Pt's progressive respiratory failure was treated with BiPAP and nebs [...] **Pt ultimately died of respiratory failure likely due to a combination of COPD, pneumonia and lung cancer**

(b)

Figure 4: Selected extracts from the de-identified medical record for Patient 1 (a) and Patient 2 (b). Emphasis added. Acronyms: POD, post-operative day. Pt, patient. RLL, right lower lobe. BiPAP, Bilevel Positive Airway Pressure

model-based roll-outs and (b) the observed reward. We then inspect trajectories for which this difference is largest. In Figure 3c, we visualize this difference across individual states and highlight some initial states for inspection. We choose one of the highlighted initial states, where 90-day mortality is observed in the actual data, but all model-based roll-outs (from the same initial state) end in 90-day survival. We investigate a clinical trajectory with this initial state.

Results: Examining trajectories alongside the medical record

In this section, we present two patient trajectories, selected using the heuristics described above. In both cases, we examine the full set of clinical notes available for the patient stay, along with their actual trajectories of vital signs. We compare this against the model-based trajectories predicted by the RL model under the learned RL policy. We can ask *whether these model-based predictions seem clinically sensible*, particularly in light of the notes.

Clinical review of patient 1: Surprisingly aggressive treatment: We present a snippet from the de-identified medical record in Figure 4a, and in Figure 5a, we present selected vital signs, along with the actions. We have the following takeaways from the notes: (i) *Cause of admission:* This patient presented with shortness of breath and chest pain after a previous visit had revealed coronary artery disease. As a result, she underwent coronary artery bypass graft (CABG) surgery and was transferred to the ICU for post-operative management on the same day as the surgery. (ii) *Treatment during ICU stay*: Around 12 hours into the ICU stay, radiology notes indicated signs of pulmonary edema. After 24 hours, the patient was recovering well, weaned from sedation, and extubated without complications. The patient received cardio-protective beta blockers on post-operative day 1, suggesting that she was in a stable condition from a hemodynamic standpoint. Stability was further evidenced by the fact that diuretics were used to gently remove fluids and bring her volume status back to pre-surgery levels. The patient was transferred out of the ICU and sent home without pulmonary edema and in good condition on post-operative day five.

In light of these points, we make the following observations regarding the trajectories in Figure 5a: (i) *The RL policy suggests unnecessary prolonged use of aggressive treatments.* In the actual trajectory, clinicians initiated IV fluids and vasopressors approximately 8 hours into the ICU stay and in the immediate post-operative period. By the time we start model-based roll-outs, 20 hours into the ICU stay, the amount of fluids and vasopressors had already been greatly reduced or completely discontinued given the patient's uncomplicated recovery.[e] Despite the benign hospital course for the patient, the RL policy recommended restarting larger dosages of both fluids and vasopressors, which would likely have increased risk of pulmonary edema and fluid overload. (ii) *Expected patient discharge while on vasopressors is clinically implausible*: In two of the model-based trajectories, the model anticipates that the patient will leave the ICU within 4 hours of giving vasopressors at the start of the roll-outs, and in most of the model-

[e] All the time steps are 4 hours apart, except there are 20 hours between steps 6 and 7.

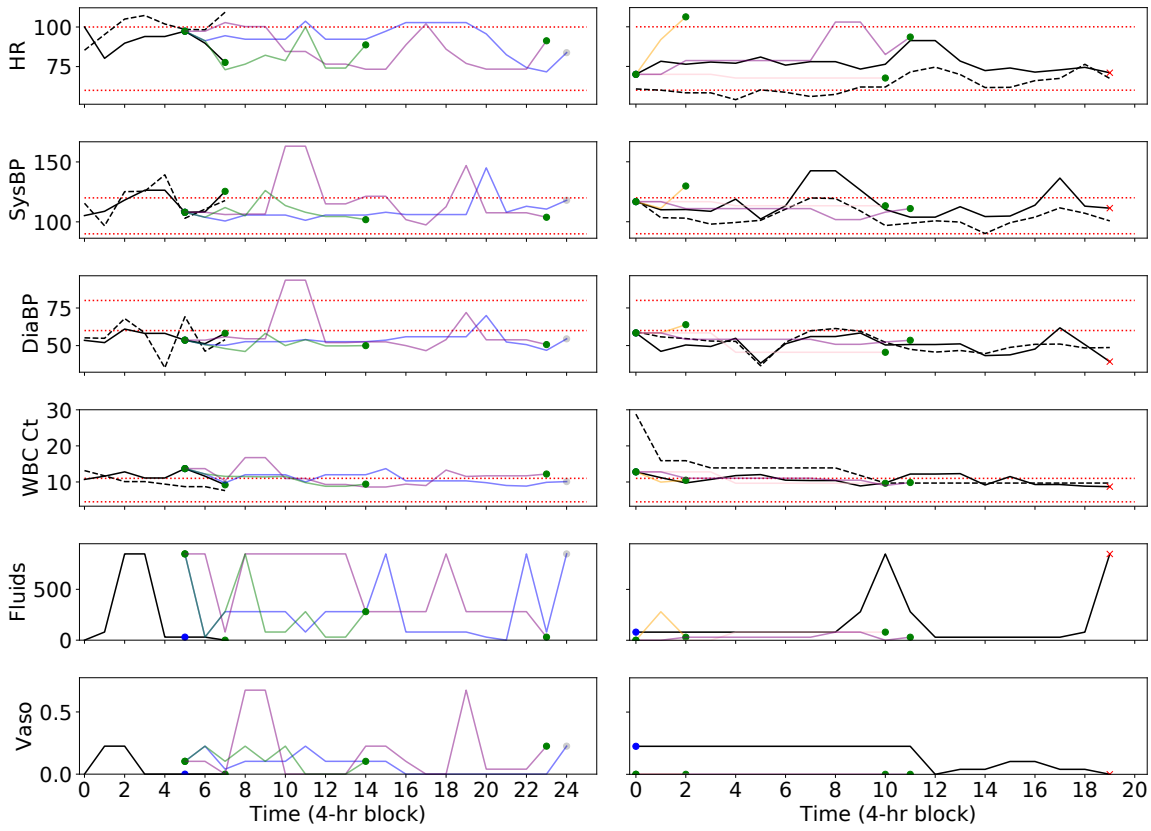


Figure 5: Comparison of actual trajectories and model-based roll-outs for patient 1 (a) and patient 2 (b). Five model-based trajectories per patient start at the blue dots: In (a), two of these end when they start at time step 5, and in (b), two end at time step 0. Model-based roll-outs are in various colors, and vitals are derived from the k-means medians. At the ends of the trajectories, green dots indicate 90-day survival, red crosses indicate 90-day mortality, and grey dots indicate maximum allowed length without discharge. Black dotted lines show actual values, while black solid lines denote median values from k-means clustering. Red dotted lines are reference ranges. HR: heart rate. SysBP: systolic blood pressure. DiaBP: diastolic blood pressure. WBC Ct: white blood cell count. Vaso: vasopressors. Fluids dosage in total milliliters over the past 4 hours and vasopressor levels in micrograms per kilogram of body weight per minute.

based trajectories, the patient is discharged while on vasopressors. This goes against clinical intuition that a patient on vasopressors should be weaned off and monitored prior to leaving the ICU. (iii) *Response in state variables is inconsistent with action.* The effect of vasopressors becomes apparent within half an hour, as evidenced by how blood pressure rose at time step 2 following a dosage at the previous time step in the actual trajectory. In the model-based trajectories, however, vasopressors are administered at time step 5, but all of the trajectories show little change in blood pressure at the next time step. For the model-based roll-out indicated in purple, 0.675 micrograms of vasopressors per kilogram of body weight per minute were administered at 32-36 hours, which is even higher than the doses for the other model-based trajectories. However, blood pressure does not rise until 40-44 hours. This may indicate that the model is not accurately modelling the drug response. The rationale for the RL policy is also unclear given the lack of a clear indication for blood pressure support.

Clinical review of patient 2: Surprisingly positive outcomes: We present a selected snippet from the medical record in Figure 4b, and in Figure 5b, we present selected vital signs, along with the actions. We summarize the major take-aways here from the full medical record: (i) *Cause of admission:* This patient was admitted after collapsing, thought secondary to either respiratory or cardiac failure. The patient was taken immediately to the cardiac catheterization lab[f],

[f] A cath lab is an exam room with diagnostic imaging equipment to visualize the heart.

where a myocardial infarction due to coronary artery disease was ruled out. Chest imaging showed a large amount of fluid around the right lung and a large mass in the lower right lobe. This was later discovered to be Stage IIIA lung cancer[g]. The sum of the diagnostic studies suggested that the most likely etiology of the patient's presentation was cardiovascular collapse[h] and a secondary post-obstructive pneumonia that were both due to the mass effect of the tumor. (ii) *Treatment before and during ICU:* The pleural effusion was felt to be multifactorial and likely due to poor forward flow as well as inflammation from the adjacent pneumonia[i]. A chest tube was placed, which subsequently drained $>$1L of exudative serous fluid. The patient's clinical status responded rapidly, suggesting the external compression from the fluid was a major contributor to his course. Upon transfer to the ICU, physicians continued to administer vasopressors and antibiotics, with the former being gradually weaned starting 44 hours into the trajectory. (iii) *Cause of death:* Despite the placement of a chest tube, the underlying problem of a large lung mass leading to cardiovascular compromise remained unaddressed. In part due to the morbidity of the necessary chemotherapy, the providers, the patient, and the family decided that further aggressive interventions would not have been in the patient's interests and he was made 'comfort measures only' 12 days after the end of this trajectory. He passed away shortly thereafter.

In light of these points, we make the following observations regarding the visualization in Figure 5b: (i) *Surprisingly early termination of several trajectories:* Two of the model-based trajectories end after the first 4 hours, and another ends after the first 12 hours. However, the average length of an ICU stay is 3.3 days[21]. (ii) *The anticipated outcomes are not credible given the medical record*: The cause of death in this patient was irreversible lung damage caused by Stage IIIA lung cancer and pneumonia. As such, the only interventions that would have resulted in survival would have been aggressive chemotherapy, careful cardiovascular support, and a short course of antibiotics. Restrictive fluid and vasopressor therapy (the RL policy) on its own would be very unlikely to have a major salutary effect on the clinical course. Yet, all model-based trajectories result in subsequent 90-day survival.

Results: Investigation of questions raised by trajectory inspection

Although our anecdotal analysis from investigating specific trajectories does not prove that something is fundamentally amiss with the model, the examples serve as inspiration for improving study design. We examine the following three questions suggested by the previous section: (i) *Why do trajectories seem to end early?* For patient 1, we observed that two of the five model-based roll-outs ended within 4 hours. Similarly for patient 2, we observed several trajectories that ended surprisingly early. This may suggest that the model is overly optimistic about how quickly patients will be discharged. (ii) *Why does the model learn to use uncommonly high vasopressor doses?* The RL-recommended aggressive treatment of patient 1 did not appear to have a clinical basis. Indeed, while selecting trajectories via our "surprising treatment" heuristic, we observed in Figure 3b that these RL-recommended actions tend to involve higher levels of vasopressor support. (iii) *Why does the model learn to expect discharge while on vasopressors?* In several model-based roll-outs for patient 1, the model predicts discharge (and positive outcomes) to occur while the patient was still on vasopressors. In the remainder of this section, we move from these anecdotal observations to a more thorough analysis across the entire dataset and seek to understand the answers to these questions.

The model is biased towards early discharge, in part due to censoring: We seek to demonstrate that early predicted discharge is common and due to the transition model rather than the RL policy. To do so, we compare the lengths of training trajectories and model-based roll-outs under the *behavior policy* (the actions taken by clinicians) in Figure 6. Because the transition model and behavior policy are derived from the training data, we should expect close matches in the distribution of trajectory lengths if the model is a good fit to the data. Instead, we make the following observations: First, a large proportion of trajectories in the training data are "censored", i.e., they do not end within the defined time window. Second, shorter trajectories are indeed more common in the model-based roll-outs, where the distribution of lengths resembles a regular decay rather than the "bell shape" observed in the training data.[j]

Third, we argue that early termination of model-based roll-outs arises from the heuristic combination of (i) censoring of trajectories after 20 time steps and (ii) the addition of a pseudo-transition at the end of each censored trajectory to

[g]Stage IIIA lung cancer is defined as spread to nearby lymph nodes, but not other organs.

[h]Cardiovascular collapse is the rapid or sudden development of cardiac failure. This was a symptom that resulted from the tumor.

[i]The pleural space is the thin fluid-filled space between the two membranes around each lung

[j]While clear when examining the distribution of trajectory lengths, this model mismatch may not be as obvious if one only looks at average length of trajectories, as done in the "Goodness of fit of the transition matrix" analysis in follow-up work[17] to the model we replicate

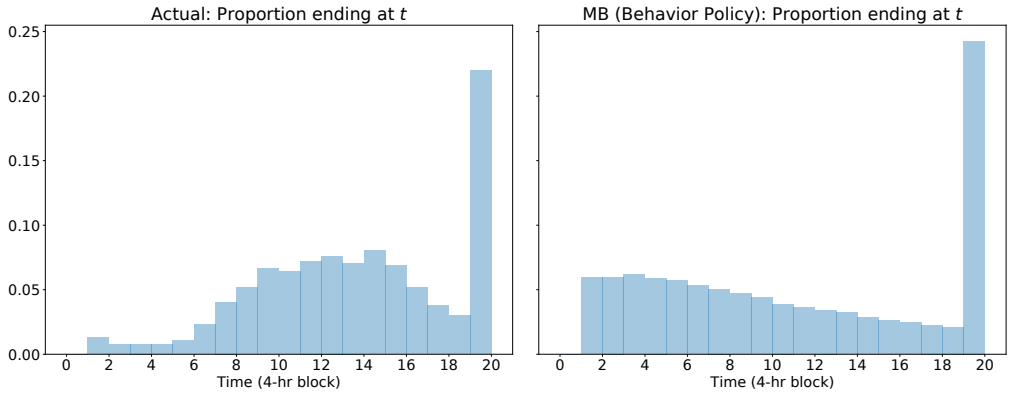
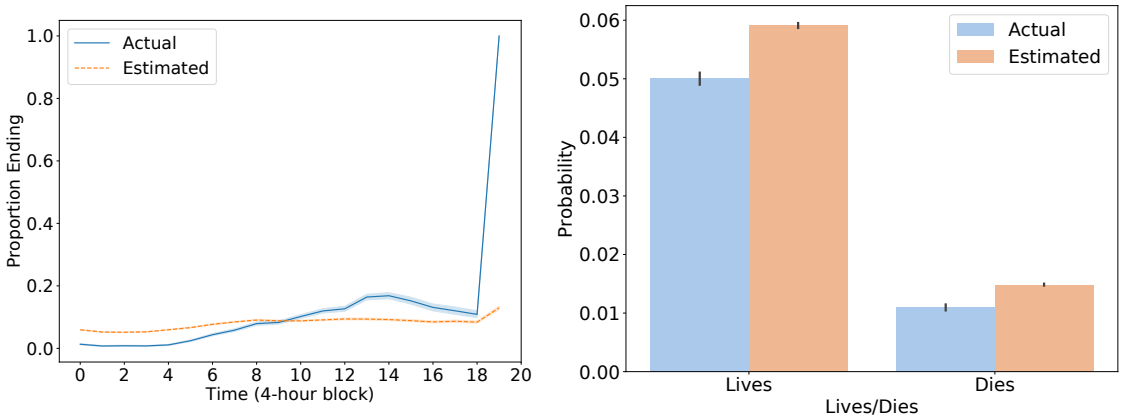


Figure 6: Histogram of trajectory lengths, comparing the training data on the left with the model-based roll-outs under the behavior policy on the right from the same distribution of initial states.



(a) **(b)**

Figure 7: (a) At each time point, the actual proportion of remaining trajectories that immediately terminate (in blue) vs the average predicted probability of immediate termination (in orange). The latter is calculated using the learned transition model and the states and actions that are observed at that time step. (b) Across all time points before the 20th step, we compute the actual and estimated average probabilities of immediate termination. For both (a) and (b), 95% confidence intervals are generated with 1000 bootstrapped samples using the seaborn package in Python.

either the 90-day survival or mortality absorbing state. We refer to this heuristic as **censoring with terminal rewards**. To demonstrate that this heuristic leads to a bias towards discharge in the learned transition model, we focus on the transition to an absorbing state (corresponding to end of trajectory with 90-day survival or 90-day mortality). Figure 7a compares the mean estimated probability of trajectory termination to the observed proportion at each time step. We note in particular that the model cannot capture the fact that at the final time step *every remaining trajectory will end*, due to the censoring with terminal rewards heuristic. To match the overall termination probability it observes, the model is forced to systematically over-estimate the probability of transition to an absorbing state in the time steps prior to the 20th one as seen in Figure 7b. If we interpret the transition to an absorbing state as the event "patient is discharged from the ICU and then lives/dies in the subsequent 90 days", then most observed "discharges" at the 20th step do not reflect an actual discharge from the ICU, and this can mislead the model. For instance, suppose that many patients are in critical condition at the 20th step and are receiving aggressive interventions. In reality, these trajectories likely continue as the physicians work to stabilize these patients. However, because the processed data implies that these patients are immediately discharged, this can lead the model to believe that when another patient is in critical condition, even at the start of a trajectory, there is some moderate probability that their stay will immediately end.

High vasopressor doses may be linked to small sample sizes: The RL model does not consider sample size. In the

work that we replicate[1], actions that are taken less than 5 times from a particular state are removed from the training data.[k] Among the training data, the RL action from the state analyzed for patient 1 was observed for only 6 patients (who survived), while the common action was observed for 148 patients who survived and 13 patients who died. Thus, the model learns that the rare RL action leads to better outcomes on average. This selection of rare (but aggressive) actions is a common phenomenon: Using the training samples, we compute the frequency with which the RL action was chosen by clinicians in each state. Among the 100 states where this proportion was smallest, the RL action is observed 1.5% of the time on average (6.4 observations per state), while the action most frequently chosen by clinicians ("common practice") is observed 35.7% of the time on average (154.5 observations per state). These states collectively make up 26.2% of the transitions in the training data, a nontrivial fraction. The RL policy tends to recommend more vasopressor treatment than common practice. For 99 of these 100 states, common practice involves zero vasopressors. Yet, the RL policy recommends vasopressors in 87 of those states, with 49 of those recommendations being large doses, which we define as those in the upper 50th percentile of nonzero amounts. Recent work has proposed various methods to constrain the RL policy to more closely resemble the behavior policy, which may be required here[5,22].

Clinically implausible discharges may be related to discretization of time and censoring: We observed for patient 1 that the model expects discharge even while on vasopressors. This is clinically surprising but not obvious to the model given the data pre-processing: Of the 8271 training trajectories that end with discharge leading to survival, 4.9% end on non-zero vasopressor dosages, and 2.6% have large dosages, as defined in the previous paragraph. This may be because the data is discretized into 4-hour time intervals, an issue noted by other authors[16], and the action captures the maximum vasopressor dosage during that interval. Thus, it is possible that vasopressors were administered briefly towards the start of that interval, and the patient was stabilized afterwards. As a result, the model contradicts clinical intuition by optimizing for discharge with vasopressors as a desirable goal: Of the model-based roll-outs (based on test trajectories) that end with discharge leading to survival, 52.8% end on non-zero vasopressor dosages, and 31.2% end with large dosages. This may be driven in part by censoring: Of the 1714 training trajectories that get censored but eventually lead to 90-day survival, 10.3% end on nonzero vasopressor dosages, and 5.6% are large dosages. Even when we consider the censored samples, discharging a patient on vasopressors is still much more common in the roll-outs than would be expected in actual clinical practice.

Other factors to investigate: We observed that the model may have insufficient knowledge about underlying conditions, including the surgical context for admission for patient 1 and specific comorbidities like lung cancer for patient 2. Features like these that could impact both treatment decisions and patient outcome may need to be added to the model to address unobserved confounding. The reward function is another part of the model that may be misspecified. For patient 2, the clinician could have already been aware during the ICU stay that patient comfort is the goal, while the RL model is still optimizing for patient survival. Learning better reward functions is an active area of research[6]. Our approach can help check whether the learned reward functions lead to models that align with expected outcomes.

Discussion

Reinforcement learning has the potential to support clinicians in patient care. To ensure that new policies are safe before deployment[5] and this exciting potential can reach fruition, we develop a strategy that allows researchers to bring clinicians into the study design process, identify potentially dangerous flaws in the model, and incorporate that information into models that are safe for deployment. To summarize, our workflow consists of three steps: (1) selecting trajectories for inspection based on surprisingly aggressive treatment recommendations or surprisingly positive outcomes, (2) inspecting the factual and model-based trajectories alongside the medical record, and (3) formulating insights for study design improvement. We demonstrate that applying our approach to a recent work on RL for sepsis suggests better reward specification, data preprocessing, and sample size considerations can help improve the design of the RL study. Our method only requires simulating from a learned model and can therefore be used in any model-based RL application, including more complex models involving neural networks[2] and latent variables[12]. Our approach can also help build trust for models that accurately reflect patient dynamics, by allowing clinicians to check that the model correctly anticipates the impact of the policy it recommends. We see our workflow as a helpful procedure for checking model behavior in various applications of RL to healthcare, including ventilation weaning[4] and HIV[7], and a key step for bringing RL into meaningful clinical decision support tools.

[k] If all actions at a state occur less than 5 times, the RL policy is set to give no treatment. We omit such states from our analysis in this paragraph.

References

1. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature medicine. 2018 Nov;24(11):1716-20.
2. Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. Machine Learning for Healthcare. 2017.
3. Raghu A, Komorowski M, Singh S. Model-based reinforcement learning for sepsis treatment. Machine Learning for Health (ML4H) Workshop at NeurIPS. 2018.
4. Prasad N, Cheng LF, Chivers C, Draugelis M, Engelhardt BE. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In: Association for Uncertainty in Artificial Intelligence. 2017.
5. Futoma J, Masood MA, Doshi-Velez F. Identifying distinct, effective treatments for acute hypotension with SODA-RL: safely optimized diverse accurate reinforcement learning. AMIA Summits on Translational Science Proceedings. 2020;2020:181.
6. Srinivasan S, Doshi-Velez F. Interpretable batch IRL to extract clinician goals in ICU hypotension management. AMIA Summits on Translational Science Proceedings. 2020;2020:636.
7. Parbhoo S, Bogojeska J, Zazzi M, Roth V, Doshi-Velez F. Combining kernel and model based learning for HIV therapy selection. AMIA Summits on Translational Science Proceedings. 2017;2017:239.
8. Hernán MA, Robins JM. Causal inference. Chapman & Hall/CRC, Boca Raton, 2020.
9. Chakraborty B. Statistical methods for dynamic treatment regimes. Springer; 2013.
10. Gottesman O, Johansson F, Meier J, Dent J, Lee D, Srinivasan S, Zhang L, Ding Y, Wihl D, Peng X, Yao J. Evaluating reinforcement learning algorithms in observational health settings. arXiv preprint arXiv:1805.12298. 2018 May 31.
11. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, Celi LA. Guidelines for reinforcement learning in healthcare. Nat Med. 2019 Jan 1;25(1):16-8.
12. Futoma J, Hughes MC, Doshi-Velez F. POPCORN: Partially observed prediction constrained reinforcement learning. In: International Conference on Artificial Intelligence and Statistics. 2020.
13. Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3(1):1-9.
14. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. Journal of artificial intelligence research. 1996 May 1;4:237-85.
15. Sutton RS, Barto AG. Reinforcement learning: an introduction. MIT press; 2018 Oct 19.
16. Jeter R, Josef C, Shashikumar S, Nemati S. Does the "artificial intelligence clinician" learn optimal treatment strategies for sepsis in intensive care? arXiv preprint arXiv:1902.03271. 2019 Feb 8.
17. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. Understanding the artificial intelligence clinician and optimal treatment strategies for sepsis in intensive care. arXiv preprint arXiv:1903.02345. 2019 Mar 6.
18. Oberst M, Sontag D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In: International Conference on Machine Learning 2019 May 24 (pp. 4881-4890).
19. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama. 2016 Feb 23;315(8):801-10.
20. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, Deutschman CS. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Jama. 2016 Feb 23;315(8):762-74.
21. Hunter A, Johnson L, Coustasse A. Reduction of intensive care unit length of stay: the case of early mobilization. Health Care Manag (Frederick). 2014;33(2):128-135. doi:10.1097/HCM.000000000000000
22. Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: International Conference on Machine Learning 2019 May 24 (pp. 2052-2062).

A Discrete Joint Model for Entity and Relation Extraction from Clinical Notes

Zongcheng Ji, Ph.D.[1], Omid Ghiasvand, Ph.D.[1], Stephen Wu, Ph.D.[1], Hua Xu, Ph.D.[1]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

Abstract

Extracting clinical concepts and their relations from clinical narratives is one of the fundamental tasks in clinical natural language processing. Traditional solutions often separate this task into two subtasks with a pipeline architecture, which first recognize the named entities and then classify the relations between any possible entity pairs. The pipeline architecture, although widely used, has two limitations: 1) it suffers from error propagation from the recognition step to the classification step, 2) it cannot utilize the interactions between the two steps. To address the limitations, we investigated a discrete joint model based on structured perceptron and beam search to jointly perform named entity recognition (NER) and relation classification (RC) from clinical notes.

Introduction

Clinical natural language processing (NLP) plays a critical role in unlocking important patient information embedded in clinical narratives of electronic health records (EHRs)[1,2]. Leveraging such information can facilitate the second use of EHRs to promote clinical and translational research. One of the fundamental tasks in clinical NLP research is to identify clinical concepts, and the relations between them[3]. Recently, several challenges have been proposed to automatically extract such information from clinical texts, such as the i2b2 2010 shared task[3], the i2b2 2012 temporal relation extraction task[4] and the 2015/2016/2017 Clinical TempEval Challenges[5], etc.

In this study, we investigated an end-to-end relation extraction task, which is to extract the clinical concepts from the texts and the relations between the extracted concepts. Existing solutions[5,6] often address the problem with two separate steps in a pipeline: first recognize the named entities and then classify the relations between any possible entity pairs. The two steps can be treated as two traditional subtasks, i.e., named entity recognition (NER) and relation classification (RC). The pipeline architecture, although widely used, has two limitations[9]. One is that errors propagate from NER to RC, and there is no feedback from the RC step to the NER step to correct for these errors. The other one is that it over-simplifies the whole task as two independent subtasks, and it cannot utilize the interactions between them.

To address the limitations of the pipeline architecture, joint models were recently proposed in the general domain and biomedical literature to perform NER and RC simultaneously[7–10]. Li and Ji[7] proposed a discrete joint model based on structured perceptron with beam search using both local and global features. Experiments conducted on Automatic Content Extraction (ACE) corpus showed that the discrete joint model significantly outperformed a strong pipelined baseline. Inspired by the work of Li and Ji[7], Li et al.[8, 9] applied similar discrete joint models to extract adverse drug events (ADEs) between drug and disease entities from PubMed abstracts.

Despite that joint models were successfully applied to address the limitations of the pipelined method for performing NER and RC in both the general domain[7, 10] and biomedical literature[7, 8], few work has been done with the clinical narratives. We are aware of two published studies on joint methods[11, 12] for recognizing some specific entities and their relations with medications. Wei et al.[12] proposed a joint method only for attribute detection, which identifies only attribute entities and classifies their relations with medications in one step. Leeuwenberg and Moens[13] employed a structured perceptron to jointly predict temporal relations between events and temporal expressions (TLINKS), and the relation between these events and the document creation time (DCTR) from clinical narratives. However, their joint model only focuses on joint extraction of different relations given gold standard entities. Li and Ji[7] did not release their code for their seminal work on joint NER and RC. Although Li et al.[8] released their code for a short period, their method only addressed two types of entities and their relations. It cannot be directly used to address multiple entities and relations. In addition, due to the different writing styles and audiences, the challenges in clinical narratives when compared with literature publications are significant[14]. Consequently, it is necessary to investigate whether or not a joint model could outperform the pipelined method when performing NER and RC from clinical narratives in EHRs. As a preliminary study, here we proposed to develop a discrete joint model for joint NER and RC from clinical narratives, using two public datasets from previous studies[3, 15].

Materials and Methods

Dataset

We used two datasets in this study, namely the i2b2 2010 shared task challenge dataset[3] and Lee et al.'s direct temporal relation extraction dataset[15]. The first dataset was collected from discharge summaries from three different hospitals and was manually annotated by experts with three types of entities including PROBLEM, TEST, and TREATMENT, and eight types of relations including *treatment improves medical problem* (TrIP), *treatment worsens medical problem* (TrWP), *treatment causes medical problem* (TrCP), *treatment is administered for medical problem* (TrAP), *treatment is not administered because of medical problem* (TrNAP), *test reveals medical problem* (TeRP), *test conducted to investigate medical problem* (TeCP), and *medical problem indicates medical problem* (PIP). out of 477 original 170 were available for download and we randomly split them in a 60:20:20 ratio for training, development, and test sets respectively. The statistics of this dataset is shown in Table 1.

The second dataset was constructed by Lee et al. to extract *direct* temporal relations from discharge summaries by leveraging the i2b2 2012 temporal relation extraction dataset[4]. This dataset contains entity types EVENT and TIMEX3 and three types of relations between them (AFTER, BEFORE, and OVERLAP), which followed the types used in the i2b2 2012 shared task[4]. In this study, in order to get a development set to tune the model, we also combined the original 190 training documents and 120 testing documents and randomly split them in a 60:20:20 ratio for training, development, and test sets respectively. The statistics of this dataset are shown in Table 2.

Table 1. Statistics of the i2b2 2010 dataset.

	Train	**Development**	**Test**
#documents	266	80	80
#sentences	27,429	7,707	8,805
#entities (PROBLEM)	12,610	3,088	3,966
#entities (TEST)	8,619	2,558	2,654
#entities (TREATMENT)	8,949	2,259	2,978
#relations (TrIP)	130	36	37
#relations (TrWP)	91	19	23
#relations (TrCP)	333	110	83
#relations (TrAP)	1,685	391	541
#relations (TrNAP)	112	28	34
#relations (TeRP)	2,061	429	563
#relations (TeCP)	330	73	101
#relations (PIP)	1,348	367	488

Table 2. Statistics of Lee et al.'s direct temporal extraction dataset.

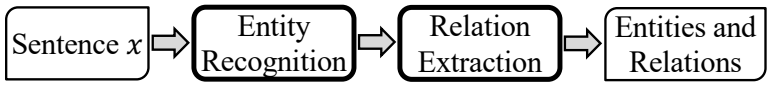
	Train	**Development**	**Test**
#documents	190	60	60
#sentences	7,888	2,613	2,610
#entities (EVENT)	12,611	4,478	4,125
#entities (TIMEX3)	2,517	882	789
#relations (AFTER)	382	133	129
#relations (BEFORE)	464	139	139
#relations (OVERLAP)	1,598	563	529

The Baseline Pipeline Architecture

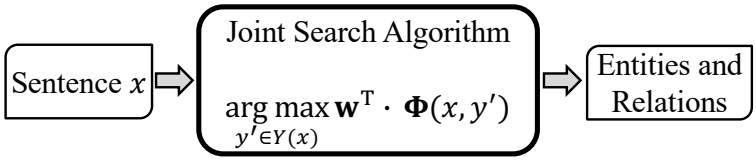
A straightforward solution to the end-to-end relation extraction task is first to recognize the entity mentions from a given sentence and then to classify the relations between any possible entity pairs. We employed this pipelined solution (shown in Figure 1(a)) for both subtasks as baselines, using the same implementation as we did in the previous challenges, in which our entries ranked top on both datasets[3, 4, 16, 17].

NER: We cast the NER task as a sequential token tagging task, adopting the well-known BIO scheme. We employed a linear-chain Conditional Random Fields (CRF) model[18] for the NER subtask, since it has shown state-of-the-art

performance in many clinical NER systems[1–4, 6] in several challenges. The CRFsuite package[19] was used to train the CRF models. The features used for the CRF model are the token-based features for NER listed in Table 3.



(a) Pipeline Architecture



(b) Joint Framework

Figure 1. Overview of the pipeline architecture and the joint framework for end-to-end relation extraction

RC: Given a sentence with the recognized entity mentions, the RC task is to classify each pair of entity mentions into one of several pre-defined relation types. We employed a support vector machine (SVM) classifier for the RC subtask, since it has shown state-of-the-art performance in many clinical RC systems[3, 4, 6, 15, 16] in several challenges. The LIBLINEAR package[22] was used to train the SVM classifiers. We also employed cost-sensitive learning[6, 15, 21] in order to counterbalance the effect of dominating number of negative instances, since the type distributions of the relations are unbalanced (see Table 1 and Table 2). To each relation type, we assigned a weight that is inversely proportional to the class frequency, adjusting the penalty factor in the SVM training[21]. The features used for the SVM classifier are the local features for RC listed in Table 3.

The Joint Framework

Inspired by previous work[7–9], we cast the whole task as a structured prediction problem, which performed the two subtasks jointly. This overcomes the two main issues with the pipeline architecture, error propagation and failure to model interactions between related subtasks.

Output Structure Representation: We first introduced a new representation for the output of the end-to-end relation extraction task. Given an input sentence x, the output structure y consists of the following two types of nodes:

- **Segment Node** $S(j, i, t)$: A segment is a sequence of tokens. j and i denote the left and right boundaries of a segment, $1 \leq j \leq i \leq |x|$, where $|x|$ is the length of a sentence x; the type t of the segment is drawn from a task-specific set of labels T_s. For example, in Lee et al.'s direct temporal relation extraction dataset, $T_s = \{EVENT, TIMEX3, O\}$. Namely, $t = EVENT$ if the segment is an event mention, $t = TIMEX3$ if the segment is a time expression mention, and $t = O$ if the segment is neither an event mention nor a time expression mention. The length of a type O segment is always 1.
- **Relation Node** $R(i_1, i_2, r)$: i_1 and i_2 denote the right boundaries of two segment nodes; $r \in T_r$ is the type of the relation node. For example, in Lee et al.'s direct temporal relation extraction dataset, $T_r = \{AFTER, BEFORE, OVERLAP, NIL\}$. Here, $r = AFTER$ if the two segment nodes have the AFTER relation, $r = BEFORE$ if they have the BEFORE relation, $r = OVERLAP$ if they have the OVERLAP relation and $r = NIL$ if they do not have any direct temporal relations.

Structured Prediction Formulation: With the new output structure representation, the end-to-end relation extraction task becomes a structured prediction problem, which is to predict the most probable output structure $\hat{y}$ for a given sentence x. Let $x \in X$ be an input sentence, and $y' \in Y(x)$ be a candidate output structure. Our goal is to predict the most probable output structure $\hat{y}$ for x. We use the following linear model to predict the most probable output structure $\hat{y}$ for x:

$$\hat{y} = \arg\max_{y' \in Y(x)} \mathbf{w}^{\mathrm{T}} \cdot \mathbf{\Phi}(x, y') \quad (1)$$

where $\mathbf{\Phi}(x, y')$ is the feature vector that characterizes the input sentence x together with a candidate output structure y', and $\mathbf{w}$ is the corresponding feature weights. With the new problem definition, the end-to-end relation extraction can be performed naturally in a joint search space simultaneously, shown in Figure 1(b).

Table 3. Summary of the features used in this work

Feature Type	Feature Description
Local Features for Named Entity Recognition (Token-based)	
Word Shape Features	The word itself, its stemmed form and its shape with converting all the numbers, capital and lowercase letters to '#', 'A', 'a'
N-gram Features	Bag-of-words or POS tags of the context window up to 5 words
Prefix / Suffix Features	Word prefixes and suffixes, from 1 to 3 characters
Sentence Features	Sentence length and whether the sentence starts with enumerate words
Section Features	Which section of the clinical note the word appears in
Regular Expression Features	Whether or not the word matches with a predefined regular expression set
Dictionary Features	Pre-label of words with a given domain dictionary based on BIO schema
Brown Clustering Features	Brown clustering features based on the 4th, 8th, and 12th bits of the path
Word Embeddings Features	Word embeddings of the context window up to 5 words
Local Features for Named Entity Recognition (Segment-based)	
Segment Shape Features	The segment itself, its stemmed form and its shape with concatenating the shape of each word in the segment
Context Features	Bag-of-words or POS tags of the preceding / following two words
Dictionary Features	Whether the segment appears in a given domain dictionary
Local Features for Relation Classification	
Entity Features	The type of an entity, the surface form and stemmed form of an entity, and the combinations of the stemmed words in both the entities involved
Context Features	The surface form and stemmed form of (1) the preceding / following two words of an entity mention and (2) the words between the two entity mentions
Position Features	The position and direction (left or right) information between the two entity mentions
Global Features for Named Entity Recognition	
Neighbor Coherence Features	Neighbor coherence between two neighboring segments
Global Features for Relation Extraction	
Neighbor Coherence Features	Neighbor coherence between two relations if an entity mention is shared

Joint Decoding Algorithm: The key step in both training and test is the decoding algorithm, which aims to search for the best output structure under the current model parameters. Since it is intractable to perform exact search in the joint framework[7], we employed a *beam-search* algorithm, an instance of inexact search, to approximate Equation (2).

Specifically, for an input sentence, the *beam-search* algorithm incrementally expands partial output structures to find the optimal output structure with the best score. The *k*-best partial output structures for x ending at the i^{th} token is:

$$B[i] = \underset{y_{[1:i]} \in Y(x,i)}{\arg\operatorname{top}^{k}} \mathbf{w}^{\mathrm{T}} \cdot \mathbf{\Phi}(x, y_{[1:i]}) \qquad (2)$$

where $y_{[1:i]}$ denotes the partial output structure whose last segment ends at the i^{th} token, and $Y(x, i)$ stands for the search space. The joint decoding algorithm is shown in Algorithm 1. For each token index i, the algorithm maintains a beam $B[i]$ for the partial output structures whose last segments end at the i^{th} token (line 11 and 22 in Algorithm 1).

Algorithm 1 Joint Decoding Algorithm based on Beam-Search

Input: a sentence $x = x_{[1...|x|]}$
k: beam size
T_s: types of a segment node
T_r: types of a relation node
$\hat{d}_t$: maximum length of a segment with type t, $t \in T_s$
Output: best output structure $\hat{y}$ for x

initialize $|x|$ empty beams $B[1...|x|]$
for $i \in 1...|x|$ **do**
  `/* recognition action */`
  $buf \leftarrow \emptyset$
  for $t \in T_s$ **do**
    for $d \in 1...\hat{d}_t$ **do**
      $j \leftarrow \max(1, i - d + 1)$
      $S(j, i, t) \leftarrow \text{New\_Segment}(j, i, t)$
      for $y_{[1:i-d]} \in B[i - d]$ **do**
        $y_{[1:i]} \leftarrow \text{Append}(y_{[1:i-d]}, S(j, i, t))$
        $buf \leftarrow buf \cup \{y_{[1:i]}\}$
  $B[i] \leftarrow k\text{-best}(buf)$
  `/* relation action */`
  for $j \in (i - 1)...1$ **do**
    $buf \leftarrow \emptyset$
    for $y_{[1:i]} \in B[i]$ **do**
      if $\text{Has\_Valid\_Pair}(y_{[1:i]}, j, i)$ **then**
        for $r \in T_r$ **do**
          $R(i1, i2, r) \leftarrow \text{New\_Relation}(y_{[1:i]}, j, i, r)$
          $y'_{[1:i]} \leftarrow \text{Append}(y_{[1:i]}, R(i1, i2, r))$
          $buf \leftarrow buf \cup \{y'_{[1:i]}\}$
      else
        $buf \leftarrow buf \cup \{y_{[1:i]}\}$
    $B[i] \leftarrow k\text{-best}(buf)$
return $B[|x|][0]$

Algorithm 2 Structured Perceptron Algorithm with Beam-Search & Early-Update

Input: training data $\mathcal{D}^{\mathcal{T}} = \{x_i, y_i\}_{i=1}^{n}$
I: maximum iteration number
Output: model parameters $\mathbf{w}$

initialize $\mathbf{w} \leftarrow \mathbf{0}$
for $t \in 1...I$ **do**
  for each $(x, y) \in \mathcal{D}^{\mathcal{T}}$ **do**
    $(y', z) \leftarrow \text{BeamSearch}(x, y, \mathbf{w})$
    if $z \neq y'$ **then**
      $\mathbf{w} \leftarrow \mathbf{w} + \mathbf{\Phi}(x, y') - \mathbf{\Phi}(x, z)$
return $\mathbf{w}$

Model Training: We employed a structured perceptron[22], an extension of the standard perceptron for structured prediction, to estimate the model parameters $\mathbf{w}$ from the training data. For each labeled example (x_i, y_i), the algorithm uses Equation (3) to search for the best output structure $\hat{y}_i$ for x_i under the current model parameters. If $\hat{y}_i$ is different from the ground truth y_i, then the parameters are updated as follows:

$$\mathbf{w} \leftarrow \mathbf{w} + \mathbf{\Phi}(x_i, y_i) - \mathbf{\Phi}(x_i, \hat{y}_i) \tag{3}$$

Huang et al.[23] proved the convergence of the structured perceptron when violation-fixing update methods such as *early-update*[24] are applied to beam search. In this work, we employed the *early-update* method for model training, as shown in Algorithm 2. For each training example (x, y), the algorithm performs *beam-search*, which is Algorithm 1

with one exception. If y', the prefix of the ground-truth y, falls out of the beam after each execution of k-best function (line 11 and 22 in Algorithm 1), then y' and the top partial output structure z in the current beam are returned for updating parameters (line 4 in Algorithm 2). In practice, we used averaged parameters to avoid overfitting[22] when decoding the test examples.

Features: We used local and global features in the joint framework as shown in Table 3.

- **Local Features**: Local features are only related to the individual segments, which include the token-based features for NER, the segment-based features for NER, and the local features for RC.

Global Features: One advantage of the joint framework is that we can easily exploit arbitrary global features from the entire output structure to capture long-distance dependencies within a task and cross-task dependencies[7]. We developed an NER-specific global feature (i.e., the Neighbor Coherence Feature for NER shown in Table 3) once a new segment node is made during decoding. The assumption of this global feature is that neighboring entity mentions tend to have coherent entity types.

Evaluation Metrics

For both NER and RC, we adopt three widely used metrics for evaluation: Precision (P), Recall (R) and F_1. P is a measure of what percentage the predicted output labels are correct, and R tells us to what percentage the gold-standard dataset are correctly labeled by the system. F_1 is the harmonic mean of P and R.

Parameters Setting

There were several parameters to be set in Algorithms 1 and 2. The maximum length $\hat{d}_t$ for each segment with type t was collected from the training data at the beginning of the training phase. Table 4 shows the maximum length of each type of segment node in the training, development and test sets. We found that the numbers collected from the training data were larger than those in both the development and test data. The beam size k and maximum iteration number I were learned from the development set. Similar to the findings in previous work[9], larger beam sizes lead to marginal increase in performance but much longer decoding time. As a trade-off, we set the beam size $k = 2$ throughout the experiments. We set the maximum number of training iterations $I = 40$ throughout the experiments.

Table 4. Maximum length of each type of segment in the two datasets.

Segment Node Type t	Maximum Length $\hat{d}_t$		
	train	development	test
PROBLEM	12	8	10
TEST	11	6	6
TREATMENT	8	7	7
EVENT	10	9	9
TIMEX3	6	5	5

Results

Results on Development Sets

Figure 3 and Figure 4 show the learning curves on the development set of the i2b2 2010 dataset and Lee et al.'s direct temporal relation extraction dataset, respectively. The learning curves compare both the NER and RC performance of the joint model with and without global features in terms of F_1. From these figures, it is clear that the global features are effective at improving the extraction performance on both tasks. We can also see that the performance gap between the model with and without global features becomes smaller when the number of iterations increases to 40. Finally, we set the number of training iterations as 20 and 11 for the two datasets based on the learning curves.

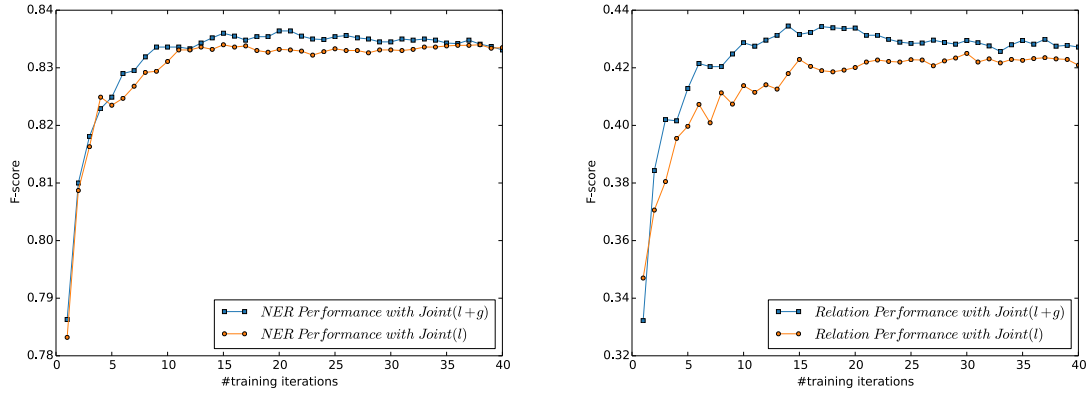


Figure 2. Learning curves on the development set of i2b2 2010 dataset.

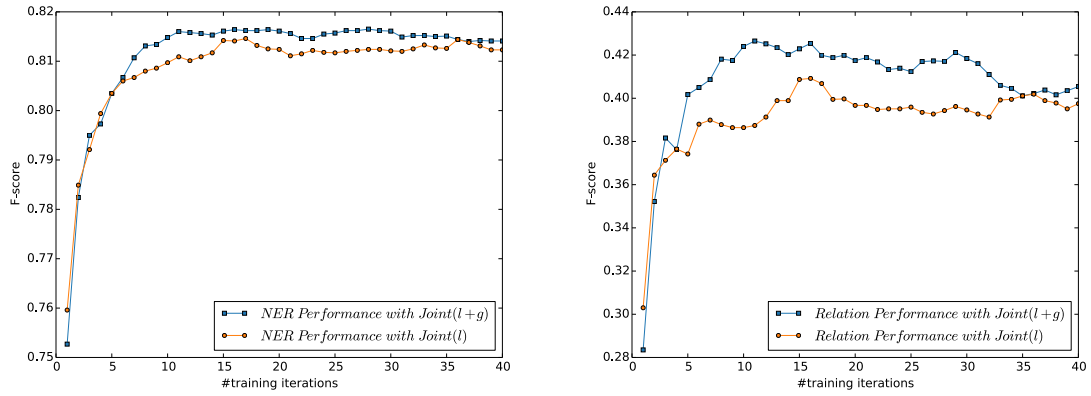


Figure 3. Learning curves on the development set of Lee et al.'s direct temporal relation extraction dataset.

Overall Performance on Test Sets

We compared the following three methods on the end-to-end relation extraction task from the two tests sets.

- *Pipeline*: This baseline method is based on the pipeline architecture.
- *Joint (l)*: This method is based on the joint framework with only local features.
- *Joint (l+g)*: This method is based on the joint framework with both local and global features.

Table 5 illustrates the overall performance on the NER and RC subtasks on the i2b2 2010 dataset and Lee et al.'s direct temporal relation extraction dataset. From the table, we observe that (1) Both *Joint (l)* and *Joint (l+g)* consistently outperformed *Pipeline* on the two datasets in precision score by 1.2-1.6% on the NER subtask and by 4.1-7.5% on the RC subtask. There was no significant improvement when comparing *Joint (l+g)* with *Joint (l)* in precision score on the NER subtask, while *Joint (l+g)* outperformed *Joint (l)* in precision score by 1.3-1.6% on the RC subtask. (2) Both *Joint (l)* and *Joint (l+g)* outperformed *Pipeline* on Lee et al.'s direct temporal relation extraction dataset in recall score by 0.5-0.6% on the NER subtask and by 0.8-1.8% on the RC subtask, while both the joint models did not improve recall on the i2b2 2010 dataset on both the NER and RC subtasks. (3) Both *Joint (l)* and *Joint (l+g)* consistently outperformed *Pipeline* on the two datasets in F_1 score by 0.7-0.9% on the NER subtask and by 0.5-3.5% on the RC subtask. Similar to the precision score, there was no significant improvement when comparing *Joint (l+g)* with *Joint (l)* in F_1 score on the NER subtask, while *Joint (l+g)* outperformed *Joint (l)* in F_1 score by 0.8-1.4% on the RC subtask.

In summary, both *Joint (l)* and *Joint (l+g)* consistently achieved higher precision and F_1 than *Pipeline*, although these joint models did not significantly improve recall on the RC subtask (in fact, recall decreased on the i2b2 2010 dataset). *Joint (l)* outperformed *Pipeline* on the two datasets in F_1 score by up to 0.8% on the NER subtask, and by up to 2.1% on the RC subtask. When the global features were introduced, *Joint (l+g)* further improved the performance and

outperformed *Pipeline* on the two datasets in F_1 score by up to 0.9% on the NER subtask and by up to 3.5% on the RC subtask.

Table 5. Overall performance on the NER and RC subtasks on the i2b2 2010 dataset and Lee et al.'s direct temporal relation extraction dataset

Method	NER			RC		
	P	*R*	F_1	*P*	*R*	F_1
i2b2 2010 dataset						
Pipeline	0.8395	0.8240	0.8317	0.4429	**0.4000**	0.4203
Joint (l)	**0.8554**	0.8223	0.8385	0.5052	0.3672	0.4253
	+1.6%	-0.2%	+0.7%	+6.2%	-3.2%	+0.5%
Joint (l+g)	0.8533	**0.8255**	**0.8392**	**0.5174**	0.3731	**0.4336**
	+1.4%	+0.2%	+0.8%	+7.5%	-2.7%	+1.3%
Lee et al.'s direct temporal relation extraction dataset						
Pipeline	0.8120	0.8026	0.8073	0.4161	0.3706	0.3920
Joint (l)	**0.8238**	0.8077	0.8157	0.4568	0.3781	0.4137
	+1.2%	+0.5%	+0.8%	+4.1%	+0.8%	+2.1%
Joint (l+g)	0.8236	**0.8085**	**0.8160**	**0.4732**	**0.3882**	**0.4265**
	+1.2%	+0.6%	+0.9%	+5.7%	+1.8%	+3.5%

Table 6. Performance on each entity and relation type on the i2b2 2010 dataset and Lee et al.'s direct temporal relation extraction dataset.

Type	*Pipeline*			*Joint (l)*			*Joint (l+g)*		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
i2b2 2010 dataset									
PROBLEM	0.8268	0.8283	0.8276	0.8419	0.8243	0.8330	0.8434	0.8323	**0.8378**
TEST	0.8528	0.8271	0.8397	0.8699	0.8240	**0.8464**	0.8663	0.8252	0.8452
TREATMENT	0.8452	0.8156	0.8301	0.8611	0.8180	**0.8390**	0.8554	0.8167	0.8356
TrIP	0.6190	0.3514	**0.4483**	0.2632	0.1351	0.1786	0.4000	0.2162	0.2807
TrWP	0.2727	0.1304	0.1765	0.3333	0.1739	**0.2286**	0.2857	0.1739	0.2162
TrCP	0.3000	0.2892	0.2945	0.3944	0.3373	0.3636	0.4247	0.3735	**0.3974**
TrAP	0.4509	0.5009	0.4746	0.5037	0.5056	0.5046	0.5112	0.5093	**0.5102**
TrNAP	0.3846	0.1471	0.2128	0.3750	0.1765	**0.2400**	0.3077	0.1176	0.1702
TeRP	0.5565	0.5595	0.5580	0.5669	0.5730	0.5699	0.5800	0.5872	**0.5836**
TeCP	0.3659	0.2970	**0.3279**	0.3857	0.2673	0.3158	0.3913	0.2673	0.3176
PIP	0.2762	0.1783	**0.2167**	0.3500	0.0430	0.0766	0.3529	0.0369	0.0668
Lee et al.'s direct temporal relation extraction dataset									
EVENT	0.8138	0.8075	0.8107	0.8263	0.8141	**0.8201**	0.8245	0.8143	0.8194
TIMEX3	0.8024	0.7769	0.7894	0.8103	0.7744	0.7920	0.8187	0.7782	**0.7979**
AFTER	0.3934	0.1860	0.2526	0.4194	0.2016	0.2723	0.4000	0.2326	**0.2941**
BEFORE	0.3902	0.2319	0.2909	0.4028	0.2101	0.2762	0.4595	0.2464	**0.3208**
OVERLAP	0.4223	0.4518	0.4365	0.4686	0.4650	0.4668	0.4861	0.4631	**0.4743**

Performance of each entity and relation type

To give a more nuanced view of the comparative performance of the models, we show the performance of each entity and relation type on the two datasets in Table 6. From the table, we see that (1) Both *Joint (l)* and *Joint (l+g)* consistently outperformed *Pipeline* on the two datasets in F_1 score in all entity types. *Joint (l+g)* further outperformed *Joint (l)* in F_1 score in the PROBLEM entity type and TIMEX3 entity type on the i2b2 2010 dataset and Lee et al.'s

direct temporal relation extraction dataset, respectively. (2) Both *Joint (l)* and *Joint (l+g)* outperformed *Pipeline* on the i2b2 2010 dataset in F_1 score in most of the relation types except TrIP, TeCP and PIP, and *Joint (l+g)* further outperformed *Joint (l)* in the TrIP, TrCP, TrAP and TeRP relation types. *Joint (l)* outperformed *Pipeline* on Lee et al.'s direct temporal relation extraction dataset in F_1 score in all relation types except BEFORE, while *Joint (l+g)* outperformed *Pipeline* and *Joint (l)* in all relation types.

In summary, both *Joint (l)* and *Joint (l+g)* outperformed *Pipeline* on the two datasets in F_1 score in all entity types and most relation types. *Joint (l+g)* further outperformed *Joint (l)* on the two datasets in F_1 score in most of the entity and relation types.

Discussion

In this study, we investigated the impact of a discrete joint model for entity and relation extraction from clinical notes, using two public datasets[3,15]. The joint model with both local and global features outperformed the state-of-the-art pipelined methods on the two datasets in F_1 score by up to 0.9% on the NER subtask and by up to 3.5% on the RC subtask. To the best of our knowledge, this is one of the initial studies to investigate discrete joint models for the end-to-end relation extraction in clinical notes.

Although the discrete joint model outperformed the pipelined method on both the NER and RC subtasks in terms of precision and F_1, it didn't improve significantly in terms of recall and actually decreased performance (compared to the pipelined approach) on the RC subtask on the i2b2 2010 dataset. We believe the main reason is that the combination of the subtasks led to feature sparsity in the discrete joint model, which requires more effective learning algorithms for training, or an alternate representation. We will try to apply *k*-best MIRA method[28], an online large-margin learning algorithm, to address this issue. Another possible solution to this issue is to introduce low-dimensional dense features (e.g., neural features) into the joint framework.

One advantage of the joint framework is that we can easily introduce arbitrary global features into the model. In this work, we introduced neighbor coherence features for both the NER and RC purpose as shown in Table 3. As shown in the overall results of both the development and test sets, *Joint (l+g)* consistently outperformed *Joint (l)* in F_1 score, indicating the effectiveness of the introduced global features. We also noticed that the RC subtask benefited much more from the global features than that of the NER subtask. One possible reason is that the RC subtask relies much more on the long-distance dependencies than that of the NER subtask, which can be well captured by the introduced global features. When we further investigated the performance of each entity and relation type on the test sets, Table 6 shows that *Joint (l+g)* outperformed *Joint (l)* in most of the entity and relation types.

From Table 6, we observed that for the i2b2 2010 dataset, the joint models were very worse than *pipeline* in the TrIP and PIP relation types, although they outperformed *pipeline* in most of the relation types. This shows the difficulty of the joint models on the two relation types. One possible reason is that the size of the data in the TrIP relation type is not enough to train a good joint model. For the data in the PIP relation type, although the size of the data is reasonable, it would be confusing for the discrete joint model to predict the relation between two problem entities. On possible solution is to introduce some global features to rescue these errors. We will leave it as future work.

Finally, our discrete joint models for the end-to-end relation extraction from clinical notes required a substantial feature engineering effort. Recently, neural joint models have alleviated these concerns[27]. Thus, we will investigate such neural joint models for joint information extraction from clinical data. We will also develop some strategies to combine the discrete joint model and neural joint model in the future.

Conclusion

In this study, we applied a discrete joint model based on structured perceptron and beam search to jointly perform NER and RC from clinical notes, in order to address the limitations of the traditional pipeline architecture. Results showed that the discrete joint model effectively improved the performance compared to its pipelined counterpart on the end-to-end relation extraction from clinical notes.

References

1. Zhang Y, Wang J, Tang B, et al. UTH_CCB: a report for semeval 2014 - task 7 analysis of clinical text. In: *SemEval*. ; 2014:802.
2. Xu J, Zhang Y, Wang J, et al. UTH-CCB: The Participation of the SemEval 2015 Challenge - Task 14. In: *SemEval*. ; 2015:311-314. http://www.aclweb.org/anthology/S15-2052.
3. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in

clinical text. *JAMIA*. 2011;18(5):552-556.
4. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *JAMIA*. 2013;20(5):806-813.
5. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 Task 6: Clinical TempEval. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics; 2015:806-814. http://www.aclweb.org/anthology/S15-2136.
6. Lee H-J, Xu H, Wang J, et al. UTHealth at SemEval-2016 Task 12: an End-to-End System for Temporal Information Extraction from Clinical Notes. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. ; 2016:1292-1297. http://www.aclweb.org/anthology/S16-1201.
7. Li Q, Ji H. Incremental Joint Extraction of Entity Mentions and Relations. In: *ACL*. ; 2014:402-412.
8. Li F, Zhang Y, Zhang M, Ji D. Joint models for extracting adverse drug events from biomedical text. In: *IJCAI*. ; 2016:2838-2844.
9. Li F, Ji D, Wei X, Qian T. A transition-based model for jointly extracting drugs, diseases and adverse drug events. In: *BIBM*. ; 2015:599-602.
10. Miwa M, Sasaki Y. Modeling Joint Entity and Relation Extraction with Table Representation. *EMNLP*. 2014:1858-1869.
11. Dandala B, Joopudi V, Devarakonda M. Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks. *Drug Saf*. 2019;42(1):135-146. doi:10.1007/s40264-018-0764-x
12. Wei Q, Ji Z, Li Z, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *JAMIA*. 2019. doi:10.1093/jamia/ocz063
13. Leeuwenberg A, Moens M-F. Structured Learning for Temporal Relation Extraction from Clinical Records. In: *EACL*. Valencia, Spain: Association for Computational Linguistics; 2017:1150-1158. http://www.aclweb.org/anthology/E17-1108.
14. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. *JBI*. 2015;57:28-37. doi:10.1016/J.JBI.2015.07.010
15. Lee H-J, Zhang Y, Jiang M, Xu J, Tao C, Xu H. Identifying direct temporal relations between time and events from clinical notes. *BMC Med Inform Decis Mak*. 2018;18(2):49.
16. Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. A hybrid system for temporal information extraction from clinical text. *JAMIA*. 2013;20(5):828-835.
17. Jiang M, Chen Y, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *JAMIA*. 2011;18(5):601-606.
18. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *ICML*. ; 2001:282-289.
19. Okazaki N. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/. Published 2007.
20. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A library for large linear classification. *J Mach Learn Res*. 2008;9(Aug):1871-1874.
21. Ben-Hur A, Weston J. A user's guide to support vector machines. In: *Data Mining Techniques for the Life Sciences*. Springer; 2010:223-239.
22. Collins M. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In: *EMNLP*. ; 2002:1-8. doi:10.3115/1118693.1118694
23. Huang L, Fayong S, Guo Y. Structured Perceptron with Inexact Search. In: *NAACL*. ; 2012:142-151. http://dl.acm.org/citation.cfm?id=2382029.2382049.
24. Collins M, Roark B. Incremental Parsing with the Perceptron Algorithm. In: *ACL*. ; 2004. doi:10.3115/1218955.1218970
25. Raj D, Sahu SK, Anand A. Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text. In: *CoNLL*. ; 2017:311-321. doi:10.18653/v1/K17-1032
26. McDonald R, Crammer K, Pereira F. Online large-margin training of dependency parsers. In: *ACL*. ; 2005:91-98.
27. Li F, Zhang M, Fu G, Ji D. A Neural Joint Model for Extracting Bacteria and Their Locations. In: *PAKDD*. ; 2017:15-26.

Recurrent Neural Networks to Automatically Identify Rare Disease Epidemiologic Studies from PubMed

Jennifer N. John[1], Eric Sid, MD, MHA[2], Qian Zhu, PhD[3]

[1]Stanford University, Stanford, CA
[2]Office of Rare Disease Research, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Bethesda, MD
[3]Division of Pre-Clinical Innovation, National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), Rockville, MD

Abstract

Rare diseases affect between 25 and 30 million people in the United States, and understanding their epidemiology is critical to focusing research efforts. However, little is known about the prevalence of many rare diseases. Given a lack of automated tools, current methods to identify and collect epidemiological data are managed through manual curation. To accelerate this process systematically, we developed a novel predictive model to programmatically identify epidemiologic studies on rare diseases from PubMed. A long short-term memory recurrent neural network was developed to predict whether a PubMed abstract represents an epidemiologic study. Our model performed well on our validation set (precision = 0.846, recall = 0.937, AUC = 0.967), and obtained satisfying results on the test set. This model thus shows promise to accelerate the pace of epidemiologic data curation in rare diseases and could be extended for use in other types of studies and in other disease domains.

Introduction

In the United States, a rare disease is defined as affecting fewer than 200,000 people.[1] It is estimated that between 6,000 and 8,000 rare diseases exist,[2] and that they affect between 25 and 30 million people in the United States.[3] Among rare diseases, there is a significant range in prevalence. Some disorders with higher prevalence rates are well-documented in the population; for instance, sickle cell disease is estimated to affect 100,000 people in the United States.[4] Other diseases are much rarer, affecting only a handful of patients. Fewer than twenty cases of Jansen's metaphyseal chondrodysplasia have been reported, for example.[5] Still others are sporadically documented only in occasional case reports. Accurate estimates of prevalence and incidence rates are critical to developing an understanding of a disease's scope and population burden. Continued epidemiological data on a greater distribution of rare diseases can help in recognizing patterns in etiology and inform decisions on research funding by providing quantifiable indications of impact.[6]

Epidemiologic data can be discovered and presented in several ways. The most complete findings are provided through epidemiologic studies, which describe the frequency of a disease in a certain population group by both geographic and demographic distribution. Such studies are often found for rare diseases whose affected population sizes range closer to the upper margins of the US rare disease definition, as they are prevalent enough to warrant a large-scale study and for results to have sufficient statistical strength. For the majority of rare diseases, however, no epidemiology studies have been conducted and population estimates are often derived solely from expert opinions and published case reports.[7] As such, remaining vigilant of newly published epidemiologic studies in these diseases is an important task in guiding research efforts focused on the broader field of rare diseases.

The Genetic and Rare Diseases (GARD) Information Center, a program managed by the National Center for Advancing Translational Sciences (NCATS) within the National Institutes of Health (NIH), aims to curate and disseminate freely accessible consumer health information on over 6,500 genetic and rare diseases.[8] Currently, GARD curators search PubMed for relevant articles and manually review them for curation, which is a tedious and error-prone process. Curators noted that searching with keywords on PubMed returned relevant results, but they found less utility in the ranking of those results and were reliant on a manual process of reviewing and selecting evidence to pick as sources for curating knowledge. By leveraging natural language processing (NLP) techniques to automatically identify rare disease epidemiologic studies from a very large volume of PubMed articles, we aim to supplement this evidence selection process and reduce the need for strict manual review of publications.

Previously, traditional machine learning approaches have been applied to classify electronic health records for epidemiologic studies.[9] Biomedical text classification has been performed using convolutional neural networks[10] and support vector machines.[11] In this work, we explored the use of a recurrent neural network (RNN) to predict the probability that a given scientific abstract on a rare disease is epidemiology related. In particular, we applied long short-term memory units,[12] a type of recurrent neural network that is well-suited for NLP because their ability to store an internal state allows them to effectively process sequential data such as text.[13] RNNs are considered state-of-the-art for sentiment analysis,[14] machine translation,[15] and speech recognition.[16] RNNs have also shown to perform well for biomedicine-related NLP tasks, such as named entity recognition for biomedical related terms[17] and chemical-protein interaction extraction from scientific papers.[18]

To our knowledge, this work represents the first attempt to automatically classify epidemiologic publications for rare diseases. Based on the performance of RNNs in related tasks, we hypothesize that this model will also be well-suited for epidemiology identification. We suspect that an RNN will be able to identify more sophisticated and semantically meaningful features than other machine learning approaches such as rule-based models or support vector machines, due to its mathematical complexity and broad success across NLP applications. This feature is important for this task because of the variation in the structure and content of epidemiologic studies, and the superficial similarities with other publication types, particularly in the limited dataset that is available. In addition, the flexibility of neural networks could allow for other types of publication classification tasks to benefit from this approach.

Methods

Dataset construction

We considered epidemiologic study identification as a binary classification task, which thus requires a positive set, containing PubMed abstracts that are rare disease-related epidemiologic studies, and a negative set, consisting of abstracts that are not rare diseases-related epidemiologic studies. As no such datasets already exist, and manually labeling articles would be labor-intensive, we utilized Medical Subject Headings (MeSH)[19] and NLP techniques to create our own datasets.

Our positive dataset was constructed from a list of reference articles with epidemiologic data curated by Orphanet, which provides datasets relating to rare diseases.[20] We selected only the references indexed by PubMed, which allowed us to retrieve their abstracts and MeSH terms through the EBI RESTful API.[21] While many of these articles were epidemiologic studies, some focused on treatments or genetic causes, and instead contained references to data obtained in previous epidemiologic studies for the disease. Others were case reports, which were excluded from this study. To filter out these types of articles, we retrieved the MeSH terms tagged to each PubMed article through the EBI API. If the PubMed article is tagged with the epidemiology-related MeSH terms including "Epidemiology (MeSH:D004813)", "Prevalence(MeSH:D015995)," or "Incidence(MeSH:D015994)," then the article was retained; otherwise, it was excluded. Articles that are categorized as case reports based on their publication types were also removed. Abstracts were retrieved from the API if available based on their PMIDs.

To construct our negative dataset, we began with a list of 6,073 rare diseases included in GARD. For each of these diseases, we invoked the EBI API to retrieve the top five associated PubMed articles. From these results, we removed articles that fall into one of the following criteria: 1) the article is part of the reference list from the Orphanet epidemiologic data; 2) the article is associated with any of the aforementioned epidemiology related MeSH terms; 3) the abstract mentions any of the keywords of "epidemiology", "prevalence", or "incidence."

We combined the above two sets and used an 80:20 training/validation split. From the Orphanet dataset, we randomly selected one hundred articles to form a test set.

Text preprocessing

Text normalization. Abstracts for epidemiologic studies often include the region in which the study was conducted and numerical statistics for prevalence data. The particular region and specific numerical values would add noise to the interpretation. Thus, we replaced all instances of percentage, geopolitical entities (countries, cities, and states), other locations, dates, times, quantities, ordinal values, and cardinal values with their entity types using the spaCy library.[22] In addition, we applied the scispaCy package[23] to normalize individual biomedical entities with their corresponding entity types, such as diseases, tissues, organs, and chemicals. We also removed stop words from the text. Figure 1 shows an example of the text normalization process for one abstract. All mentions of the specific disease, numeric values and geographic locations in this example have been normalized by their entity types.

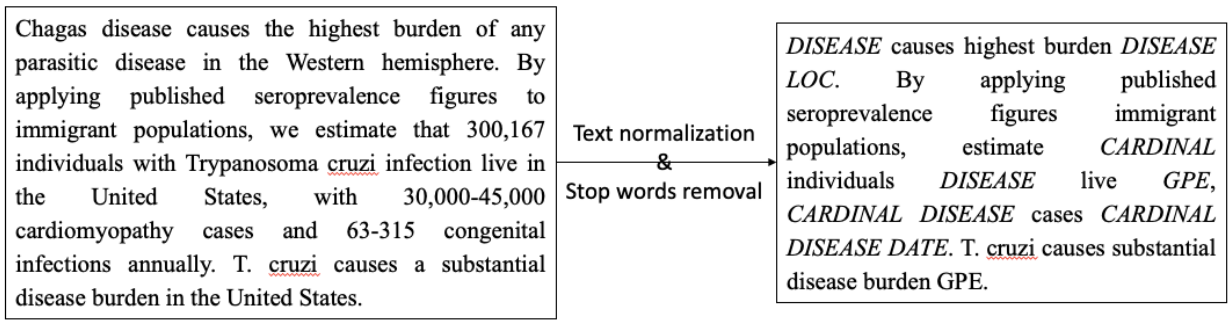


Figure 1: Text normalization example (the abstract is from [24])

Tokenization. We also fit a preprocessing tokenizer from the TensorFlow library on the training set. We limited our vocabulary size to 5,000, and the words that are not within the 5,000 most frequently used words in the training set are replaced with the <OOV> (out of vocabulary) token. In the example abstract shown in Figure 1, "seroprevalence," "immigrant," and "cruzi" were all replaced with "<OOV>," as these words do not occur frequently in rare disease texts. The tokenizer additionally vectorizes the set of abstracts and adds padding to standardize the length of the abstracts.

Recurrent neural network

We fit a shallow recurrent neural network on the training set. Figure 2 diagrams the model architecture. The network begins with an embedding layer, which converts the input into dense vectors representing the meaning of the abstract. The embedding layer is followed by two long short-term memory layers, the first with 64 units, and the second with 32 units. The output of the second LSTM layer feeds into a fully-connected (dense) layer with a ReLU activation function.[25] The final output layer is followed by the softmax activation function, which adjusts the output to create probabilities.[26] We used two LSTM layers as we found that this improved the model performance compared to one layer, and given the size of the dataset, we suspected that additional layers could cause overfitting. We begin with 64 units in the first LSTM layer to match the dimensionality of the embedding layer, and we decrease the dimensionality in the second LSTM layer to 32 to more densely represent the data. The model was compiled using the sparse categorical cross entropy loss function, and the Adam stochastic optimization function is applied.[27] To reduce overfitting, we used early stopping[28] with validation loss as the monitor. We set the maximum number of epochs to 10, as the preliminary results suggested that overfitting would compromise the performance with further epochs.

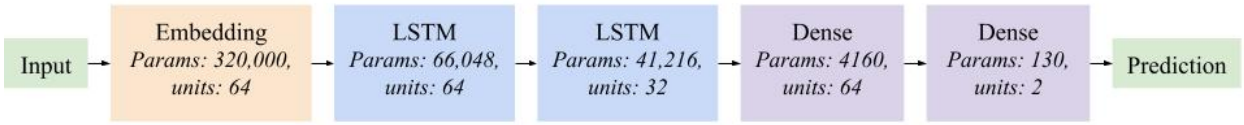


Figure 2: The RNN model architecture. "Params" indicates the number of trainable parameters in the layer, and "units" indicates the number of basic computational nodes.

Evaluation

We conducted three steps to evaluate our model. 1) The model was evaluated on the hold-out validation set of 5,275 abstracts, of which 295 were epidemiologic studies. From this set, we calculated precision, recall, F1 score, and area under the ROC curve (AUC). 2) One GARD curator manually validated the predictive results on the test set consisting of 100 abstracts, none of which were included in the training or validation sets. 3) To further assess the performance of the model with practical cases, we performed five case studies with five rare diseases, namely Tay-Sachs disease, Turner syndrome, sickle cell disease, cystic fibrosis, and Ehlers-Danlos syndrome. Specifically, for each disease, we identified epidemiologic studies from their top fifty PubMed articles retrieved via EBI API. We sorted the articles in

order of their predicted epidemiology probability and compared with our baseline results, which are the top five results by searching for the disease name and epidemiology related MeSH terms from PubMed.

Results

Dataset preparation

From the Orphanet epidemiology dataset, we extracted 10,845 articles with corresponding PMIDs. There are 4,691 PubMed articles associated with any MeSH terms. Of these, 1506 articles have been tagged with epidemiology-related MeSH terms ("Epidemiology," "Prevalence," or "Incidence") and were not categorized as case reports based on their publication types. After excluding 93 articles without abstracts, 1,413 articles comprised our positive set. Manual inspection on a sample set that confirmed that these articles represent epidemiologic studies. Figure 3 shows the results of creating the positive dataset.

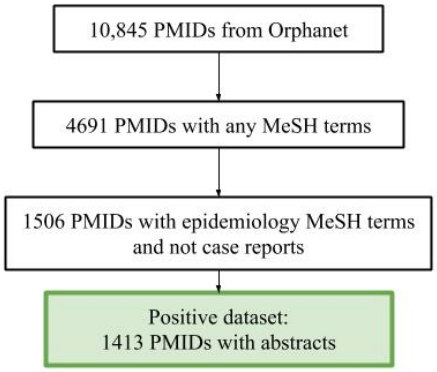


Figure 3: Stepwise results for the preparation of the positive dataset.

28,515 PubMed articles were retrieved for the 6,073 GARD rare diseases. Of these articles, we excluded 469 articles that are part of the Orphanet epidemiology dataset, and 3,056 articles with epidemiology related MeSH terms or keywords, leaving 24,990 articles in the negative dataset. Manual examination on randomly selected articles was performed and showed that they cover a wide spectrum of topics, including case reports, treatment explanations, genetic analyses, and general literature reviews of disorders. The results of the negative dataset preparation are shown in Figure 4.

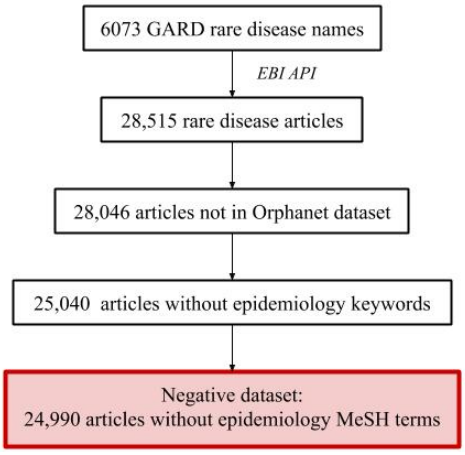


Figure 4: Stepwise results for the preparation of the negative dataset.

Table 1 provides the breakdown statistics of the dataset. In Discussion, we discuss the reason of having the imbalance in the training set and its influence on the model performance. Note that the total positive and negative dataset sizes were slightly reduced from the aforementioned numbers as articles in the test set were removed.

Table 1: The composition of the training and validation sets.

	Positive dataset (epidemiologic studies)	**Negative dataset (not epidemiologic studies)**	**Total**
Training set	1119	19,981	21,100
Test set	268	5007	5275
Total	1387	24,988	26,375

Holdout validation set evaluation

The recurrent neural network achieved promising results on the holdout validation set. Early stopping halted training after three epochs because of an increase in loss in the validation set. At this point, the precision on the validation set was 0.846, the recall was 0.937, the F1 score was 0.886, and the AUC was 0.967. The receiver operating characteristic (ROC) curve is given in Figure 5.

Overall, while the average epidemiology probability among the true positives was 0.966, the false positives received an average epidemiology probability of 0.892. Conversely, the epidemiology probability among the true negatives was 0.0229, while it was 0.0563 for false negatives. Of the 28 false negatives, only eight abstracts included epidemiologic information based on our manual review. Thus, given the focus of our study, the other twenty should be considered as true negatives, as our classification used only the abstract.

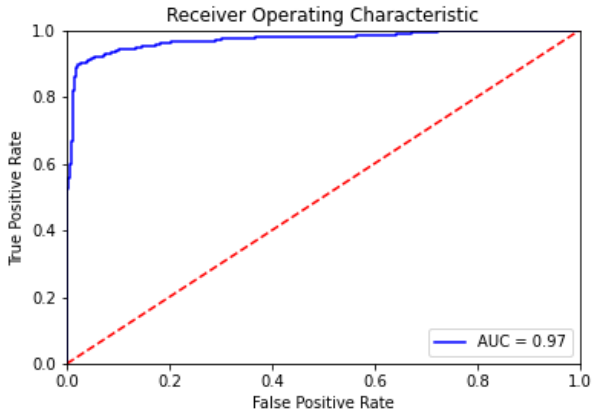


Figure 5: ROC curve for the holdout validation set.

Manual evaluation

A GARD curator manually validated the predictive results on the test set consisting of 100 articles, and these results suffered slightly compared to results on the holdout validation set. The precision was 0.726, the recall was 0.700, the F1 score was 0.701, and the AUC was 0.751. We discuss the reasons behind this discrepancy in the Discussion section.

Of the twenty false negatives based on the test result, twelve articles described epidemiologic information only in the full text, instead of in the abstracts themselves, such as with the two article "Chromosome 1p36 deletions: the clinical phenotype and molecular characterization of a common newly delineated syndrome"[29] and "Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype".[30] Thus, these errors were likely an artifact of the differing

focus of the manual evaluation. The false positives included genealogy and genetics studies, a case report, and two epidemiologic studies in geographic regions that were too constricted for use by GARD.

Case studies

On the case studies we performed for five rare diseases, our model generally successfully identified epidemiologic studies from PubMed. We set the probability threshold for an epidemiology article to be 0.5, and additionally included the exact probability for analysis. In most cases, the results returned with our method were more relevant than those found via filtering by epidemiology related MeSH terms from PubMed. The PubMed search query was composed as "(((epidemiology[MeSH Terms]) OR (prevalence[MeSH Terms])) OR (incidence[MeSH Terms])) AND (Disease Name)", where "Disease Name" is replaced with the specific disease name.

Tay-Sachs disease.[31] Four of the five articles that are predicted as epidemiologic studies by our model contain epidemiologic information, as shown in Figure 6. The article without epidemiologic information was ranked fourth of the five and had an epidemiology probability of 0.644. In contrast, out of the top five results from the manual PubMed search for Tay Sachs with epidemiology related MeSH terms, only one article titled "Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population"[32] was epidemiology related and contained minimal information on Tay-Sachs disease. None of the four epidemiology articles discovered by our model appeared in the PubMed search results.

Turner syndrome.[33] Two articles were predicted as epidemiologic studies. The first article does in fact give the prevalence of the syndrome,[34] while the second article described the risk of coronary artery disease, a known clinical complication amongst Turner syndrome patients.[35] The manual PubMed search does not include any epidemiologic studies in the top five results; notably, two of them were not related to Turner syndrome at all, and another two articles detail bone fragility and autoimmune thyroid disease in Turner syndrome, but are not epidemiologic studies.

Sickle cell disease (SCD).[36] Of the four articles predicted as epidemiologic studies for SCD, one stated a rough estimate for its prevalence in the United States,[37] one referred to the millions of patients affected worldwide,[38] one compared the prevalence of priapism in those with and without SCD,[39] and one detailed an approach to treatment.[40] None of the results from the manual PubMed search were epidemiologic studies or provided epidemiologic information in their abstracts.

Cystic fibrosis.[41] One positive result from the model for cystic fibrosis described the prevalence of fungal disease within the disorder.[42] None of the results from the manual PubMed search were epidemiologic studies, although one provided an estimate for the worldwide prevalence of the disease.[43]

Ehlers-Danlos syndrome.[44] One of the two positive results generated from the model, detailed the prevalence of cardiovascular disorders in patients with this syndrome.[45] The other was did not involve epidemiology.[46] One result classified as negative did include a prevalence statistic, but the topic of the article was surgical outcomes.[47] None of the manual search results were epidemiologic studies or included epidemiologic information for Ehlers-Danlos syndrome.

	PMID	Title	Probability of epidemiology	Relevant text
0	32302469	The incidence and carrier frequency of Tay-Sachs disease in the French-Canadian population of Quebec based on retrospective data from 24 years, 1992-2015.	0.999	This corresponds to an incidence of 1/218,144, which in turn corresponds to a carrier frequency of 1/234.
1	29943104	Presentation of central precocious puberty in two patients with Tay-Sachs disease.	0.998	The disease is very rare in Turkey, with an incidence of 0.54/100,000
2	30506202	Prenatal Diagnosis of Tay-Sachs Disease.	0.823	TSD is more prevalent among Ashkenazi Jewish (AJ) individuals and some other genetically isolated populations with carrier frequencies of approximately ~1:27 which is much higher than that of 1:300 in the general population
3	30616450	Patient-Derived Phenotypic High-Throughput Assay to Identify Small Molecules Restoring Lysosomal Function in Tay-Sachs Disease.	0.644	None
4	31076878	Amyotrophy, cerebellar impairment and psychiatric disease are the main symptoms in a cohort of 14 Czech patients with the late-onset form of Tay-Sachs disease.	0.622	(a calculated birth prevalence of 1 per 325,175 live births)

Figure 6: Predictive results generated for the case study of Tay-Sachs disease.

Discussion

Epidemiologic studies provide insights and directions for basic and clinical research to determine the causes and mechanisms of rare diseases and develop methodologies for prevention, diagnosis, and treatment. However, epidemiologic data curation in the rare disease field continues to rely heavily on human effort, from identification of epidemiologic studies from PubMed to data curation. In this study, we presented a computational model by applying recurrent neural networks and NLP techniques to programmatically identify epidemiologic studies from PubMed. This work can reduce the human effort required from the epidemiologic data curation process and holds promise for other applications beyond rare diseases and with other types of studies.

Quantitatively, our model performed very well on the holdout validation set, with a high AUC of 0.967. Our manual inspection of the results further proved that our model can consistently assign high epidemiology probabilities (above 0.98) for standard epidemiologic studies, and strong correlation is found between the predicted epidemiology probability and the amount of epidemiologic information mentioned in the abstract. For example, an article titled "Birth prevalence of Prader-Willi syndrome in Australia", whose abstract details an epidemiologic study,[48] obtained an epidemiology probability of 0.999. However, the article titled "Th17 cytokine deficiency in patients with Aspergillus skull base osteomyelitis", which is a molecular study,[49] is predicted to have an epidemiology probability of 0.00956. In addition, the five case studies demonstrated that this model was effective at surfacing epidemiologic studies for individual diseases. Compared to the baseline results with manual PubMed search, our model captured more epidemiologic studies, which were not part of the top five results, or were even not found in the entire list of PubMed search results. However, we observed the performance of the model on the test set was not as promising as the holdout validation set. Our analysis indicated this discrepancy was likely due to our focus on the content of the abstract, while the curators often examined the full text in addition to the abstract when labeling the dataset.

Notably, our model reached satisfying performance even with a dataset that is small and imbalanced: non-epidemiology articles outnumbered epidemiologic studies by roughly 20:1. Initially, we expanded our positive dataset by including articles tagged with epidemiology-related MeSH terms that were not referenced by Orphanet. However, this did not significantly improve the performance. This was likely because some of the MeSH terms may have been assigned incorrectly, whereas restricting the dataset to those also used by Orphanet added another layer of confirmation that the articles were likely related to epidemiology. The success of our model in light of this illustrates that the features of an epidemiologic study are easily identifiable and significantly distinct from those characteristic of case reports, clinical guidelines, genetic analyses, and other types of studies. For instance, of the 1702 case reports in the validation set, only 17 were predicted as epidemiologic studies. Since case reports rarely include epidemiology information about a disease, this result suggests that the model was able to identify features distinguishing case reports from epidemiologic studies.

Given the lack of available training data relating to epidemiology, we used a combination of Orphanet data, MeSH terms, and keyword searches to generate our dataset. This approach could introduce bias based on the types of sources selected by Orphanet and the process used to assign MeSH terms. The strategy of generating the negative set by excluding abstracts containing epidemiology keywords set might also bias the model toward over-relying on keywords to generate its predictions rather than more sophisticated linguistic features. We did not observe significant negative impact as a result, but a follow-up analysis could better characterize any bias. Relatedly, a more robust evaluation of the model from a larger and more consistently labeled dataset would assist in confirming our results.

The computational approach established in this study will be able to support the task of supplementing epidemiology curation for GARD and other applications in multiple ways. First, our model can identify and rank epidemiologic studies relating to rare diseases. This would allow curators to begin by reviewing the articles with the highest predicted epidemiology probability, rather than searching for relevant articles manually. Second, the model could be integrated into an alert system to notify curators about the publication of new epidemiologic studies. From a set of epidemiologic studies identified by the model, we could apply information extraction to their text following previous work[50], which could lead to a process to fully automate the curation of epidemiology data.

Furthermore, there are several directions for expanding this work. A deeper analysis into the results of our model could reveal features or patterns in its predictions that would allow the model to be refined to achieve better performance, as the interpretability of the model at present is limited. The addition of more data, particularly epidemiology articles, could also improve performance. In this study, we limited the dataset to articles addressing rare diseases as this was the immediate use case of the model, and this approach accounts for any unique structural and content features of rare disease epidemiologic abstracts. In future work, epidemiologic studies addressing diseases that are not rare may also be included. Because the text processing steps remove the specific disease features, this

change will likely improve the capacity of the model to identify rare disease epidemiologic studies, as the benefit of increasing the size of the dataset could outweigh any noise that is introduced. Furthermore, an expanded dataset could allow for more advanced approaches such as Bidirectional Encoder Representations from Transformers (BERT)[51] or a deeper neural network architecture; these approaches were not used in this study due to concerns about overfitting on a limited dataset. In order to capture epidemiologic information beyond epidemiologic studies, our model framework could be applied to identify case reports, as these can be aggregated to determine case or family counts. When we combined case reports with epidemiologic studies in our dataset, the model performance suffered, likely because the structure and content of case reports differ significantly from epidemiologic studies. However, case reports could be considered independently in a separate model. Similarly, because of the generalizability of neural networks, our approach could also be used to develop classifiers for natural history studies or clinical trials, and in other domains beyond rare diseases.

Conclusion

In this paper, we demonstrated that a recurrent neural network with long short-term memory architecture achieved good performance in classifying epidemiologic studies of rare diseases. Our model can be leveraged to greatly shorten the manual curation process for evidence selection in curating epidemiologic information. We hypothesize that the success of our model suggests that our approach can be applied to other similar tasks such as classifying natural history studies and in other medical domains.

Acknowledgements

This research was supported in part by the Intramural/Extramural research program of NCATS, NIH. The authors thank Karen Hanson, from ICF International, Inc. for her help on manual evaluation; Dac-Trung Nguyen, from Division of Pre-Clinical Innovation, NCATS, participated in the valuable discussion. Dr. Anne Pariser, as Director of the Office of Rare Disease Research (ORDR), at NCATS, supported this work and also participated in the valuable discussion. Lastly, we thank the NIH Office of Data Science Strategy and the HHS Civic Digital Fellowship for supporting these efforts.

References

1. Rare Diseases Act of 2002 Congress, 107th Sess. (2002).
2. Dawkins HJ, Draghia-Akli R, Lasko P, et al. Progress in rare diseases research 2010–2016: an IRDiRC perspective. Clinical and Translational Science. 2018;11(1):11.
3. Griggs RC, Batshaw M, Dunkle M, et al. Clinical research for rare disease: opportunities, challenges, and solutions. Molecular Genetics and Metabolism. 2009;96(1):20-6.
4. Hassell KL. Population estimates of sickle cell disease in the US. American Journal of Preventive Medicine. 2010;38(4):S512-S21.
5. Jansen Type Metaphyseal Chondrodysplasia: NORD - National Organization for Rare Disorders; 2018 [Available from: https://rarediseases.org/rare-diseases/jansen-type-metaphyseal-chondrodysplasia/
6. Boat TF, Field MJ. Rare diseases and orphan products: Accelerating research and development: National Academies Press; 2011.
7. Wakap SN, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. European Journal of Human Genetics. 2020 Feb;28(2):165-73.
8. GARD Information Center [Available from: https://rarediseases.info.nih.gov/.
9. Schuemie MJ, Sen E, 't Jong GW, van Soest EM, Sturkenboom MC, Kors JA. Automating classification of free-text electronic health records for epidemiological studies. Pharmacoepidemiology and Drug Safety. 2012;21(6):651-8.
10. Rios A, Kavuluru R, editors. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics; 2015.
11. Cohen AM. An effective general purpose approach for automated biomedical document classification. AMIA annual symposium proceedings; 2006: American Medical Informatics Association.
12. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735-80.
13. Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:150600019. 2015.
14. Tang D, Qin B, Liu T, editors. Document modeling with gated recurrent neural network for sentiment classification. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015.

15. Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:160908144. 2016.
16. Graves A, Jaitly N, editors. Towards end-to-end speech recognition with recurrent neural networks. International Conference on Machine Learning; 2014.
17. Li L, Jin L, Jiang Z, Song D, Huang D, editors. Biomedical named entity recognition based on extended recurrent neural networks. 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2015: IEEE.
18. Lu H, Li L, He X, Liu Y, Zhou A. Extracting chemical-protein interactions from biomedical literature via granular attention based recurrent neural networks. Computer Methods and Programs in Biomedicine. 2019;176:61-8.
19. Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000;88(3):265.
20. Epidemiological Data. In: Orphanet, editor. orphadata.org2020.
21. Burke M, Armstrong D, Carvalho-Silva D, et al. EMBL-EBI, programmatically: take a REST from manual searches. European Bioinformatics Institute (EMBL-EBI); 2017.
22. spaCy: Explosion AI; 2020 [Available from: https://spacy.io/.
23. Neumann M, King D, Beltagy I, Ammar W. Scispacy: Fast and robust models for biomedical natural language processing. arXiv preprint arXiv:190207669. 2019.
24. Bern C, Montgomery SP. An estimate of the burden of Chagas disease in the United States. Clinical Infectious Diseases. 2009;49(5):e52-e4.
25. Nair V, Hinton GE, editors. Rectified linear units improve restricted boltzmann machines. ICML; 2010.
26. Bridle J. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. Advances in Neural Information Processing Systems. 1989;2:211-7.
27. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
28. Caruana R, Lawrence S, Giles CL, editors. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in Neural Information Processing Systems; 2001.
29. Shapira SK, McCaskill C, Northrup H, et al. Chromosome 1p36 deletions: the clinical phenotype and molecular characterization of a common newly delineated syndrome. The American Journal of Human Genetics. 1997;61(3):642-50.
30. Zollino M, Orteschi D, Murdolo M, et al. Mutations in KANSL1 cause the 17q21. 31 microdeletion syndrome phenotype. Nature Genetics. 2012;44(6):636-8.
31. Tay-Sachs disease [Available from: https://rarediseases.info.nih.gov/diseases/7737/tay-sachs-disease.
32. Rivas MA, Avila BE, Koskela J, et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. PLoS Genetics. 2018;14(5):e1007329.
33. Turner syndrome [Available from: https://rarediseases.info.nih.gov/diseases/7831/turner-syndrome.
34. Abu-Halima M, Oberhoffer FS, El Rahman MA, et al. Insights from circulating microRNAs in cardiovascular entities in turner syndrome patients. PLoS One. 2020;15(4):e0231402.
35. Funck KL, Budde RPJ, Viuff MH, et al. Coronary plaque burden in Turner syndrome a coronary computed tomography angiography study. Heart Vessels. 2020.
36. Sickle cell disease [Available from: https://www.genome.gov/genetics-glossary/Sickle-Cell-Disease.
37. Fantasia HC, Morse BL. Voxelotor for the treatment of sickle cell disease. Nurs Womens Health. 2020;24(3):233-7.
38. Pavan AR, Dos Santos JL. Advances in sickle cell disease treatments. Curr Med Chem. 2020.
39. Idris IM, Abba A, Galadanci JA, et al. Men with sickle cell disease experience greater sexual dysfunction when compared with men without sickle cell disease. Blood Adv. 2020;4(14):3277-83.
40. Herity LB, Vaughan DM, Rodriguez LR, Lowe DK. Voxelotor: a novel treatment for sickle cell disease. Ann Pharmacother. 2020:1060028020943059.
41. Cystic fibrosis [Available from: https://rarediseases.info.nih.gov/diseases/6233/cystic-fibrosis.
42. Cuthbertson L, Felton I, James P, et al. The fungal airway microbiome in cystic fibrosis and non-cystic fibrosis bronchiectasis. J Cyst Fibros. 2020.
43. Baiardini I, Steinhilber G, DI Marco F, Braido F, Solidoro P. Anxiety and depression in cystic fibrosis. Minerva Med. 2015;106(5 Suppl 1):1-8.
44. Ehlers-Danlos syndrome [Available from: https://www.cedars-sinai.org/health-library/diseases-and-conditions/e/ehlers-danlos-syndrome-eds.html.
45. Paige SL, Lechich KM, Tierney ESS, Collins RT. Cardiac involvement in classical or hypermobile Ehlers-Danlos syndrome is uncommon. Genet Med. 2020.

46. Miller AJ, Schubart JR, Sheehan T, Bascom R, Francomano CA. Arterial elasticity in Ehlers-Danlos syndromes. Genes (Basel). 2020;11(1).
47. Louie A, Meyerle C, Francomano C, et al. Survey of Ehlers-Danlos patients' ophthalmic surgery experiences. Mol Genet Genomic Med. 2020;8(4):e1155.
48. Smith A, Egan J, Ridley G, et al. Birth prevalence of Prader-Willi syndrome in Australia. 2003;88(3):263-4.
49. Delsing CE, Becker KL, Simon A, et al. Th17 cytokine deficiency in patients with Aspergillus skull base osteomyelitis. BMC Infectious Diseases. 2015;15(1):140.
50. Karystianis G, Thayer K, Wolfe M, Tsafnat G. Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews. Journal of Biomedical Informatics. 2017;70:27-34.
51. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

Generative Adversarial Networks for Creating Synthetic Free-Text Medical Data: A Proposal for Collaborative Research and Re-use of Machine Learning Models

Suranga N. Kasthurirathne, PhD[1,2], Gregory Dexter[3], Shaun J. Grannis MD, MS[1,2]
[1]Regenstrief Institute, Indianapolis, IN, USA; [2]Indiana University School of Medicine, Indianapolis, IN, USA; [3]Purdue University, Indianapolis, IN, USA

Abstract

Restrictions in sharing Patient Health Identifiers (PHI) limit cross-organizational re-use of free-text medical data. We leverage Generative Adversarial Networks (GAN) to produce synthetic unstructured free-text medical data with low re-identification risk, and assess the suitability of these datasets to replicate machine learning models. We trained GAN models using unstructured free-text laboratory messages pertaining to salmonella, and identified the most accurate models for creating synthetic datasets that reflect the informational characteristics of the original dataset. Natural Language Generation metrics comparing the real and synthetic datasets demonstrated high similarity. Decision models generated using these datasets reported high performance metrics. There was no statistically significant difference in performance measures reported by models trained using real and synthetic datasets. Our results inform the use of GAN models to generate synthetic unstructured free-text data with limited re-identification risk, and use of this data to enable collaborative research and re-use of machine learning models.

Introduction

Rapid uptake of Health Information Systems (HIS) has enabled the accessibility and availability of structured and unstructured electronic health data. These data, together with the rapid evolution of Artificial Intelligence (AI) and various analytical and machine learning toolkits has led to the widespread development of machine learning solutions(1, 2) designed to address organizational-level challenges using organizational-level data. However, the current U.S. regulatory framework limits sharing of Patient Health Identifiers (PHI) outside the healthcare organization(3). Limited or burdensome data access hinders (a) sharing and re-using machine learning solutions across larger audiences, (b) promoting inter-organizational collaboration addressing various healthcare challenges, and (c) building generalized machine learning models targeting diverse populations.

There have been significant efforts to de-identify structured and unstructured patient data for research and dissemination purposes(4, 5). Traditional de-identification efforts focus on the perturbation of potentially identifiable patient demographic attributes such as names, addresses, identifiers, and contact information via randomization, suppression or generalization(6, 7). However, such efforts are not foolproof – patient records scrubbed of PHI may be susceptible to re-identification based on residual clinical information contained in symptoms, diagnosis, medications or lab results(8). This significantly impacts de-identification of structured data due to difficulty in identifying potentially sensitive information from free-text data. Researchers have proposed various approaches for creating synthetic data that mimics clinical patterns in medical records as a solution to re-identification risk based on clinical information(9). A synthetic patient dataset that has been scrubbed of any PHI elements using traditional de-identification methods would be significantly harder to re-identify than a real dataset that has only been scrubbed of PHI elements. However, previous synthetic data generation efforts have resulted in data that are not sufficiently realistic for machine learning(7).

Generative Adversarial Networks (GAN) are a class of deep learning algorithms that offer significant promise to improve synthetic data generation. GAN algorithms are implemented by a system of two neural networks(10). One neural network, the generator, attempts to create synthetic data, while the other neural network, the discriminator, seeks to distinguish between synthetic data and real data. As these networks are trained, the generator network successfully develops synthetic data that cannot be flagged by the discriminator. Initial GAN models were designed to mimic real-valued data(10). As such, they have been used to produce high quality categorical(11) and image datasets(12, 13). In the healthcare domain, GAN models have been used to generate numerical clinical data that is statistically similar to real data(7, 14).

Recent improvements to GAN algorithms enable them to generate synthetic free-text data(15). Researchers have applied these models to successfully generate text data such as molecules encoded as text sequences, musical melodies(16), reviews, dialogues(17), poetry and image captions(18). These innovations offer much potential to the medical field, where a large quantity of clinical information may be trapped within unstructured free-text(19, 20).

We evaluate the potential to leverage GAN models to produce synthetic unstructured free-text medical data that closely reflect characteristics of real data, and thus, may be used to develop machine learning models that approximate similar models created using original data. Next, we will assess the re-identification potential of these synthetic, informationally similar, unstructured datasets.

Materials and methods

Test data selection

We extracted all laboratory messages pertaining to cases of Salmonella reported to the Indiana Network for Patient Care (INPC)(21) during 2016-2017. The INPC is a statewide Health Information Exchange (HIE) that facilitates interoperability across 117 hospitals, 38 health systems, other free-standing laboratories, and physician practices across the state of Indiana. We parsed these messages, which were obtained in the form of Health Level Seven (HL7) version 2 messages, and extracted the free-text report data included in each message. Laboratory messages for salmonella were selected due to the semi-structured nature of the HL7 messages, which allowed us to separate PHI from the unstructured text, as well as the brevity of the free-text laboratory messages. Each message was manually reviewed, and labelled as positive or negative for Salmonella. We randomly selected 90% of each of the positive and negative salmonella messages, hereafter known as positive (train) and negative (train) datasets for training GAN models. The remainder of the datasets, hereafter known as the positive (holdout) and negative (holdout) datasets, were used to test the performance of GAN generated data.

Development of GAN models for synthetic data generation

We adopted SeqGAN, a GAN algorithm designed to generate textual data(22). SeqGAN models approach the sequence generation procedure as a sequential decision-making process. The generative model is treated as an agent of reinforcement learning; the state is the generated tokens while the action is the next token to be generated. The discriminator evaluates the sequence and feeds back the evaluation to guide the learning of the generative model(22). GAN models consist of a number of parameters that can be fine-tuned to optimize model performance. We explored model performance by training multiple GAN models using the positive (train) and negative (train) datasets, and varying several parameters (Appendix A). We adopted a Gaussian distribution as the default initial parameter for all generators. Performance of these models were compared using two document similarity based metrics; embedding similarity, which measures similarity between two documents using similarity between word embeddings, and NLL-test, which evaluates a model's capacity to fit real test data(15). Optimal models selected using this approach were used to generate positive (synthetic) and negative (synthetic) laboratory messages. To build compatible decision models, we generated n positive synthetic reports, where n equals the number of positive (train) messages, and m negative synthetic reports, where m equals the number of negative (train) messages. We trained SeqGAN models using Texygen, a benchmarking platform for GAN based text generation models(15).

Machine learning process

Features extracted from the positive (train) and negative (train) datasets, (jointly known as the real dataset), and positive (synthetic) and negative (synthetic) datasets (jointly known as the synthetic dataset) were used to train multiple classification models using the following approach.

Feature extraction

We mimicked the feature extraction process adopted in our previous work on predicting cancer cases using free-text data obtained from the INPC(23, 24). We developed a Perl script to parse the positive and negative training datasets, and identify all unique stemmed tokens present within these reports. Next, we used the Negex algorithm(25) to identify the context of use (positive or negative) for each stem. We counted the presence of each feature in positive and negated context, and used this data to prepare an input vector for each laboratory message. A similar approach was used to generate vectors of counts representing each message in the synthetic dataset.

Decision model building and evaluation

We applied the Gini impurity(26) metric to rank features in the real and synthetic datasets by order of importance. We used subsets of the top 5, 10, 15 and 20 features selected from the real and synthetic datasets to train a series of decision

models using the Random Forest classification algorithm(27). Random forest was selected due to its proven track record in health care decision-making applications(24, 28, 29). The real and synthetic decision models were tested using feature vectors derived from the positive and negative holdout datasets. We calculated sensitivity (True Positive Rate or Recall), specificity (True Negative Rate), F1-Measure and Area Under the ROC Curve (AUC) for each decision model. Paired t-tests were used to compare the performance of synthetic and real data-based decision models.

Evaluation of re-identification risk

Risk of presence disclosure, aka membership inference assesses an attackers ability to determine if any real patient records in their possession were used to train GAN models by comparing these records against the synthetic patient dataset(30, 31). Assessing risk of presence disclosure ensures the privacy of individuals whose data was used to train a decision model(32), as well as the interests of the healthcare entity where the individual received treatment(33). Thus, presence disclosure is a widely evaluated measure of re-identification risk(34, 35). We assessed risk of presence disclosure using the following experiment(7); we re-purposed vectors that represented the training and synthetic datasets using binary values representing the presence or absence of each feature in positive and negative context. We compared each synthetic record with all training messages using various hamming distance cutoff scores(36), a measure of the minimum number of substitutions required to change one string into the other. The identification of a synthetic record that matched with any training message using a hamming distance equal or smaller to the hamming score threshold would label it as a 'match' to the synthetic record under study. We computed the frequency of matches across each hamming score threshold, and used these metrics to evaluate re-identification risk.

Figure 1 presents our complete workflow, from data extraction to decision model evaluation.

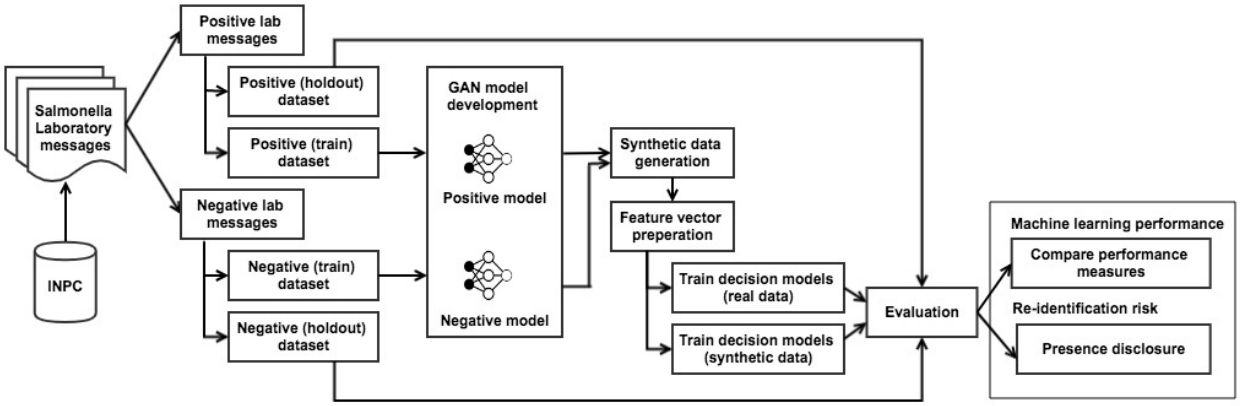


Figure 1. A workflow depicting our study approach from laboratory message extraction to decision model evaluation.

Results

We identified a total of 6,770 laboratory messages pertaining to salmonella. Manual review labelled 1,213 (17.91%) of these messages as positive, and 5,557 (82.08%) as negative. We identified optimal SeqGAN models for generating positive and negative laboratory messages using hyperparameters identified in appendix A. Using these models, we generated 1092 positive synthetic messages and 5001 negative synthetic messages to correspond with the 90% training messages for each dataset. Appendix B presents representative samples from the positive (train), negative (train), positive (synthetic) and negative (synthetic) lab report sets. These samples were manually reviewed, and any PHI elements masked. As seen in appendix B, the only PHI elements identified within these report sets were date, time and report identifier fields. We computed several Natural Language Generation (NLG) measures to evaluate similarity between real and synthetic datasets; a) Bilingual Evaluation Understudy (BLEU) scores(37) are widely used to compare similarity between real and synthetic datasets. We calculated BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores that evaluated the quality of synthetic datasets using 1-gram, 2-gram, 3-gram and 4-gram matches respectively. b) Google-BLEU (GLEU) scores, a measure that seeks to address limitations in BLEU score calculations and are better suited for sentence level comparisons(38). The GLEU score is a composite of all 1-grams, 2-grams, 3-grams and 4-gram matches (table 1).

NLG measure	Positive (train) vs. Positive (synthetic)	Negative (train) vs. Negative (synthetic)
BLEU-1	0.913	0.944
BLEU-2	0.675	0.742

BLEU-3	0.480	0.552
BLEU-4	0.331	0.409
Google-BLEU	0.249	0.328

Table 1. Comparison of real and synthetic datasets using various NLG measures.

These results are comparable to those produced by prior researchers(15), and indicate considerable similarity between real and synthetic datasets. Further, NLG measures comparing negative (train) and negative (synthetic) datasets were higher than the positive (train) and positive (synthetic) datasets. We hypothesize this is because negative reports are more similar due to uniform text documenting negative status. The positive (train) dataset comprised of 2,551 unique stemmed features. 1,827 (71.6%) of these features were present within the positive (synthetic) dataset. The negative (train) dataset comprised of 5,803 unique stemmed features. 4,093 (70.5%) of these stemmed features were present within the negative (synthetic) dataset. With stop words and dates removed, the overall training dataset of positive and negative reports consisted of 3810 unique stemmed features. 2651 (69.6%) of these were present within the overall synthetic dataset. Appendix C lists the top 20 features identified across the real and synthetic datasets using gini impurity scores. Appendix D presents the overlap between the top 5, 10, 15, 20, 50 and 100 features identified across the real and synthetic datasets. We note significant similarity between real and synthetic datasets with between 70% to 80% overlap across each of the feature subsets being compared. Figure 2. presents the sensitivity, specificity, F1-measure and Area under the ROC curve scores reported by decision models built using the top 5, 10, 15 and 20 real and synthetic features upon being tested using the holdout test datasets.

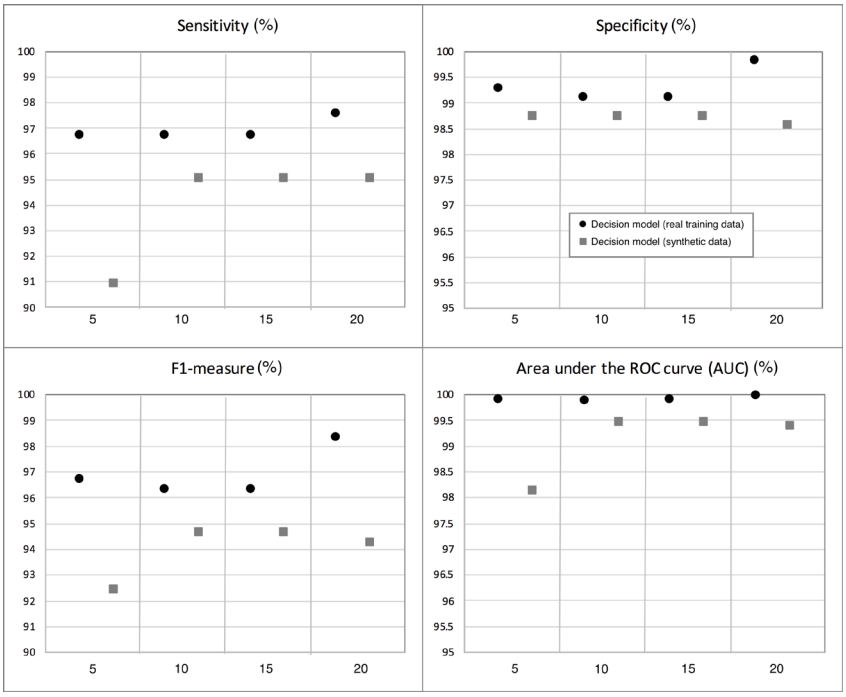


Figure 2. Sensitivity, specificity, F1-measure and Area under the ROC curve scores reported by decision models built using the top 5, 10, 15 and 20 real and synthetic features upon being tested using the holdout test datasets.

Due to the discriminatory power of the features, each model achieved high-performance measures despite being trained on a small number of features. Further, paired t-tests reported statistical significance levels (alpha) of > 0.65. As such, there is no significant difference between performance measures reported by real and synthetic decision models built using any of the feature subset sizes. Given high overlap between top 50 and 100 feature sets (appendix D), we hypothesize that decision models built using the top 50 and 100 real and synthetic features would also report statistically similar performance measures.

Evaluation of re-identification risk

Results of the presence disclosure test are presented in appendix E. We conclude that these results indicate acceptable levels of re-identification risk given that the number of positive matches identified by a hamming threshold of 10 was reasonably small.

Discussion

Our results further two challenges; the use of GAN models to generate synthetic free-text medical data with limited re-identification risk, and use of these datasets to develop machine learning models with statistically similar performance metrics to models developed using the original test data, thereby enabling cross-institutional collaboration and broader dissemination of machine learning models.

Comparison of unique features across test and synthetic datasets revealed that the synthetic dataset contained only 69.6% of the features in the test dataset. We attribute this to the mode collapse problem(39), which leads to reduced diversity of synthetic data(15, 17). NLG scores reported by our models were compatible to scores reported by other efforts to generate synthetic text data extracted from non-medical sources(15). However, we note that the synthetic data presented reduced syntactic/grammatical correctness (appendix B), a common pitfall in deep learning based text generation approaches(40). However, this was irrelevant for our purposes as we only sought to demonstrate that synthetic data could be used to replicate machine learning performance, and not as a tool for training or teaching of humans. Thus, no human evaluation of the synthetic reports was performed. The synthetic dataset also contained a quantity of recurring phrases such as hospital and laboratory test names. The recurrence of such phrases may have positively influenced NLG scores. Despite these limitations, there was 70-80% overlap between the top 5, 10, 15, 20, 50 and 100 features extracted from both datasets (appendix D). Performance measures generated by models trained using top 5, 10, 15 and 20 features extracted from the test and synthetic datasets were high, as well as statistically similar. These results present the possibility of using synthetic datasets to share machine learning solutions, and foster cross-institutional collaboration on various challenges.

Our findings help inform data de-identification efforts. As discussed previously, de-identification efforts involve (a) removal of PHI elements, and (b) addressing re-identification risk based on clinical information in patient records. Adoption of GAN models alone do not result in de-identified data. However, synthetic data generation reduces re-identification risk by creating new patient records with similar, but different content. It also removes any 1-to-1 mapping between test and synthetic reports. Our results using presence disclosure tests confirmed that synthetic datasets pose a small chance of re-identification based on clinical information. However, synthetic data produced by these efforts must undergo rigorous de-identification of PHI elements before they can be distributed for public use. Removal of PHI elements will not impact decision model performance as the top 100 features listed in appendix D did not include any PHI elements.

An alternate approach to evaluate re-identification risk is attribute disclosure, which evaluates an attackers ability to derive additional attributes (features) for a patient based on a subset of attributes they are aware of(41). We did not evaluate our datasets for attribute disclosure as a) unlike longitudinal patient datasets that consist of varied clinical diagnoses that are not necessarily related, the salmonella lab reports consisted of very specific features that are often highly correlated. Thus, it would be relatively easy to predict missing features in our dataset based on those present. Secondly, a considerable number of features presented very low prevalence across the laboratory reports. Thus, predicting 'absence' of these features was relatively easy. However, we argue that these factors pose low risk to the patient because unlike studies that deal with clinical diagnosis that if revealed, may impact the patient's privacy, our dataset focusses on salmonella alone. As an example, discovery of any of the top features listed in appendix C using an attribute disclosure attack would not lead to any harm beyond the awareness that the patient was tested for, and diagnosed as positive or negative for Salmonella. In contrast, attribute disclosure across a different dataset may lead to discovery of multiple clinical diagnosis, patient demographics or other treatment information. We propose the following hypothetical scenario to demonstrate how our approach could be applied in a real-life setting; An organization that possesses rich free-text data sources, but lacks adequate machine learning expertise can leverage our approach to create synthetic data. They de-identify and share the synthetic data with experts who use it to build machine learning models. Once optimal models have been identified, they can be implemented across the original dataset with compatible performance measures.

We identified a number of limitations in our study. Our test dataset consisted of structurally similar reports describing a very specific illness. This, together with the overall simplicity of our predictive outcomes (positive vs. negative for salmonella) may have contributed to our positive results. Datasets that are not structurally similar, nor restricted to a specific illness, or consist of more colloquial language may be harder mimic, and thus, produce less optimal results. Such datasets may require more robust decision models built using other free-text friendly GAN models such as Maximum-Likelihood augmented discrete Generative Adversarial Networks (MaliGAN)(42) or Long Text Generative Adversarial Networks (LeakGAN)(18), and more complex feature vectors consisting of n-grams. Further, our approach was restricted to mimicking synthetic free-text data. It is unclear if our models are able to learn or mimic the

significance of various numeric values such as age or other measurements present in free-text data. This limitation did not impact the performance of our current effort as no numerical values were selected as top features. However, it may impact models built using other datasets.

Future research avenues include use of GAN models to create truly de-identified synthetic free-text data that does not require additional de-identification, and expansion of our work across other more challenging healthcare datasets. Other researchers have demonstrated the ability to mimic numerical and categorical patient data using GAN models(7, 14). Integrating these efforts with ours would enable researchers to share comprehensive synthetic patient health records consisting of both structured and unstructured data for secondary research purposes. Furthermore, our study did not include any analysis of the readability or syntactic/grammatical accuracy of the synthetic reports. As such, these results are suitable for machine learning, and not for teaching or learning resources. Next steps include a) manual assessment of the readability and correctness of synthetic reports using human experts (Turing test), and b) investigation of other word and grammar-based measures that inform synthetic data assessment.

Conclusions

GAN models can be used to generate synthetic unstructured free-text medical data that can be used to replicate the performance of machine learning models with high, as well statistically similar results. Further, synthetic datasets poise limited risk of re-identification based on clinical features. As such, these synthetic datasets can be easily de-identified, and used to champion cross-organizational collaboration efforts.

Appendices

Appendix A. List of hyperparameters evaluated as part of the SeqGAN training process.

Parameter name	Description	Variations attempted
Pre-training epochs	The generator is trained for n epochs, followed by n epochs for the discriminator	Increments of 5 between 10 and 100
Adversarial epochs	Number of adversarial epochs	Increments of 5 between the values 5 and 50
Embedding dimensions	Dimensionality of embedding layer	32, 64, 128
Hidden dimensions	Number of neurons in hidden layer	32, 64, 128
sequence length	Length of each training sequence	Increments of 10 between the values 10 and 120

Appendix B. Representative samples of the train and synthetic datasets with HL7 tags removed.

Positive (train) messages

A) culture in progress. identifications performed by maldi tof mass spectrometry were developed and performance characteristics were determined by pcl alverno hammond in. salmonella species numerous. susceptibility not routinely performed. gastroenteritis due to non typhoidal salmonella spp. is generally self limiting in patients without underlying medical issues. for salmonella typhi isolates azithromycin is the drug of choice. identified by maldi tof mass spectrometry. sent to indiana state department of health.
B) additional organisms present as probable contaminants. salmonella species 100000 cfu/mlthis strain tested resistant to naladixic acid. treatment of extraintestinal salmonella infections may not be eradicated by fluoroquinolone treatment. therefore ciprofloxacin and levofloxacin are reported as resistant.

Negative (train) message

A) identifications performed by maldi tof mass spectrometry were developed and performance characteristics were determined by pcl alverno hammond in. no salmonella shigella aeromonas plesiomonas edwardsiella isolated. no predominant growth of klebsiella oxytoca present. one or more organisms were isolated and found to be normal flora through maldi tof mass spectrometry. campylobacter jejuni numerous. identified by maldi tof mass spectrometry. drugs of choice are ciprofloxicin erythromycin clindamycin tetracycline.
B) one or more organisms were isolated and found to be normal flora through definitive biochemical testing. no salmonella shigella plesiomonas edwardsiella or campylobacter isolated. no predominant growth of klebsiella

oxytoca present. aeromonas species moderate. susceptibility not routinely performed. aeromonas spp. are associated with gastrointestinal disease. symptoms are usually mild and self limiting. individuals with impaired immune systems or underlying malignancy are susceptible to more severe infection. antibiotics maybe indicated if symptoms are prolonged and in system icinfections. identified by maldi tof mass spectrometry.

Positive (synthetic) messages

A) client services present. moderate salmonella spp result progress called faxed to dr chang stroman at <time> on <date> by dr maritza office dr cahans office salmonella species sent to state lab salmonella group ser. jg salmonella spp
B) identifications performed by maldi tofmass spectrometry were developed and performance characteristics were determined by pcl alverno hammond in. salmonella species identified by maldi tof mass spectrometry. salmonella species sent to indiana indiana by felicia stipp [317 . e. coli o157 no campylobacter isolated sent to indiana state department of health reportable confirmed by indiana state dept. of further testing performed. salmonella performed performed on up please contact the laboratory if serotyping is required.

Negative (synthetic) messages

A) normal gi flora present no enteric pathogens isolated stool screened for salmonella shigella staphylococcusaureus campylobacter and sorbitol negative e. coli o157 this culture is a prior result no further 15d0662599 date called to difficile rn at dr. nichols office on <date>
B) test culture stool specimen type stool specimen received <date> <time> est final reports verified date/time <date> <time> final reports verified date/time <date> <time> no salmonella shigella species isolated no salmonella species no shigella species isolated no salmonella or shigella plesiomonas isolated. no shigella aeromonas plesiomonas edwardsiella or campylobacter isolated not routinely cultured is desired. no campylobacter specie scalled to indiana dept. of at <time>

Appendix C. List of top 20 features selected from the real and synthetic datasets using gini impurity scores.

Rank	Real (train) dataset	Synthetic dataset
1	Shigella	Salmonella
2	Salmonella	Speci
3	Speci	Shigella
4	Isol	Health
5	Campylobact	Campylobact
6	Health	Isol
7	Indiana	Sct
8	Group	State
9	Suscept	Group
10	Typhi	Confirm
11	Confirm	Chslb
12	Depart	Indiana
13	Ml	Suscept
14	Spp	Typhi
15	Call	Call
16	Cultur	Depart
17	Stool	Sent
18	Chslb	Test
19	Coli	Self
20	Enter	Cultur

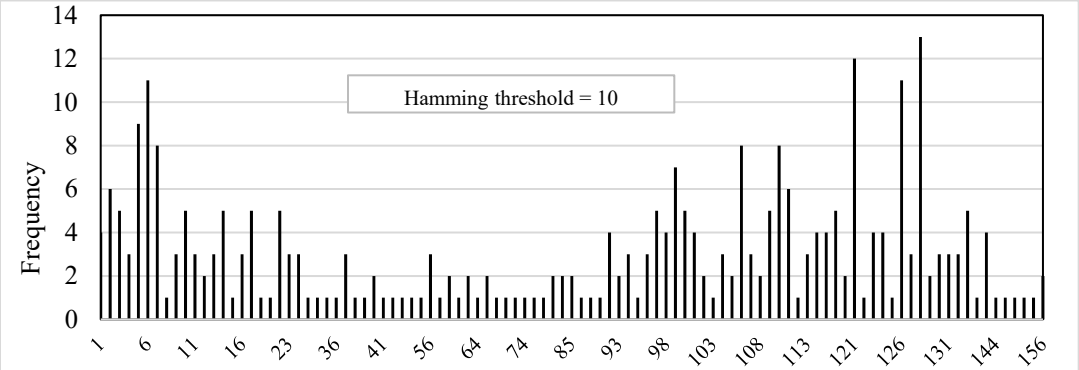
Appendix D. Intersection of top 5, 10, 15, 20, 50 and 100 features selected from the real and synthetic datasets using gini impurity scores.

Feature subset size	# features present in both datasets	List of features present in both datasets

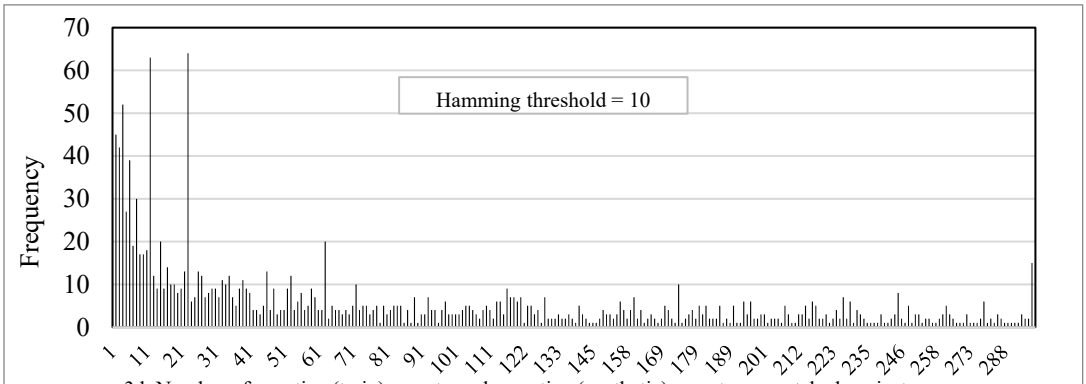
5	4 (80%)	salmonella, speci, shigella, campylobact
10	7 (70%)	speci, health, isol, group, salmonella, shigella, campylobact
15	12 (80%)	speci, indiana, campylobact, confirm, health, isol, group, suscept, typhi, salmonella, shigella, call
20	14 (70%)	chslb, speci, indiana, shigella, campylobact, confirm, health, isol, group, suscept, depart, salmonella, typhi, call
50	35 (70%)	sct, speci, non, due, isol, suscept, typhi, coli, report, chslb, call, indiana, cultur, confirm, diseas, issu, gener, azithromycin, sent, health, gastroenter, without, self, progress, depart, salmonella, stool, campylobact, enter, medic, test, final, group, shigella, tofmass
100	79 (79%)	perform, sct, present, speci, characterist, non, laboratori, due, isol, result, pathogen, infect, suscept, typhi, coli, identifi, report, chslb, serogroup, call, indiana, cultur, thi, lab, confirm, enzym, sourc, aerob, diseas, growth, issu, drug, gener, azithromycin, moder, maldi, specimen, sent, gastroenter, health, without, underli, serotyp, spectrometri, routin, self, toxin, numer, aeromona, follow, salmonella, progress, depart, stool, tofmass, ser, enter, campylobact, normal, develop, shiga, medic, patient, salsp, determin, test, mani, choic, mass, final, group, usual, see, board, date, shigella, access

Appendix E. Presence disclosure test

We computed frequency of a synthetic report matching with 1-to-n many real reports using a hamming score threshold of 10, which was selected based on its use in prior research(7). We *hypothesize*d that negative synthetic reports stood a much greater chance of matching with negative (train) reports because negative (train) reports tend to be similar to each other due to uniform text used to report a negative outcome. Thus, separate tests were performed against positive and negative datasets. We tabulated counts of how many positive reports were matched to each synthetic report. Next, we plotted the frequency of n synthetic reports matching with m training reports.



3.a Number of positive (train) reports each positive (synthetic) report was matched against



3.b Number of negative (train) reports each negative (synthetic) report was matched against

Figure 3. (3.a). Frequency of positive (synthetic) reports matched with positive (train) reports (hamming threshold <=10). (3.b). Frequency of negative (synthetic) reports matched with negative (train) reports irrespective of report status (hamming threshold <= 10).

We determined that re-identification risk is greater when,

a) A synthetic report is matched with a smaller number of real reports. Linking a synthetic report to a smaller number of real reports offer attackers a greater chance of pinpointing true matches via manual review. Re-identification risk falls as the number of real reports matched with a single synthetic report increases, as attackers must manually review each of these matches to pinpoint patients.
b) Synthetic reports are matched with real reports using smaller hamming cutoff thresholds. Smaller hamming distance thresholds indicate smaller differences between records, and thus, raises the likelihood that two matched reports are the same.

An evaluation of matches across a hamming distance of 10 presents that positive synthetic reports were matched to positive real reports (figures 3.a) at a lower rate than negative reports (3.b). As such, they poise significantly low chance of re-identification. We hypothesize that negative reports were matched with more certainty because they included boilerplate phrases reporting negative status. As anticipated, chances of matching a negative synthetic and real reports were larger than matching positive synthetic and real reports.

References

1. Callahan A, Shah NH. Machine Learning in Healthcare. Key Advances in Clinical Informatics: Elsevier; 2018. p. 279-91.
2. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. The American journal of gastroenterology. 2010;105(6):1224.
3. Hodge Jr JG, Gostin LO, Jacobson PD. Legal issues concerning electronic health information: privacy, quality, and liability. Jama. 1999;282(15):1466-71.
4. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC medical research methodology. 2010;10(1):70.
5. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017;24(3):596-606.
6. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. bmj. 2015;350:h1139.
7. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. arXiv preprint arXiv:170306490. 2017.
8. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PloS one. 2011;6(12):e28071.
9. McLachlan S, Dube K, Gallagher T, editors. Using the caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. Healthcare Informatics (ICHI), 2016 IEEE International Conference on; 2016: IEEE.
10. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al., editors. Generative adversarial nets. Advances in neural information processing systems; 2014.
11. Camino R, Hammerschmidt C, State R. Generating Multi-Categorical Samples with Generative Adversarial Networks. arXiv preprint arXiv:180701202. 2018.
12. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H, editors. Synthetic data augmentation using GAN for improved liver lesion classification. Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on; 2018: IEEE.
13. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434. 2015.
14. Beaulieu-Jones BK, Wu ZS, Williams C, Greene CS. Privacy-preserving generative deep neural networks support clinical data sharing. BioRxiv. 2017:159756.
15. Zhu Y, Lu S, Zheng L, Guo J, Zhang W, Wang J, et al. Texygen: A Benchmarking Platform for Text Generation Models. arXiv preprint arXiv:180201886. 2018.
16. Guimaraes GL, Sanchez-Lengeling B, Outeiral C, Farias PLC, Aspuru-Guzik A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. arXiv preprint arXiv:170510843. 2017.

17. Xu J, Sun X, Ren X, Lin J, Wei B, Li W. DP-GAN: Diversity-Promoting Generative Adversarial Network for Generating Informative and Diversified Text. arXiv preprint arXiv:180201345. 2018.
18. Guo J, Lu S, Cai H, Zhang W, Yu Y, Wang J. Long text generation via adversarial training with leaked information. arXiv preprint arXiv:170908624. 2017.
19. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. Jama. 2014;311(24):2479-80.
20. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems. 2014;2(1):3.
21. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, et al. The Indiana network for patient care: a working local health information infrastructure. 2005;24(5):1214-20.
22. Yu L, Zhang W, Wang J, Yu Y, editors. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. AAAI; 2017.
23. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. 2016;60:145-52.
24. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. Journal of biomedical informatics. 2017;69:160-76.
25. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BGJJobi. A simple algorithm for identifying negated findings and diseases in discharge summaries. 2001;34(5):301-10.
26. Breiman L, Friedman J, Olshen R. Stone, cj (1984) classification and regression trees. Wadsworth, Belmont, California. 2009.
27. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
28. Kasthurirathne SN, Vest JR, Menachemi N, Halverson PK, Grannis SJ. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. Journal of the American Medical Informatics Association. 2017;25(1):47-53.
29. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. Journal of biomedical informatics. 2016;60:145-52.
30. Nergiz ME, Clifton CJIToK, Engineering D. δ-presence without complete world knowledge. 2010;22(6):868-83.
31. Shokri R, Stronati M, Song C, Shmatikov V, editors. Membership inference attacks against machine learning models. Security and Privacy (SP), 2017 IEEE Symposium on; 2017: IEEE.
32. Dwork C, McSherry F, Nissim K, Smith A, editors. Calibrating noise to sensitivity in private data analysis. Theory of cryptography conference; 2006: Springer.
33. Truex S, Liu L, Gursoy ME, Yu L, Wei W. Towards demystifying membership inference attacks. arXiv preprint arXiv:180709173. 2018.
34. Rahman MA, Rahman T, Laganiere R, Mohammed N, Wang Y. Membership Inference Attack against Differentially Private Deep Learning Model. Transactions on Data Privacy. 2018;11(1):61-79.
35. Backes M, Berrang P, Humbert M, Manoharan P, editors. Membership privacy in MicroRNA-based studies. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016: ACM.
36. Hamming RWJBStj. Error detecting and error correcting codes. 1950;29(2):147-60.
37. Papineni K, Roukos S, Ward T, Zhu W-J, editors. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics; 2002: Association for Computational Linguistics.
38. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:160908144. 2016.
39. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X, editors. Improved techniques for training gans. Advances in Neural Information Processing Systems; 2016.
40. Chen L, Dai S, Tao C, Zhang H, Gan Z, Shen D, et al., editors. Adversarial Text Generation via Feature-Mover's Distance. Advances in Neural Information Processing Systems; 2018.
41. Matwin S, Nin J, Sehatkar M, Szapiro T. A review of attribute disclosure control. Advanced Research in Data Privacy: Springer; 2015. p. 41-61.
42. Che T, Li Y, Zhang R, Hjelm RD, Li W, Song Y, et al. Maximum-likelihood augmented discrete generative adversarial networks. arXiv preprint arXiv:170207983. 2017.

Quantification of BERT Diagnosis Generalizability Across Medical Specialties Using Semantic Dataset Distance

Mihir P. Khambete[1,2], William Su, MD[1,3], Juan C. Garcia, PhD[1], Marcus A. Badgeley, PhD[1,4]

[1] nference LLC, Cambridge, MA; [2] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA; [3] Department of Radiation Oncology, Penn Medicine, University of Pennsylvania Health System, Philadelphia, PA; [4] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA

Abstract

Deep learning models in healthcare may fail to generalize on data from unseen corpora. Additionally, no quantitative metric exists to tell how existing models will perform on new data. Previous studies demonstrated that NLP models of medical notes generalize variably between institutions, but ignored other levels of healthcare organization. We measured SciBERT diagnosis sentiment classifier generalizability between medical specialties using EHR sentences from MIMIC-III. Models trained on one specialty performed better on internal test sets than mixed or external test sets (mean AUCs 0.92, 0.87, and 0.83, respectively; p = 0.016). When models are trained on more specialties, they have better test performances (p < 1e-4). Model performance on new corpora is directly correlated to the similarity between train and test sentence content (p < 1e-4). Future studies should assess additional axes of generalization to ensure deep learning models fulfil their intended purpose across institutions, specialties, and practices.

Introduction

Natural Language Processing (NLP) models could enhance clinical research and patient care by automatically curating electronic health record (EHR) notes. The state-of-the-art NLP models are high capacity deep neural networks, and other deep neural networks (for image recognition) have been shown to be particularly prone to overfitting in healthcare contexts (Badgeley et al. 2019; Zech et al. 2018)[11,12]. Generalizability is the capacity of a model to perform comparably on test sets derived from domains not used for model training. NLP models that better generalize would enable models trained on existing labelled datasets to be deployed to new research questions and EHRs.

A major reason for the generalizability issue is that medical notes are largely entered as free text, and therefore lack the organization typical of structured data in the EHR. Different hospital systems may use different note styles, as can different medical specialties or caregivers within an institution. Thus, any generalizable NLP model must make accurate predictions on a wide range of note styles after training on medical notes from a small number of institutions.

Current approaches to work around the generalizability problem have included using domain adaptation and model retraining, which must be applied each time a model is used on data from a new hospital system (Wu et al, 2014)[10]. This precludes large-scale application of NLP for medical note curation.

Some prior studies have assessed how well NLP models generalize across institutions, but not at other levels of healthcare organization. Cancer diagnosis has been shown to be generalizable when using training data from multiple hospitals, at the expense of a slight loss in performance on test sets derived from hospitals used for model training (Santus et al. 2019)[7]. Similarly, temporal reasoning can also be considered generalizable when a diverse training set is used (Velupillai et al. 2015)[8]. Another group trained a model to predict the prognosis of ICU patients at an academic medical center, and found that results generalized better to another academic medical center than a local community hospital (Marafino et al. 2018)[5]. Marafino et al note that case mix and documentation practices may differ between academic and community hospitals, thus possibly explaining some of the difference in AUCs. On the other hand, existing models for negation detection have not generalized well across institutions (Wu et al. 2014)[10].

It remains unclear whether NLP models are generalizable on finer axes of generalization such as medical specialty and the type of EHR note. In this study, we used MIMIC-III, an ICU EHR dataset from Beth Israel Deaconess Medical Center (Johnson et al. 2016)[3]. We trained state-of-the-art NLP models (Bidirectional Encoders, BERT)[2] to classify

the sentiment of possible diagnoses across three medical specialties. Using training and test sets containing sentences from either one, two, or all three medical specialties, we answer questions about NLP generalizability at the medical specialty level. We investigated the effect of increasing the training set diversity, defined as the number of specialties represented in our study. Additionally, we examined how semantic similarity between a training set and a test set is correlated to model performance.

Methods

Dataset

Our study used both structured data and text from MIMIC-III, an EHR dataset from the Beth Israel Deaconess Medical Center (BIDMC) critical care units between 2001 and 2012. We primarily used the NOTEEVENTS table which contains over 2 million medical notes' text and metadata fields such as note type (e.g ECG, Nursing, Discharge Summary).

Our main objective was to assess whether an NLP model could generalize to sentences from medical specialties other than the one(s) which the model was trained on. We performed 2 types of analyses 1) unsupervised clustering of the EHR note embeddings and 2) supervised modeling of diagnosis sentiment for three specialties. We randomly sampled 5,020 "background" notes and fetched 2,955 sentences containing a medical diagnosis term. The unsupervised analysis used both background sampling and sentences containing diagnoses, whereas model training and testing only involved sentences containing diagnoses.

We identified sentences containing disease names from three medical specialties: oncology, cardiology, and pulmonology. These specialties were selected because of the high prevalence of disease in critical care patient populations. Disease tokens for these specialties were identified from the NIH TCGA project and the International Classification of Diseases-10 (ICD-10) (https://www.cancer.gov/tcga and World Health Organization, 2019 respectively). Disease names for cardiology and pulmonology were simplified to match shorthand commonly used by clinicians in EHR notes. For example, in our cardiology disease names, we selected the simplified term "heart failure" instead of specific variants such as "systolic (congestive) heart failure", "right heart failure", and "end stage heart failure". The most common disease names found for each specialty are included in Supplementary Table 2.

Language Processing

Each medical note was broken into sentences using the SpaCy English sentencizer. Sentences were then broken into words (excluding punctuation characters). Continuous stretches of 1-6 words were matched with the disease names from the medical specialty being searched for, to account for all disease names containing 5 or less words e.g. clear cell renal cell carcinoma or chronic obstructive pulmonary disease. If a match was found, the sentence, name of the disease, and the associated note metadata were collected. If a sentence contained multiple disease names from the same specialty, the sentence was evaluated for the first disease in the sentence. For example, the sentence "The patient has cardiomyopathy and heart failure" would be evaluated for cardiomyopathy. Sentences with more than 512 tokens were excluded, since our model architecture had a limit of 512 tokens per sentence. All sentences were unique.

Unsupervised Techniques

We used a deep bidirectional encoder model (BERT) to create language embeddings (Devlin et al. 2018)[2] The variant we used is called SciBERT, which was pre-trained on text from scientific publications (Beltagy, Lo, and Cohan 2019)[1] and is available in the python transformers package (Wolf et al. 2019)[9]. The transformer generates embeddings for individual words. To compute sentence or document embeddings we apply mean pooling as previously described and implemented in the sentence-transformers package (Reimers and Gurevych 2019)[6].

BERT embeddings of sentences and documents are 768-dimensional numerical representations of the texts' semantic content. We projected these embeddings into 2-dimensional spaces using t-distributed Stochastic Neighbor Embedding (tSNE, (Maaten and Hinton 2008)[4] and Principal Component Analysis (PCA) for unsupervised analyses. tSNE and PCA projections were generated and visualized using the scikit-learn and plotly packages. tSNE was used to identify

Table 1: Train and Test Set Compositions by Specialty

Medical Specialty Sets	Oncology		Cardiology		Pulmonology	
	Train	Test	Train	Test	Train	Test
Oncology Only	783	337	0	0	0	0
Cardiology Only	0	0	631	271	0	0
Pulmonology Only	0	0	0	0	653	280
Oncology and Cardiology	392	168	316	136	0	0
Oncology and Pulmonology	392	168	0	0	326	140
Cardiology and Pulmonology	0	0	316	136	326	140
All Three Specialties	261	112	210	90	217	93

local clusterings of notes; we used a perplexity value of 15 when creating tSNE projections to allow separation of points into separate clusters when visualized. PCA was used to retain distances between points and project queried sentences into an unbiased background distribution. The PCA rotation was fit on 1,004 randomly sampled sentences from MIMIC-III, and then the learned rotation was applied to the 2,955 specialty sentences.

Semantic similarity between sentences was measured by computing the cosine distance between pairs of sentences' 768-dimensional embeddings. The similarity of 2 data partitions was estimated by measuring the cosine distance between all pairs of sentences and computing the median: the Median Cosine Distance (MCD). Median rather than mean cosine distance was used to mitigate the effect of outliers.

Supervised Model Development

Sentence Annotation

Sentences with diagnoses were annotated by one of two authors: a resident physician (WS) or a physician scientist (MAB). Given a sentence and the disease contained in that sentence, the annotator marked the sentence as one of three ground truth labels: "Yes" (indicating the patient had the disease, or if a fetus had the disease in obstetric cases), "No" (if the patient did not have the disease), or "Maybe" (if there was insufficient evidence to exclude or confirm the disease). No inter-annotator agreement was used, since overlap between sentences annotated by both the resident physician and the physician-scientist was minimal.

An example "Yes" sentence is "pt was found to have **cardiomyopathy** and is in heart failure" , since the patient has a positive diagnosis for cardiomyopathy. An example "No" sentence is "No episodes of **tachycardia** at this time". An example "Maybe" sentence is "Some left **heart failure** cannot be excluded since the patient is supine"; there is insufficient evidence to mark this sentence as either positive or negative for heart failure. If a sentence contained information for a family member (which is routinely collected as part of the medical history), the sentence was marked as "No" even if the relative had the indication.

The initial round of labeling saw an overwhelming majority of sentences annotated as "Yes". To mitigate the class imbalance, we collected additional sentences from MIMIC-III that the preliminary BERT model predicted as "Maybe" or "No", and our annotators (WS, MAB) confirmed or corrected each of the model's predictions.

Train-Test Split

We created seven different combinations of sentences from the 3 specialties - and split each into a train and test set. For each medical specialty, approximately 70% of sentences were used for training (selected at random) while the rest were used for testing. Each train-test set contained sentences from one, two, or three medical specialties (see Table 1). For datasets containing sentences from multiple specialties, we used roughly the same number of sentences from each of the component specialties while keeping the total number of sentences consistent from train set to train set and from test set to test set. We trained sciBERT classification models on each of the 7 training datasets and then tested each model on all 7 test datasets, yielding 49 different train-test performance tests.

We define the following relations for train-test set pairs based on whether the same or different medical specialties are included: 1) native, 2) partial, and 3) external. "Native" performance tests include the same set of specialties in training and test sets. If there is no overlap between the specialties included in the training and test set, we designate the performance test as "external". All performance tests with any but not complete overlap are designated "partial".

Model Architecture

Our model included three modules. The first module was a cased sciBERT transformer which generated 768-dimensional embeddings from an input sentence. The embeddings were fed into a linear module with three output units. Lastly, a 3-class softmax unit generated confidences (probabilities) for each of the three output classes, such that the predicted class was that with the highest confidence.

Training used a categorical cross-entropy loss for each sentence using the ground truth labels and the model outputs from the softmax layer. Our model used a learning rate of 3e-5 in order to preserve the pre-training weights while fine-tuning the model. We chose hyperparameters of dropout rate = 0.9, and used the Adam optimizer along with a warmup schedule and LayerNorm. 10% of the training data was held out and used as a validation set.The model was trained until the validation set balanced accuracy worsened, up to 4 epochs.

Evaluating Model Performance

We generated receiver operator characteristic (ROC) and precision-recall (PR) curves for each train-test pair on each output class using the scikit-learn package. The primary performance metric was the area under the ROC curve (AUC) on the epoch with the lowest validation loss.

We compared how well 7 models performed on different tests to assess how model performance is affected by generalization across specialties and the diversity of training data. There are 3 train-test set relations for generalizability across specialties and up to 3 specialties included in training, so we use the ANOVA repeated measures (extension of a paired t-test) to see if the explanatory variable has a significant impact on a model's relative performance across tests. Repeated-measures ANOVA test was implemented with the python pingouin package.

Finally, we analyzed the relationship between AUC and MCD between a train-test set to determine whether similarity between a train-test set is correlated to how well models generalize. Significant associations between model test performance and MCD were evaluated using a two-tailed t-test implemented with the python statsmodels package.

Results

Data Characteristics

We queried medical notes from the MIMIC dataset to obtain 1) a random subset of 5,020 documents for unsupervised note analysis and 2) 2,955 sentences containing medical diagnosis terms from one of 3 medical specialties to train diagnosis models. The distribution of note types was unbalanced in both cohorts, and the notes containing oncology diagnoses were notably enriched in discharge summaries (Supplementary Table 1). The medical diagnoses that were found most frequently are shown in Table S4.

Documents and sentences were split into subword tokens. Sentences with pulmonology, cardiology, and oncology diagnoses had an average of 43, 48, and 65 tokens per sentence (see full distributions in Figure S10).

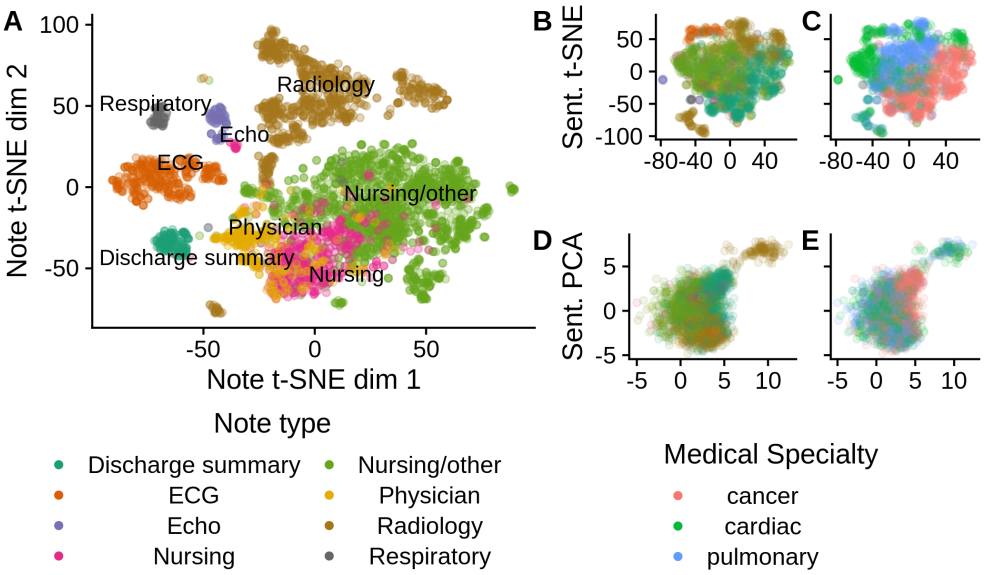


Figure 1: Unsupervised analysis of medical notes. (A) Document-level tSNE for 5,020 randomly sampled medical notes. Sentence-level (B, C) tSNE and (D, E) PCA for 2,955 sentences containing diagnoses. Each point represents a document or sentence that was embedded using a sciBERT model and projected into 2-dimensions with tSNE (A, B, C) and PCA (D, E). Colors indicate the (A, B, D) type of note (e.g. nursing, physician, pharmacy) and (C, E) diagnosis specialty (cancer, cardiac, pulmonary).

Documents were vectorized using the sciBERT model which was pre-trained on scientific corpora. The tSNE projection of SciBERT embeddings of randomly sampled notes are shown in Figure 1A. Each point, whose xy-coordinates represent the 2-dimensional projection of a document-vector embedding, was subsequently colored according to note type in the document metadata. Document-vector embeddings cluster into neighborhoods primarily based on the type of note. General note types were intermixed and overlapping (e.g. nursing and physician notes). Specialty notes are packed into single isolated clusters (e.g. ultrasound and respiratory notes) or distributed into several subpopulations (e.g. radiology and ECG). For example, the radiology cluster centered roughly around (-50, -75) in the tSNE projection corresponds to notes for patients undergoing catheterization for angiography and other vascular procedures, while the nursing cluster centered around (10, -70) contains notes related to the care of newborns in the ICU.

Sentences containing diagnoses were similarly embedded and visualized with SciBERT. Like the document tSNE, the sentence clusters coincide with the type of note they came from (Figure 1B); however, the sentence-level clustering is much less pronounced than the document-level clustering. The specialty coloring (Figure 1C) reveals the disproportionate note types for each specialty. Manual review of sentence clusters reveals content-based neighborhoods. For example, cardiology sentences projected near (-50, 0) discuss arrhythmias such as tachycardia and ventricular fibrillation.

In addition to tSNE, we visualized our sentence embeddings using PCA, which preserves longer distance relations between high-dimensional data points (Figures 1C, D). PCA was fitted using an unbiased distribution of approximately 1,000 sentences sampled from across MIMIC-III. There is a primary group of sentences with all 3 specialty diagnoses as well as a secondary group of sentences from radiology notes that only contains cardiac and pulmonary sentences. As in tSNE, we see that sentences are more separated based on note type than the diagnosis token medical specialty. The bivariate principal component distributions of cardiac and pulmonary specialty sentences appear more similar than the oncology sentence distribution.

Model Performance

Model performance was assessed by measuring AUC for each train-test pair and label (Supplementary Table 3), as well as by constructing precision-recall curves (Supplementary Figure 2).

We first investigated whether models could generalize across medical specialties. We summarize test set performance for models trained on a single specialty by macro-averaging AUC across labels and averaging across train-test relationships (native, partial overlap, and external; see methods). We find that AUC monotonically decreases as the overlap between specialties decreases (Figure 2A; repeated measures ANOVA p = 0.0163).

Unsupervised Analysis of Clinical Text Embeddings

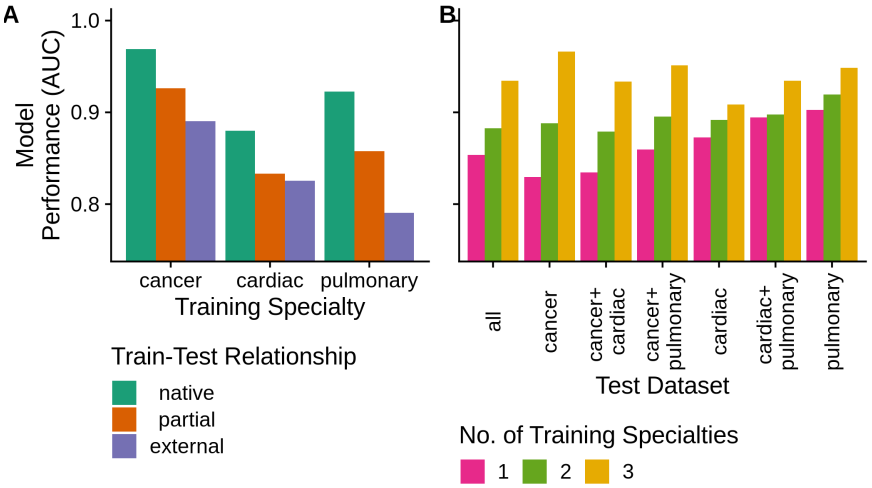


Figure 2: AUC bar plots of model performance tests with different groupings to test primary hypotheses. (A) AUC vs train-test set relation for each single-disease model, averaged over labels and test sets. (B) AUC vs how many specialties were used to train the model. Bars are colored by the number of medical specialty diagnoses used in model training and grouped by test set name, averaged over labels and training specialties.

To test whether the diversity of training data improves test performance, we grouped models based on how many specialties they were trained and summarized performance on each test set. Specifically, we take a macro-average across labels and average the results for models trained on diagnoses from 1, 2, or 3 specialties (Figure 2B). We found that for every single test set, the average model performance monotonically increased as more training data specialties are used to train the models. Using the repeated-measures ANOVA test, we found that the differences in macro-average AUC were statistically significant ($p < 1e\text{-}4$).

Secondary Performance Evaluation

We used MCD to measure the semantic similarity between different train-test sets and investigated how this affects model performance. We measured MCD both between and within every combination of train and test datasets (see supplementary tables 2 and 3). We group dataset pairs using the following relations: intra-dataset, internal, partial, and external (the latter three relations are between training set-test set pairs only). We found significant differences between the mean MCDs of native, partial, and external train-test pairs (mean MCDs: 0.230, 0.237, and 0.245, respectively; one-way ANOVA p = 5.78e-5). However, we did not find significant differences between intra-set and native train-test MCDs (mean MCD 0.230 vs 0.231; one-sample t-test p = 0.819). We validated that MCD is not different for subsamples from the same distribution, and that it progressively increases for train-test sets from 1) same distribution, 2) distribution mixtures, and 3) separate distributions.

Model test AUCs are shown as a function of the MCD for every combination of train-test sets in Figure 3. Using

ordinary least-squares (OLS) regression, we find an inverse relationship between MCD and AUC for unequivocal diagnoses (p=6.5e-6 and 4.5e-5 for sentences labelled "Yes" or "No" respectively) but no difference when a diagnosis is uncertain (p = 0.70 for "Maybe") (Table 2). The median cosine distances and AUCs for each train-test set are provided in Supplementary Table 3. Baseline median cosine distances between each dataset and itself are given in Supplementary Table 2.

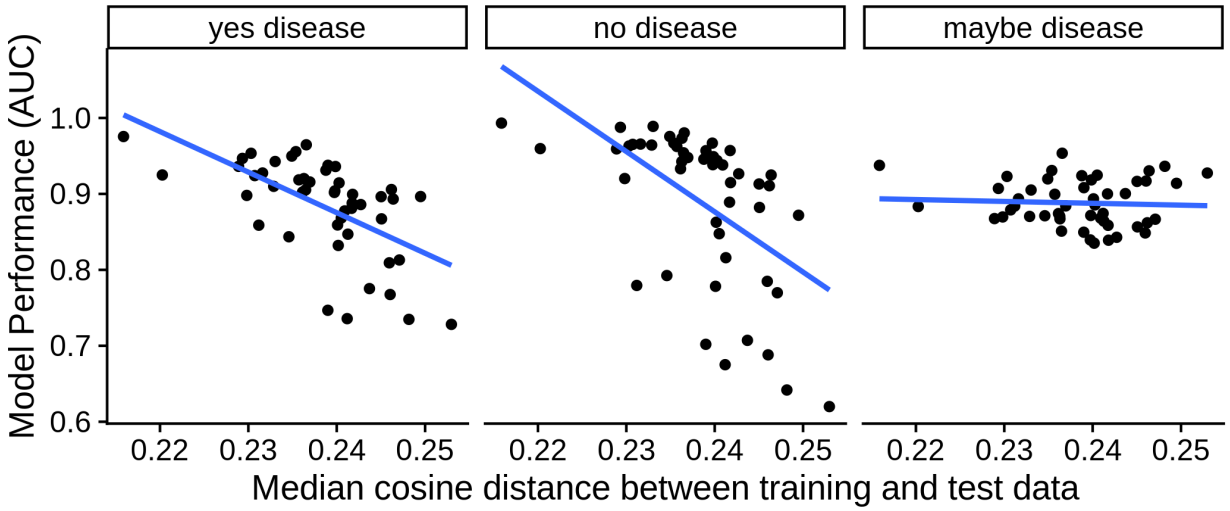


Figure 3: AUC vs. semantic similarity between a model's training and test datasets. Semantic similarity is measured with median cosine distance. Each point represents one performance test (train-test pair) for each label. Trend lines are fitted using OLS.

Table 2: Ordinary Least-squares regression statistics

Diagnosis Label	Slope	R^2	p-value
"Yes"	-5.3	0.354	0.000006
"No"	-7.9	0.301	0.000045
"Maybe"	-0.2	0.003	0.700000

Alternative Choice of Performance Measure

We also investigated the use of a precision (also known as positive predictive value or PPV) as a performance measure. For each train-test set pair, we measured PPV at a recall cutoff of 0.9. We repeated the analyses in the Model Performance and Secondary Performance Evaluation sections using PPV rather than AUC (see Supplementary Figures 6-7 and Supplementary Table 7).

Discussion

Comparison with Previous Studies

Generalizability continues to be one of the most important concerns in biomedical natural language processing, since a resolution of the generalizability problem would allow for widespread deployment of NLP models to both research and clinical practice. Previous studies have qualitatively highlighted the difficulty of making NLP models generalize across institutions in applications such as negation detection, cancer diagnosis, and temporal reasoning (Wu et al, Santus et al, Velupillai et al respectively)[7,8,10]. Here we provide systematic and quantitative testing that reveals that even within a single medical unit there are generalizability differences across medical specialties and labels.

Wu et. al examined whether a rule-based negation detection model could generalize to clinical text corpora on which it

was not trained[10]. The authors established that while a model could be optimized to increase F1-score on an individual corpus of text, that model could not then be deployed on another corpus with comparable performance without first undergoing domain adaptation.

Santus et al investigated the generalizability of a breast cancer diagnosis model across institutions. Their model used a convolutional neural network (CNN) coupled with a logistic regression unit to predict the presence of breast cancer based on a medical note (represented as a matrix of sentence embeddings)[7]. Santus et. al. combined data from 2 or 3 specific institutions to "increase diversity" and showed improved accuracy on test data from an unseen institution.

Our study improved upon the work from Santus et al[7] in several ways. We used a state of the art NLP classification architecture (sciBERT), designed training and test cohorts in a combinatorial fashion, and used more refined performance metrics and distance measures between datasets. A BERT-based architecture allowed our model to consider the relationship between any two tokens in an input sentence as opposed to older language models. We trained models on all combinations of 2 (or all 3) medical specialties to better estimate model performance trends and computed group-comparison statistics. We used AUC as a performance measure rather than accuracy since AUC can be computed separately for each class and measures. We use MCDs (which are mentioned descriptively in Santus et al[1]) to explain and statistically test differences in model generalization across this continuous measure of train-test dataset difference.

We discovered that generalization is even an issue within a single critical care unit when models are tested on new medical specialties, how well a model will generalize is related to training and testing conditions as well as train-test data-distance, and there are finer differences between specific labels' generalizability.

Unsupervised Result Interpretation

We used tSNE and PCA to examine whether the sciBERT embeddings would cluster according to some property of the corresponding sentences/notes.

tSNE demonstrated that both note and sentence embeddings clustered based on the type of EHR note (Figure 1A, B), but is markedly more pronounced at the document level. This could be explained by the smaller size of individual sentences compared to entire notes, and is likely reinforced by different medical specialties using different sets of EHR note templates.

We found that sentences/notes clustered in the tSNE projection by note type rather than medical specialty. This is consistent with the notion that different note types require different vocabulary and may be written by different types of caregivers (e.g respiratory therapists for respiratory notes, nurses for nursing notes).

PCA confirmed the notion that note type is the primary signal in clustering sentence embeddings (Figure 1D, 1E). In addition, cardiology and pulmonology sentences are more similarly distributed in the PCA projection than either is to oncology sentences. This is not unexpected since cardiac and pulmonary conditions share underlying physiology and often present together in ICU patients.

Supervised Result Interpretation

Using the repeated-measures ANOVA test on the single-specialty models, we were able to show that the differences in macro-averaged AUC between internal, partial overlap, and external test sets were statistically significant as shown in Figure 2A ($p = 0.016$). Model performance was thus found to decrease when the test set had a smaller percentage of sentences from the same specialty as those in the training set. This is a confirmation of the generalizability issue at the medical specialty level, since models trained on a single specialty could not achieve comparable performance on test sets containing sentences from unseen specialties. We note that the relatively high performance of oncology-trained models in Figure 2A may be due to an overrepresentation of oncology articles in the corpora used to pretrain sciBERT compared to articles on other specialties.

However, we have also found that increasing the number of specialties represented in the training set upfront can improve model generalizability on unseen test specialties. Thus, our result corroborates the central finding from Santus et al[7] and is also supported by a statistical significance test. Unlike Santus[7], we created specialties for train and

test sets in a combinatorial fashion, providing a more comprehensive view of the effect of training set diversity. Our finding suggests that a general purpose diagnosis model might have to be fine-tuned using a heterogenous training set to allow the model to generalize on a wider range of test sets.

We showed that train set - test set similarity is positively correlated with increased classification performance for sentences with positive and negative sentiment ("Yes" and "No" ground truth labels). In general, greater train-test distance resulted in lower classification AUC score for "Yes" and "No" labels but not for "Maybe" labels (Figure 3), a surprising result. Thus, it is easier to generalize on "Maybe" sentences from a dissimilar test set since there is relatively small loss of AUC compared to "Yes" and "No" sentences. This suggests that medical informatics models may be more brazenly deployed to new specialties for screening or population health applications than diagnostic applications.

We believe that cosine distance between the unsupervised embeddings of datasets provides invaluable information for machine learning practitioners during model deployment and development. In model deployment, a practitioner can assess whether a model can generalize to a new dataset without requiring any labelled data. Similarly, during iterative model development, the practitioner can measure the distance between a new, unlabeled dataset and an existing training set to ensure maximal gain in data diversity at every iterative step of model development. In instances where it is expensive to acquire labelled data, such as in healthcare settings, this approach will greatly reduce the time and resources required to develop robust NLP models.

Limitations

Our analysis is limited by our sentence collection methodology, performance measures, and use of MCD as a proxy for dataset similarity.

A limitation of the dataset is that MIMIC-III is only contains ICU records. A dataset containing sentences from all departments of the hospital would provide a larger and more diverse pool of sentences.

The AUC score weighs a binary classifier's sensitivity and specificity across all probability cutoffs, but medical diagnoses and treatment decisions require a single cutoff. Cutoffs can be selected to provide high sensitivity in screening tests and high specificity in diagnostics. In a clinical setting, one could define a positive predictive value (PPV, also known as precision) cutoff based on the particular classification task at hand (as we did in the supplementary results section) or use the average precision derived from precision-recall curves (see Supplementary Figures 3-4).

We found expected baseline MCD trends between datasets and showed that this proxy of dataset similarity was significantly associated with how well a model will generalize to an arbitrary test dataset. But this single statistic is a narrow interpretation of dataset similarity; a collection of metrics may afford a richer picture of dataset similarity.

Future Directions

Follow-up studies could include assessing generalizability of specialty sentences along other metadata axes from the MIMIC-III database, such as note type or patient demographics. Additionally, one could ask other questions about the sentences collected for this study, such as patient prognosis given a sentence and the name of a disease contained in that sentence. We only used sciBERT for further fine-tuning since it was the only model supported in our company's cloud model tuning toolkit. One could also examine the effect of using different pretrained BERT models, such as BioBERT or an EHR-pretrained BERT on AUCs and generalizability. We tried unsupervised analyses with pretrained bioBERT and saw similar tSNE clustering as with sciBERT embeddings. Based on the unsupervised results across embeddings, we believe that the supervised phenomenon we've reported on SciBERT would similarly apply to embeddings with different tunings.

Conclusion

Healthcare delivery currently requires extensive manual review of patients' medical record notes, and medical informatics has models that could help but they may not work as well when deployed to new institutions. This study found that models also may not perform as well on corpora from the same institution that involve different medical special-

ties. We identified the median cosine distance as an indicator of how well a model will perform on new unlabelled data so that medical informatics can be deployed to support healthcare providers and researchers with reasonably similar corpora.

Acknowledgements

We'd like to acknowledge support for this work from Nference. The healthcare AI company has provided fantastic apps to accelerate this research on making biopharmaceutical and clinical data more computable.

We'd like to thank Joseph Lehar for reviewing the manuscript and providing feedback.

Supplemental Tables and Figures

Supplemental figures and tables can be found in the pre-print: https://arxiv.org/pdf/2008.06606.pdf.

Authors' Disclosures of Conflicts of Interest

Marcus Badgeley: Employee of Neurable at the time of writing.

References

1. Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. "SciBERT: A pretrained language model for scientific text." arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1903.10676.
2. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: pre-training of deep bidirectional transformers for language understanding." arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1810.04805.
3. Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. "MIMIC-III, a freely accessible critical care database." Scientific Data 3 (May): 160035.
4. Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing data using T-SNE." Journal of Machine Learning Research: JMLR 9 (Nov): 2579–2605.
5. Marafino, Ben J., Miran Park, Jason M. Davies, Robert Thombley, Harold S. Luft, David C. Sing, Dhruv S. Kazi, et al. 2018. "Validation of prediction models for critical care outcomes using natural language processing of electronic health record data." JAMA Network Open 1 (8): e185097.
6. Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: sentence embeddings using siamese BERT-Networks." arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1908.10084.
7. Santus, Enrico, Clara Li, Adam Yala, Donald Peck, Rufina Soomro, Naveen Faridi, Isra Mamshad, et al. 2019. "Do neural information extraction algorithms generalize across institutions?" JCO Clinical Cancer Informatics 3 (July): 1–8.
8. Velupillai, Sumithra, Danielle L. Mowery, Samir Abdelrahman, Lee Christensen, and Wendy W. Chapman. 2015. "Towards a generalizable time expression model for temporal reasoning in clinical notes." AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium 2015 (November): 1252–59.
9. Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2019. "HuggingFace's transformers: state-of-the-art natural language processing." arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1910.03771.
10. Wu, Stephen, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. "Negation's not solved: generalizability Versus optimizability in clinical natural language processing." PLoS ONE. https://doi.org/10.1371/journal.pone.0112774.
11. Badgeley, Marcus A., John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. 2019. "Deep learning predicts hip fracture using confounding patient and healthcare variables." NPJ Digital Medicine 2 (April): 31.
12. Zech, John R., Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study." PLoS Medicine 15 (11): e1002683.

VERTIcal Grid lOgistic regression with Confidence Intervals (VERTIGO-CI)

Jihoon Kim, MS[1,*], Wentao Li, MS[2,*], Tyler Bath, BS[1], Xiaoqian Jiang, PhD[2], Lucila Ohno-Machado, MD, MBA, PhD[1]
[1]University of California San Diego Health System Department of Biomedical Informatics, La Jolla, CA 92130, USA.
[2]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[*]These authors contributed equally.
Corresponding Author: Lucila Ohno-Machado, MD, MBA, PhD (lohnomachado@health.ucsd.edu)

Abstract

Federated learning of data from multiple participating parties is getting more attention and has many healthcare applications. We have previously developed VERTIGO, a distributed logistic regression model for vertically partitioned data. The model takes advantage of the linear separation property of kernel matrices of a dual space model to harmonize information in a privacy-preserving manner. However, this method does not handle the variance estimation and only provides point estimates: it cannot report test statistics and associated P-values. In this work, we extend VERTIGO by introducing a novel ring-structure protocol to pass on intermediary statistics among clients and successfully reconstructed the covariance matrix in the dual space. This extension, VERTIGO-CI, is a complete protocol to construct a logistic regression model from vertically partitioned datasets as if it is trained on combined data in a centralized setting. We evaluated our results on synthetic and real data, showing the equivalent accuracy and tolerable performance overhead compared to the centralized version. This novel extension can be applied to other types of generalized linear models that have dual objectives.

Introduction

With the of adoption of electronic health records (EHRs) in the US and advances in health information technology (HIT), a vast amount of health data is being generated rapidly. These data come from different sources (e.g., hospitals, cohort studies, disease registries, health insurance providers, and DNA/RNA sequencers). The conventional solution is first to gather datasets from multiple sources at a central site and then conduct analyses to answer a clinical/research question. However, such a centralized approach is not always viable because of potential harm to patient privacy, regulations, and policies, mistrust among participants, etc. If analyses could be conducted with data that are maintained in different places, this would greatly mitigate these factors.

A dataset can be partitioned in two ways: horizontally or vertically. Datasets are horizontally partitioned if all participating sites have the same set of features from different individuals. For example, a risk score model for coronary heart disease collects demographic, cholesterol, blood pressure, diabetes and smoking status from different institutions to develop or validate the model. Horizontally partitioned datasets[1] occur in multi-site clinical trials, clinical data research networks (CDRNs), registries, and risk prediction models with non-overlapping development and validation sites[2]. On the other hand, a dataset can be partitioned vertically in two or more different features from the same individual and the subset of features of it can be stored in different sites. For example, a Strong Heart Study, the largest epidemiology study of cardiovascular disease in American Indians, store the the genotype data in one institution and the phenotype data in another institution, allowing access only to approved researchers[3]. Booming direct-to-consumer (DTC) genetic testing[4] companies keep the individual's genetic data in their storage server, but the clinical information of the patients are stowed in the patient registry or EHR system. However, the association test of these genetic data can be performed only when they are linked with phenotypes, typically EHRs, thus physically separated from the genotype data. While healthcare claims data are saved in health insurance companies, detailed patient data are located in hospital EHR system. Current protocol of data access involves a lengthy process of request, review, approval, and monitoring limiting the opportunity of clinical research. Even when datasets can be centralized, transferring these to a central site is not trivial with genomic, imaging, and patient-generated health data from mobile phones and devices because these datasets can be very large in size. For this reason, commercial cloud computing platforms provide com-

monly used public genomics datasets such as 1000 genome[5] or The Cancer Genomics Atlas (TCGA)[6] datasets so that users can bypass the redundant transfer of such large datasets, which are costly and choke up the network.

Many algorithms were developed for federated analytics for both horizontal and vertically partitioned datasets[1,7,8]. For vertically partitioned datasets, secure matrix product algorithms are widely adopted[9–11]. None of these methods used dual optimization to perform interval estimation for vertically partitioned datasets. Dual optimization has been used for support vector machine classifier[12], but the logistic regression model is preferred method in genetics. VERTIcal Grid lOgistic regression (VERTIGO) is a distributed algorithm to build a logistic regression model on vertically partitioned datasets using dual optimization[13]. However, VERTIGO provides the only point-estimates, so no confidence interval is provided, and the statistical significance of the estimate in the form of a P-value is not provided. This study is an extension of our previous work, namely VERTIGO, to add standard errors to derive the interval estimates and express the parameter's statistical significance. This paper introduces a novel way of generating and transmitting confidence intervals along with coefficients. We describe our proposed algorithm, provide the mathematical proof, and demonstrate the algorithm performance on both simulated and real datasets.

Methods

Synthetic data generation

Synthetic data for 2000 samples and 20 features were created with some distributional assumptions, as follows:

1. Generate two independent matrices, X_1 and X_2, of the dimension 2000 examples $\times$ 20 features, using a Uniform$[0, 1]$ distribution
2. Derive a linear combination, $X = 1 + 2X_1 + 3X_2$, of the above two matrices
3. Generate random ground truth parameter vector β with size (20×1) using a Uniform$[0, 1]$ distribution
4. Apply the sigmoid function to calculate the probabilities for a binary outcome, $p = 1/(1 + e^{-X\beta})$
5. Generate the binary outcome y with probability p in step 4 using a Bernoulli distribution

Then the generated samples were assigned to mutually exclusive partitions, where the number of partitions, k, was varied from 2 to 4 and each partition represented a client site.

Real data BURN1000

A synthetic data about a burn study was obtained from R package aplore3. It is included in a companion data archive for the textbook by Hosmer and Lemeshow[14]. The burn data had eight variables and 1000 samples. The outcome was death, a binary variable of alive or dead. The seven features were age, gender, race, burn facility, total burn surface area, burn involved in inhalation injury, and flame involved in a burn injury.

PennCath

A real data was from the Foulkes lab (`http://www.stat-gen.org/`), and this is the PennCATH cohort data, which arises from a Genome-wide association (GWA) study of coronary artery disease (CAD) and cardiovascular risk factors based at the University of Pennsylvania Medical Center[15]. First of all, the quality control process is performed on the genotype data to check sex discrepancy, minor allele frequency, Hardy-Weinberg equilibrium, and relatedness. In the end, the sample size of the data shrinks from 3850 to 1280. Then the whole dataset was split into two clients, phenotype and genotype. The binary outcome is the disease condition, yes or no. The phenotype data includes age, sex, and additional covariates for each individual, while the genotype data contains 10 principal components for SNPs. Those 10 components along with phenotypes data and one genotype data in 1,000 SNPs, will be put into the VERTIGO-CI algorithm. To evaluate the computation time, we designed studies for 3 batches of trials using 10, 100, and 1000 SNPs.

Model

The logistic model is defined as

$$P(y = \pm 1|X, \beta) = \frac{1}{1 + \exp(-y\beta^\top X)}$$

where y is a binary outcome, X is the design matrix of sample-by-feature, and β is the model parameter. The goal is to find the estimate for β given observed data X and y. The best estimate for β is the maximizer of the log-likelihood function

$$\underset{\beta}{\text{argmax}}\, l(\beta) = \underset{\beta}{\text{argmax}}\ \log \pi(yX\beta) - \frac{\lambda}{2}\beta^T\beta$$

where π is the sigmoid function and $\frac{\lambda}{2}\beta^T\beta$ is the regularization penalty term to avoid overfitting. Since the above equation cannot be used for a vertically partitioned dataset in its current form, VERTIGO algorithm adopts reparameterization using the dual form of the original optimization equation

$$\underset{\alpha}{\text{argmin}}\, J(\alpha) = \underset{\alpha}{\text{argmin}}\, \frac{1}{2\lambda}\|y\alpha X\|_2^2 - L(\alpha), \quad L(\alpha) = -\alpha^T \log(\alpha) - (1-\alpha)^T \log(1-\alpha)$$

This dual form of the maximum likelihood function is generating the same results by optimizing dual parameters with respect to samples rather than features, keeping the information intact[16]. The next step is to update the parameters α using Newton's method[17] by iterating

$$\alpha^{(s+1)} = \alpha^{(s)} - \frac{J'(\alpha^s)}{H(\alpha^s)}$$

where $J'(\alpha)$ and $H(\alpha)$ are, respectively, the first and second derivative of dual object function $J(\alpha)$, defined as

$$J'(\alpha) = \lambda^{-1} y\alpha^T yXX^T + \log\frac{\alpha}{1-\alpha}$$

$$H = \lambda^{-1}\text{diag}(y)XX^T\text{diag}(y) + CI$$

Note that H, the Hessian matrix, has been changed in this situation for calculation convenience, and such changes will not harm the convergence as it only changes the step size[18]. C is a positive constant that enables the Hessian matrix to be full rank so its inverse matrix exists. When dual parameters α converges, the desired primal form parameter vector β can be obtained by its relationship to α,

$$\beta = \lambda^{-1}\alpha y^T X$$

This study's novel contribution is producing the standard errors of the point estimates that can be used to report statistical significance by P-Values or confidence intervals. The standard error of the coefficient can be represented as

$$(X'VX)^{-1/2}, \quad V = \text{diag}\left(\frac{e^{X\hat{\beta}}}{1+e^{X\hat{\beta}}}\left(1 - \frac{e^{X\hat{\beta}}}{1+e^{X\hat{\beta}}}\right)\right) \tag{1}$$

with the setting of vertically partitioned assumption on X, we have $X = (X_1, X_2, \cdots, X_k)$.

Since V is not separable for its own, the intermediate-term $e^{X_i\hat{\beta}_i}$ can be used to calculate V, by sending each term to clients, so that the final matrix V can be computed. Additionally, V should not be known by the center server because the information of X can be reverse-engineered with using previously seen data. So, at this step, the matrix V must be kept secret from the server.

The first connected client to the closed network acts as a lead-client and collects the first intermediate matrix, $e^{X_i\hat{\beta}_i}$, from the other clients. This lead-client generates V and sends it back to all clients. Finally, each client sends the second intermediate matrix, $X_iV^{1/2}$, back to the server. Since the matrix V is hidden to the server, the individual-level data are protected. Since $X'VX$ is separable as follows

$$X'VX = \begin{pmatrix} X_1'VX_1 & X_1'VX_2 & \cdots & X_1'VX_k \\ X_2'VX_1 & X_2'VX_2 & \cdots & X_2'VX_k \\ \vdots & \vdots & \ddots & \vdots \\ X_k'VX_1 & X_k'VX_2 & \cdots & X_k'VX_k \end{pmatrix} \tag{2}$$

where k is the number of clients, directly interpretable statistics such as the Z score can be calculated as $Z = \beta/\text{diag}((X'VX)^{-1/2})$, and confidence intervals and P-values can be derived. The pseudo-code is presented in **Algorithm 1**. Since $X_i'VX_j$ has a different size, the problem turns into a 'puzzle solving' to update the partial block matrices. Thus, putting those matrices in the right places is important. See the matrices-puzzle-solving pseudo-code in **Algorithm 2**. As an example, when $k = 3$, the algorithm will be executed as shown in **Figure 1**. 'Row_Block i' is defined as $[X_iVX_1, X_iVX_2, \ldots, X_iVX_k]$ binding k matrices column-wise where k is the number of clients.

Algorithm 1 VERTIGO-CI

Input: Data matrix of each client X_i (n samples by p_i features), shared outcome Y, and penalty parameter λ ($i = 1, \cdots, k$)
Output: Coefficient β^*, their standard errors and confidence intervals
Procedure:

1. Each client i: sends gram matrix $K_i = X_iX_i^T$ to the server
2. Server: combines the global gram matrices to have $K = \sum_i K_i$'s , initializes dual parameters $\alpha^{(0)} = \mathbf{0}$ and broadcasts these parameters back to the clients
3. Initialize step $s = 0$
4. Repeat while changes in $\alpha <$ predetermined threshold:
   (a) The client i: Computes $E_i^{(s)} = \lambda^{-1} y{\alpha^{(s)}}^T yK_i$ and send the intermediate matrix to the server
   (b) Server: Combines and calculates $E^{(s)} = \sum_i E_i^{(s)}$, $J'(\alpha^{(s)}) = E^{(s)} + \log \dfrac{\alpha^{(s)}}{1-\alpha^{(s)}}$
   (c) Server: Computes Hessian matrix $H^{(s)}(\alpha^{(s)}) = \lambda^{-1}\text{diag}(y)K\text{diag}(y) + CI$ and calculates the inverse ${H^{(s)}}^{-1}$
   (d) Server: Updates the dual parameters using Newton's method $\alpha^{(s+1)} = \alpha^{(s)} - J'(\alpha^{(s)}){H^{(s)}}^{-1}$, then sends the updated $\alpha^{(s+1)}$ back to clients
   (e) $s = s + 1$
5. Set the final alpha as α^*, the optimal value of α
6. Each client i: Calculates the global optimization $\beta_i^* = \lambda^{-1}\alpha^* y^T X_i$ and sends it to the server
7. Server: Combines the global optimum estimates from each client $\beta^* = ({\beta_1^*}^T, \cdots, {\beta_k^*}^T)^T$,
8. Client-to-Client communication:
   (a) The client i: Calculates $e^{X_i\beta_1^*}$ and sends it to client 1
   (b) Client 1: Combines the statistics $e^{X\beta} = \prod_i e^{X_i\beta_i^*}$ and calculates V as above (1), then sends the V back to clients $2, 3, \ldots, k$
9. Each client i: Calculates $X_iV^{1/2}$ and sends to the server
10. Server: Combines and calculates the standard errors, p-values, and confidence intervals

Implementation

We implemented the VERTIGO-CI in Python 3.7, using the *numpy*, *pandas*, and *scipy* modules to perform the mathematical computations. We utilized the asyncio module for network programming to allow asynchronous operations. All testing was performed on Amazon Web Service (AWS) EC2 instance of r5a.2xlarge (64 GB Memory, 8 CPUs) with Ubuntu 18.04 instances in different data centers in five continents (Asia: Seoul, Australia: Sydney, Europe: Dublin, North America: Oregon/Virginia, and South America: Sao Paulo).

Algorithm 2 Matrices-puzzle-solving

Input: Intermediate matrix of each client $X_i'V^{1/2}$, number of clients k
Output: The completed intermediate matrix $X'VX$ for calculation of Standard Deviation.
for $i = 1 : k$ **do**
  for $j = 1 : k$ **do**
    $RowBlock[i] = [RowBlock[i],\ X_i'V^{1/2} \cdot (X_j'V^{1/2})^T]$
  end for
  $X'VX = [(X'VX)^T,\ RowBlock[i]^T]^T$
end for

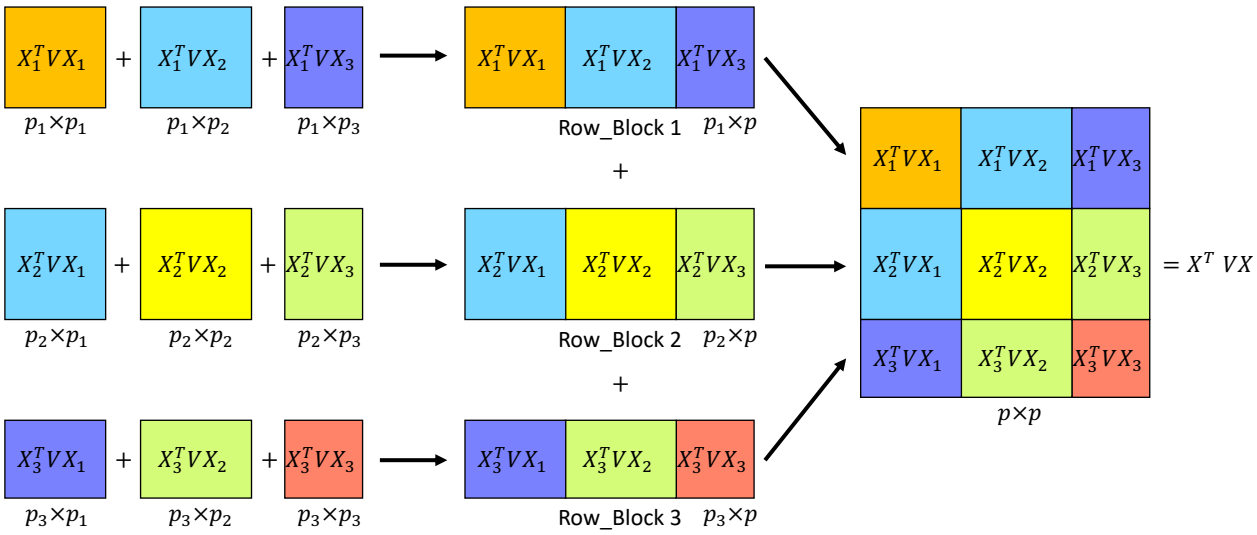


Figure 1: Example for 3 clients VERTIGO-CI matrices puzzle combination. Here the dimensions of X_1, X_2, X_3, V are $n \times p_1, n \times p_2, n \times p_3$, and $n \times n$ where n is the number of patients and p_i is the number of variables in the client i. And $p = p_1 + p_2 + p_3$ is the total number of variables. 'Row_Block i' is defined as $[X_iVX_1, X_iVX_2, \ldots, X_iVX_k]$ binding k matrices column-wise where k is the number of clients.

Results

The proposed method's correctness is reported in **Table 1** using the the maximum absolute distance from the ground truth of the 20 estimates for 20 features from the synthetic dataset. All 20 coefficients specified in the simulation model achieved the near-perfect agreements. The runtime of the proposed method increased exponentially with an increase in the sample size and an increase in the number of clients(**Figure 2**). The runtime increased slightly when the number of features was increased. The effect of physical distance among clients and the server was evaluated using different cloud service providers' data centers. Six different Amazon Web Service (AWS) data centers were selected to co-locate all four clients, while keeping the server in Virginia, US. All four clients were scattered in four different places (blue line), which took the longest execution time. As a baseline (pink), the co-location of all four clients and the central server in one data center (Virginia) achieved the shortest computation time. From Dublin to Sydney, a remote data center was tested to observe the effect of the client data center's physical distance from the server data center, Virginia. Interestingly, trans-US (Oregon - Virginia) took longer run times than trans-Atlantic (Dublin - Virginia) or trans-America (Sao Paulo - Virginia). The reasons may lay on multiple jump boxes along with the

Number of Clients	Difference in Coefficient	Difference in Std Error
2	1.34×10^{-6}	5.31×10^{-8}
3	1.34×10^{-6}	5.22×10^{-8}
4	1.34×10^{-6}	5.34×10^{-8}

Table 1: The difference in parameter estimates in synthetic data. The difference was measure in the L_∞ norm, the maximum absolute distance from the ground truth of the 20 estimates. The dataset had 2000 samples, and 20 features were used.

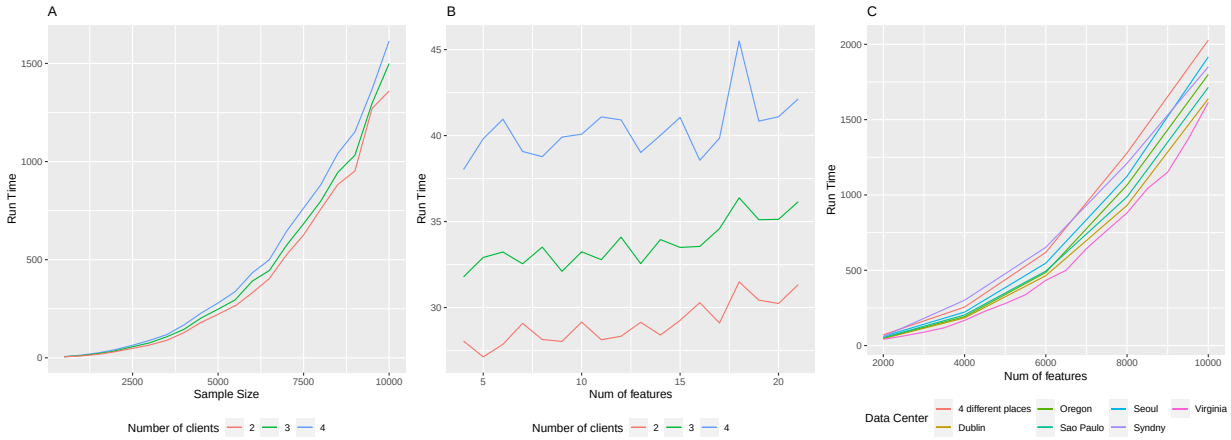


Figure 2: Computation time of the synthetic data. The time includes intermediate file transfer in two ways, client-to-client and client-to-server. **A**: Both sample size and number of clients varied under a fixed number of features = 20. **B**: Both feature and client numbers varied under a fixed sample size = 2000. **C**: Runtime by different AWS data centers. The blue line represents the run time of all four clients scattered in four different data centers away from Virginia, where server is located. The other colors represent the two data centers, one for co-locating all four clients and the other for the server site.

connection between Oregon and Virginia, while the submarine cables are connected directly. In BURN1000 (the first real dataset), VERTIGO-CI achieved the near-perfect agreement between the estimates and the ground truth (**Table 2**). Its average runtime varied between 12 and 15 seconds, with the number of clients ranging from 2 to 4 (**Table 3**). In PENNCATH (the second real dataset), the proposed method showed a good agreement between the federated and centralized coefficient estimates (**Table 4**). However, the estimated difference in standard error was the one order of magnitude larger than for the coefficient.The runtime increased linearly with the increase in the number of SNPs, and the mean running time for each trial can be seen in **Table 5**.

Discussion

We proposed a novel method of embedding the client-to-client part to enhance the interpretation of VERTIGO with hypothesis statistics like standard error, Z-score, p-value as well as confidence intervals for each coefficients. Using both synthetic and real datasets, we demonstrated the correctness of VERTIGO-CI by showing that its estimates are identical to those from the logistic regression with acceptable runtime with a small to mid-size number of features. Our proposed method's novel contribution is the standard error of the point estimates, which allows statistical decisions using P-value and confidence intervals. As the previous VERTIGO implied, the implementation of a fixed-Hessian matrix on Newton's method can highly reduce the computation complexity. However, the inversion of fixed-Hessian matrix is still non-trivial. And another potential problem is the size of gram-matrix during communication, gram matrix with $10,000 \times 10,000$ size can take up to 60 GB size. We have successfully implemented our VERTIGO-CI on a server in different sites but there is still a room for improvement in runtime to handle a very large number of features as in genomics data.

Variable	Coeff	Difference in Coeff	SE	Difference in SE
Intercept	-3.819841	4.978316e-07	0.296338	-5.805270e-08
facility	-0.176201	-3.277626e-07	0.139130	-3.553973e-08
age	2.075578	3.407076e-08	0.217424	-9.323797e-09
tbsa	1.741145	-5.004354e-08	0.179537	-1.389442e-08
gender_male	-0.069838	-2.018457e-08	0.142060	-9.766352e-09
race_white	-0.347684	-3.992930e-08	0.153023	-7.573084e-09
inhalation_injury	0.439069	7.644087e-08	0.118723	-1.191069e-08
flame_involved	0.291130	-1.952836e-07	0.178000	-2.107444e-08

Table 2: Accuracy of VERTIGO-CI in BURN1000 data. total burn space area (TBSA 0-100 %), flame involved in burn injury (flame), burn involved in inhalation injury (inh_inj), and Standard Error (SE)

Number of Clients	Mean running time (s)
2	12.4515
3	14.1357
4	15.9227

Table 3: The runtime in BURN1000 data with varied number of clients

Variable	Coeff	Difference in Coeff	SE	Difference in SE
sex	-1.200262	2.007786e-07	0.005065	1.391678e-01
age	-0.032013	1.330223e-06	0.144231	1.393521e-01
tg	0.011913	1.658586e-08	0.001869	7.267426e-04
hdl	0.015559	2.796433e-07	0.004879	1.855911e-04
ldl	0.006471	8.162119e-07	0.001142	7.268230e-04
pc1	1.057632	6.040747e-06	2.405702	3.457518e-05
pc2	-3.234316	1.851210e-05	2.378695	1.801008e-02
pc3	-2.172853	1.293278e-05	2.404689	2.596270e-02
pc4	-1.136879	7.012701e-06	2.392821	1.190317e-02
pc5	1.449743	8.630703e-06	2.408969	1.611641e-02
pc6	0.060668	8.945382e-08	2.401000	8.005713e-03
pc7	2.508987	1.462683e-05	2.424523	2.348583e-02
pc8	-3.037303	1.840741e-05	2.449719	2.515850e-02
pc9	-2.629828	1.576750e-05	2.422779	2.698205e-02
pc10	-0.983910	6.774366e-06	2.396671	2.614376e-02

Table 4: Accuracy of VERTIGO-CI in PENNCATH data. high-density lipoprotein (hdl), low-density lipoprotein (ldl), principal component (pc), tryglyceride (tg), and Standard Error (SE)

Number of SNPs	Mean runtime (s)	Standard deviation of runtime
10	260.2662	51.4977
100	2625.8944	13.6653
1000	26159.9742	0.6426

Table 5: The runtime with PENNCATH data with varied number of SNPs

Contributors

JK designed the study. JK and WL designed and implemented the core regression component. WL and TB developed the network programming component. XJ and LOM critically reviewed and edited the paper. All authors contributed to the manuscript preparation.

References

1. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. Journal of the American Medical Informatics Association. 2012;19(5):758–764.
2. Pletcher MJ, Forrest CB, Carton TW. PCORnet's collaborative research groups. Patient related outcome measures. 2018;9:91.
3. Monsey L, Best LG, Zhu J, DeCroo S, Anderson MZ. The association of mannose binding lectin genotype and immune response to Chlamydia pneumoniae: The Strong Heart Study. PLoS One. 2019;14(1):e0210640.
4. Charbonneau J, Nicol D, Chalmers D, Kato K, Yamamoto N, Walshe J, et al. Public reactions to direct-to-consumer genetic health tests: A comparison across the US, UK, Japan and Australia. European Journal of Human Genetics. 2019;p. 1–10.
5. Consortium GP, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
6. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology. 2015;19(1A):A68.
7. Chen F, Wang S, Jiang X, Ding S, Lu Y, Kim J, et al. Princess: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions. Bioinformatics. 2017;33(6):871–878.
8. Lu CL, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association. 2015;22(6):1212–1219.
9. Karr AF, Lin X, Sanil AP, Reiter JP. Secure regression on distributed databases. Journal of Computational and Graphical Statistics. 2005;14(2):263–279.
10. Slavkovic AB, Nardi Y, Tibbits MM. " Secure" Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases. In: Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). IEEE; 2007. p. 723–728.
11. Nardi Y, Fienberg SE, Hall RJ. Achieving both valid and secure logistic regression analysis on aggregated data from different private sources. Journal of Privacy and Confidentiality. 2012;4(1).
12. Yu H, Jiang X, Vaidya J. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. In: Proceedings of the 2006 ACM symposium on Applied computing; 2006. p. 603–610.
13. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. Vertical grid logistic regression (VERTIGO). Journal of the American Medical Informatics Association. 2016;23(3):570–579.
14. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. vol. 398. John Wiley & Sons; 2013.
15. Reilly MP, Li M, He J, Ferguson JF, Stylianou IM, Mehta NN, et al. Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. The Lancet. 2011;377(9763):383–392.
16. Minka T. A comparison of numerical optimizers for logistic regression (Technical Report). Microsoft Research. 2003;.
17. Seber GA, Lee AJ. Linear regression analysis. vol. 329. John Wiley & Sons; 2012.

18. Snyman JA, Wilke DN. Practical Mathematical Optimization: Basic Optimization Theory and Gradient-Based Algorithms. vol. 133. Springer; 2018.

Comparing Suicide Risk Insights derived from Clinical and Social Media data

Rohith K. Thiruvalluru*, MS[1], Manas Gaur*, MS[2], Krishnaprasad Thirunarayan, PhD[3], Amit Sheth, PhD[2], Jyotishman Pathak, PhD[1]
[1]Department of Population Health Sciences, Weill Cornell Medicine, USA;
[2]Artificial Intelligence Institute, University of South Carolina, USA;
[3]Kno.e.sis Center, Wright State University, USA

Abstract

Suicide is the 10^{th} leading cause of death in the US and the 2^{nd} leading cause of death among teenagers. Clinical and psychosocial factors contribute to suicide risk (SRFs), although documentation and self-expression of such factors in EHRs and social networks vary. This study investigates the degree of variance across EHRs and social networks. We performed subjective analysis of SRFs, such as self-harm, bullying, impulsivity, family violence/discord, using >13.8 Million clinical notes on 123,703 patients with mental health conditions. We clustered clinical notes using semantic embeddings under a set of SRFs. Likewise, we clustered 2180 suicidal users on r/SuicideWatch (∼30,000 posts) and performed comparative analysis. Top-3 SRFs documented in EHRs were depressive feelings (24.3%), psychological disorders (21.1%), drug abuse (18.2%). In r/SuicideWatch, gun-ownership (17.3%), self-harm (14.6%), bullying (13.2%) were Top-3 SRFs. Mentions of Family violence, racial discrimination, and other important SRFs contributing to suicide risk were missing from both platforms.

Introduction

Suicide is one of the leading causes of death in the US[1]. With an estimated increase of 61% in mental health patients per mental healthcare providers (MHPs) by 2025, it is hard to maintain patient engagement with treatment[2,3]. Further, predicting when someone will attempt suicide has been nearly impossible. Prior research has identified suicidal behavior using health insurance claims and electronic health record (EHR) data[2,3]. It is widely recognized that such data may have low sensitivity to detect suicidal behavior due to coding practices, reimbursement patterns, issues concerning ethics and safety, and the uncertainty of the patient's intent. Nevertheless, most of the clinically relevant data in EHRs, such as signs and symptoms and condition severity, are frequently available in narrative text from MHPs but not in structured and coded form. For instance, most clinically relevant information on mental health conditions, such as depression and suicidal ideation, is available in unstructured clinical notes. However, understanding and preventing suicide at an early stage requires data collected in real-time or through a source where individuals can express their life events and mood-related symptoms without the fear of social stigma. Social media platforms (e.g., Twitter, Reddit) can provide a rich source of insights on linguistic, interactional, and expressiveness features, complementing and supplementing clinical notes and interviews[4]. Similarly, the information derived from social media posts created by those experiencing suicidal ideation or attempt suicide may differ from what is typically documented in EHRs by MHPs, given the freedom of expression and its timeliness. For instance, a user makes the following post on r/SuicideWatch: "Really struggling with my *bisexuality*, which is causing chaos in my *relationship* with a girl. Being a fan of the LGBTQ community, I am equal to *worthless* for her. I'm now starting to *get drunk* because I can't cope with the *obsessive*, *intrusive thoughts*, the *need to isolate myself*, and *sleep forever*". Note its clinical relevance: it signals *Drug Abuse*, *Obsessive Compulsive Disorder*, *Suicidal Ideations*, and *Borderline Personality Disorder* (BPD). While the information on suicidality derived from unstructured EHR text can support point-of-care clinical decision making for suicide prevention, social media text can provide additional perspectives for public health interventions. In our research, we seek to demonstrate the supplementary and complimentary relationships between EHR and social media in recognizing individuals at risk of suicide. Leveraging the list of suicide risk factors (SRFs) identified by Jashinsky et al.[5], we showcase the similarities and dissimilarities in their manifestation on social media and EHR data. For this task, we utilize social media and EHRs from Reddit and Weill Cornell Medicine Ambulatory EHR clinical notes (WCM EHR), respectively. Recent research on identifying users with depression[6], estimating the severity of mental illness[4] or analyzing the change in user's expression as they change topics of their conversation depending on their current conditions, mental health communities (or MH-subreddits) on Reddit have been effective in gleaning actionable insights. The creation of communities specific to mental illnesses on Reddit (e.g., r/Depression, r/Autism,

r/PTSD, r/SuicideWatch) has facilitated clinical inferencing because of flexibility in the length of the post and the trust because of human moderation. Such characteristics have made Reddit provide nuanced content over Twitter. WCM EHRs are the unstructured and heterogeneous clinical notes are rich in content with granular details related to suicide risk. The user and content distributional similarity between Reddit and WCM EHR actuated our study to compare and contrast mentions of SRFs in both the platforms (see Datasets section).

In this study, we extract and compare SRFs derived from unstructured clinical text data and Reddit. While we mainly focus on user postings in r/SuicideWatch, we identify and gather semantically similar postings made by the user in other MH-subreddits. In this study, we label these emerging SRFs as "Other Important SRFs". Considering these key insights, we discuss the complementary or supplementary relationship between r/SuicideWatch and WCM EHRs in discerning expressions of SRFs. Subsequently, we report the limitations of our analytical study as a recommendation to improve future research focusing on associating clinical settings and social media for prompt assessment of suicidality of an individual.

Suicide risk prediction from unobtrusively gathered and up-to-date social media data has been beneficial in understanding suicide-related behaviors of users suffering from mental health conditions[5,7]. A study by Alvarez et al. showed that mediated social media-based therapy could successfully support early interventions for patients with the first episode of psychosis[7]. Platforms such as Twitter, Reddit, and Facebook have provided data relevant to depression, suicidality, postpartum depression, and post-traumatic stress disorders[8]. Twitter has been investigated for depression symptoms, suicide ideations, and SRFs for potential insights on early intervention in emergency[9]. Comparing the patient's perspectives gleaned from social media can suggest suicide risk factors that can supplement and complement the findings from EHR[10,11]. Merchant et al. conducted a study on $\sim$1000 consenting patients on Facebook suffering from 21 categories of mental conditions, including anxiety, depression, and psychosis, to show the utility of Facebook language as the screening tool estimate the onset of disease and conduct early interventions. The study concluded that contrasting insights from social media with EHR is necessary to leverage the findings for clinical use[8].

On the other hand, Roy et al.[12] and Gaur et al.[13] underlined that prior studies had ignored the use of clinical guidelines in their research, curtailing adoption of methods to practice . Studies from Chen et al.[14], Howard et al.[15] statistically explored natural behavioral language processing methods (e.g., use of first-person pronouns, sentiments, emotions, language models) and general lexicons (e.g., Linguistic Inquiry and Word Count (LIWC)) to study suicide risk without entailing clinical knowledge in the form of either medical resources or subject matter expert. We improve upon the limitations of previous work by incorporating relevant clinical information in the way of lexicons and medical knowledge bases in mental health care[16].

Datasets

The posts explicitly and implicitly expressing SRFs were extracted from Reddit using a set of domain-specific keywords that describe each SRFs independently. For instance, Depressive Feelings was expressed with following set of keywords: *thoughts, emotions, ranting, hopeless, ocd*, Self Harm was expressed with following set of keywords: *cuts, hurt, pills, overdose, tear, knife*. A more detailed list of keywords in provided on this link. The content covers 2,624,846 unique users and 2,469,893 posts across 15 mental health subreddits (MH-subreddits) (Bipolar, Borderline Personality Disorder, Depression, Anxiety, Opiates, Opiates Recovery, Self Harm, Stop Self Harm, BipolarSOs, Addiction, Schizophrenia, Autism, Aspergers, Crippling Alcoholism, BipolarReddit, SuicideWatch) over 11 years (2006-2016). This study focuses on users who have expressed suicidal tendencies on r/SuicideWatch and other mental health-related subreddits. We consider postings of a r/SuicideWatch in other subreddits as transient because they implicitly reflect a user's mental health status. As a consequence, we obtain posts with words related to some SRFs but posts are not related to SRFs. For instance: "People accidentally cutting while shaving" will be mis-labeled with following SRF "injury of unknown intent", if words like "hair", "shave", "slack", "accidentally" were not used to filter out. To filter such posts, we use a list of exclusion terms to prevent false positives while labeling posts with SRFs. A complete list of exclusion terms is provided on this link. Considering this as a proxy of understanding co-morbidity on Reddit, it is essential to identify a user's content in other subreddits and measure similarity with content often posted by other users in r/SuicideWatch. We followed a quantitative procedure, termed as *semantic relatedness* (a variant of cosine similarity measure). Thus, our final dataset contains 416,154 posts from 195,836 users with an average of $\sim$460 words per post, that is significantly larger and substantial than twitter dataset in Jashinsky et al.[5]. The Reddit dataset

is available for download on this link.

For the EHRs data, we used the EpicCare® Ambulatory EHR platform (used by Weill Cornell Medicine's (WCM)). The platform documents clinical care in its outpatient settings, which constitutes the EHR data used in this study. In particular, we extracted all clinical notes for n=123,703 patients who either had a diagnosis of major depressive disorder or have been prescribed an antidepressant between 2007 to 2017. Our corpus of all clinical records comprising more than 13.8 Million documents was authored by clinicians from multiple specialties, such as internal medicine, psychiatry, anesthesiology, pain medicine, across WCM outpatient clinics. Notes were heterogeneous in their content and level of detail and unstructured in their format. We subsequently generated representation of these datasets using word embedding models, which have shown to capture each word's meaning in context and derive clusters of semantically related words and phrases[17,18]. Finally, a comparative analysis was performed to meaningfully probe auxillary relationship between clinical and social media setting to understand the suicide risk factors.

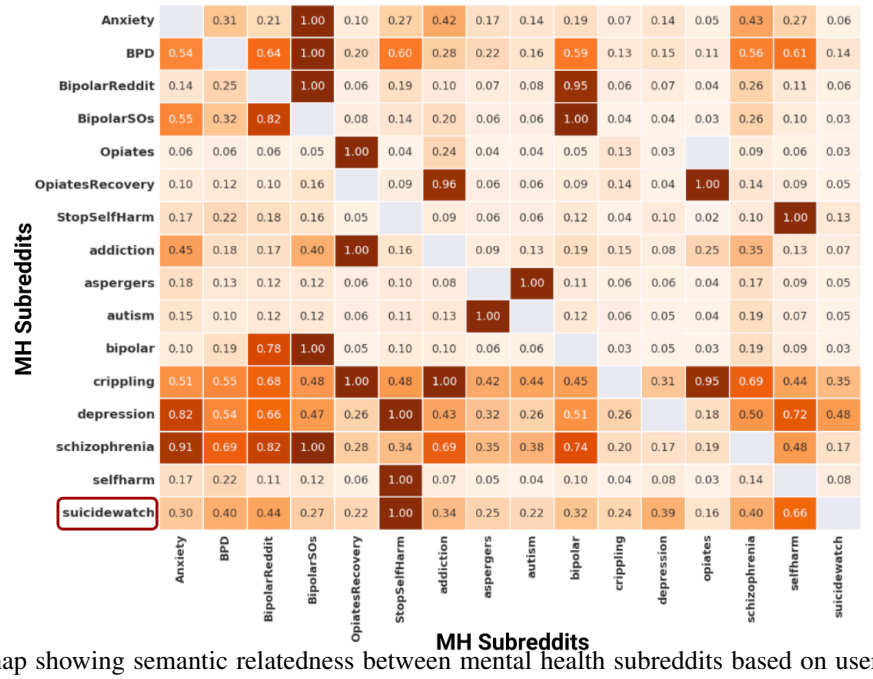


Figure 1: Heatmap showing semantic relatedness between mental health subreddits based on user overlap. For instance, r/SuicideWatch and BPD has a score 0.40, which signifies, of the number of users in common to both subreddits, content of 40% of users overlaps. Semantic relatedness is measured following equation 1. We have ignored comments in subreddits as they added minimal information gain.

Methods

We develop a method to quantify the semantic relatedness of the content produced by a user in r/SuicideWatch and other mental health subreddits (see Datasets section). A fundamental challenge that social media platform raise is false positives. A segment of resilient users on r/SuicideWatch shares their experiences to support others. Recognizing and separating supportive users from potentially suicidal users is essential for a reliable comparison with clinical notes. We utilize the suicide risk severity lexicon built using the Columbia-Suicide Severity Rating Scale (C-SSRS) to eliminate supportive content[19]. Consequently, we build an SRF-related lexicon as a composition of relevant lexicons created in past studies to recognize medical concepts in r/SuicideWatch subreddit and clinical notes[19]. Such a process is termed as entity (or concepts in lexicon) normalization and it requires a numerical representation in the form of a vector of length $\mathcal{L}$ ($V_{\mathcal{L}}$) (where $|V_{\mathcal{L}}|$ can be of either dimensions $\{300, 200, 100, 50\}$ and $V_{\mathcal{L}} \in \mathbb{R}$). This numerical vector representation or embedding of the word are generated using a word embedding model[19]. Since our task is specific to mental health and suicide risk, we require fine-tuning the word embedding model. For this purpose, we utilize SRF-related lexicon. After that, we used a non-parametric clustering approach to cluster embeddings and associate SRFs with them based on their semantic similarity (details in the Methodology section). On the analyzing the clusters labeled with one or set of SRFs, we found that SRFs such as "gun ownership" and "suicide around individual"

are often mentioned together in clinical notes, suggesting that gun was the medium for suicide. On the other hand, "gun ownership" frequently occurred with SRFs such as "depressive feelings", "psychological disorders", "suicide ideations" in the r/SuicideWatch. It suggests that SRFs: "depressive feelings", "psychological disorders", "suicide ideations" cause owning the gun. Hence, complementing the findings of clinical notes with Reddit would provide a better picture of the severity of an individual's suicide risk. Our analysis showed that the list of SRFs stated in Jashinsky et al. is not sufficient for a complete comparison of the two platforms. For instance, SRFs such as "poor performance in school", "relationship issues", "racial discrimination" are major contributors of suicide ideations but are not specified in Jashinsky et al.'s list of 12 SRFs: "depressive feelings", "depression symptoms", "drug abuse", "prior suicide attempts", "suicide around individual", "suicide ideation", "self-harm", "bullying behavior", "gun ownership", "psychological disorder", "family violence and discord", and "impulsivity"[5]. It is measured between two sub-reddits as the overlap in content made by users in common to both the subreddit. In our study, we formalize semantic relatedness (SR(S_1, S_2)) between two subreddits (S_1 and S_2) as follows:

$$\text{SR}(S_1, S_2) = \frac{\sum_{u \in S_1, S_2} \frac{\sum_{p_i \in \text{posts}_{(S_1)}, p_j \in \text{posts}_{(S_2)}} \delta(\vec{u}_{p_i}^{S_1}, \vec{u}_{p_j}^{S_2})}{|\text{posts}_{(S_1)}| + |\text{posts}_{(S_2)}|}}{\text{N}_u}; \delta(\vec{v}_x, \vec{v}_y) = \begin{cases} 1; \cos(\vec{v}_x, \vec{v}_y) > 0.9 \\ 0; \text{otherwise} \end{cases} \quad (1)$$

where $\vec{u}_{p_i}^{S_1}$ is the vector representation of a post (p_i) made by a user (u) in a subreddit (S_1) and N_u is the number of users common to both the subreddits. The threshold for the similarity is 0.9, which is empirically defined based on domain expert judgment. We follow this process over all the MH subreddits, as shown in Figure 1. From Figure 1, an SR score of 1.0 between r/SuicideWatch and r/StopSelfHarm suggest that users in common to these subreddits have more similar content compared to r/SuicideWatch and r/Opiates (SR score = 0.16). Based on the threshold of 0.40 set on the SR score, we extracted and gathered posts of r/SuicideWatch users in Depression, Addiction, Anxiety, Bipolar, Stop Self Harm, Self Harm, Borderline Personality Disorder (BPD), and Schizophrenia subreddits. The aggregated suicide-related content contains posts which have negations and conjunctions. Generating embeddings of these posts is erroneous as its is difficult to generate a semantic-preserving representations of posts with negations and conjuctions. Thus, we identified these posts and remove them for the study. Further, we leveraged a suicide risk severity lexicon (See details under Methods) to filter our posts which are not suicide-risk-related. With this pre-processing method we identified a cohort of 2180 (2.5% of 93K, $\sim$100K posts) users who were potentially suicidal through expressions of suicide risk factors and associated mental health conditions[20]. We extracted the content of these users in other MH subreddits and aggregated to create the dataset for the study. The reliability of the dataset was evaluated through an annotation performed over a randomly sampled 500 users ($\sim$30K Posts). The annotation was performed by psychiatrists using 5-labels: Supportive, Indicator, Ideation, Behavior, and Attempt, of which {Ideation, Behavior, Attempt} are defined in the Columbia-Suicide Severity Rating Scale. The content was annotated at the user-level and at the post-level. The inter-rater reliability score was recorded through pairwise and groupwise agreement using the Krippendorff metric[21]. Pairwise agreement is conducted between pairs of annotators and the annotator with high agreement score is selected for groupwise agreement. In this annotation agreement scheme, the annotations of the selected annotator is compared with mutually agreed annotations from an incremental group of annotators (in our case {2,3}). If there is a substantial agreement between the selected annotator and other groups of annotators with varied sizes, we consider the annotation, else process is repeated with next best annotator in pairwise scheme. Both the agreement schemes, together achieve robustness in the annotation task. Both at the user-level and post-level, we obtained a substantial inter-rater reliability score by measuring pairwise and groupwise agreement. At user-level, pairwise agreement was 0.79 and groupwise agreement was 0.69. In post-level, pairwise agreement was 0.88 and groupwise agreement was 0.76.

We describe an unsupervised and clinically grounded SRF-labeling methodology to identify and compare the different SRFs expressed in the voluminous r/SuicideWatch posts and clinical notes in EHRs. The proposed methods inputs sentence-level embeddings of the posts and clinical notes, and word embedding of the concepts in the SRF lexicon. The outcome, independent clusters of r/SuicideWatch posts, and clinical notes were associated with an SRF or set of SRFs by measuring the similarity between the embeddings of the centroid of the clusters and SRFs. The common and disparate SRFs were identified from the two platforms and compared (see Figure 2).

To identify SRF-related concepts (words or phrases) from online conversations on r/SuicideWatch, we require semantic

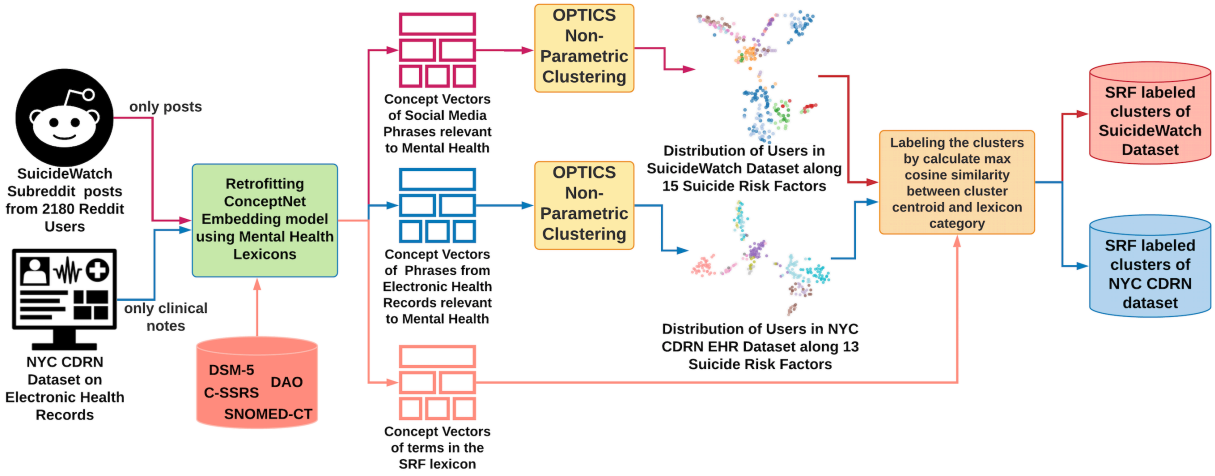


Figure 2: Overall workflow of an unsupervised approach to understand suicide risk factors (SRFs) in r/SuicideWatch posts and WCM EHR clinical notes using ConceptNet and semantic lexicons.

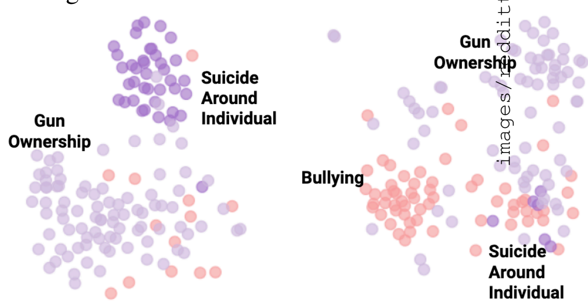
lexicons created specifically for suicide risk. We employ lexical-resources based on Columbia Suicide Severity Rating Scale (C-SSRS)[21] and Diagnostic and Statistical Manual of Mental Disorders (DSM-5) developed by Gaur et al.[19]. We used the concepts in the created SRF lexicon as the seed concepts to enrich it with terms in SNOMED-CT, and Drug Abuse Ontology (DAO)[19] following a guided markovian random walk procedure. An instance of the created SRF lexicon associates " Suicide Ideations" with "intrusive thoughts". To verify this association, we trace the path from "Obsessive-compulsive disorder" ([SNOMEDCT: 1376001]) to "intrusive thoughts." Obsessive-compulsive disorder [SNOMEDCT: 1376001] is associated with Suicidal Ideations [SNOMEDCT: 425104003], and the concept "Intrusive Thoughts" [SNOMEDCT: 225445003] is a child of parent concept "Disturbance in Thinking" [SNOMEDCT: 26628009], which is a child concept of "Obsessive-Compulsive Disorder." Thus proving the association between "intrusive thoughts" and "Suicidal Ideations". The SRF lexicon acts as a component to fine-tune a generic word embedding model, ConceptNet[21], to generate contextualized representations of r/SuicideWatch posts and clinical notes in WCM EHRs. This post-processing technique is called retrofitting, and it reinforces the embedding of words by minimizing the distance between concepts that are relevant in describing SRFs[20]. For example, in the retrofitted ConceptNet, "hopeless" and "depressive feelings" are in proximity compared to "hopeless" and "harassment." The proximity suggests that depressive feelings are expressed with the term "hopeless" more often than "harassment" in suicide risk-related conversations. Another example is the semantic proximity of "impulsivity" to "bullying" rather "family violence and discord."

After retrofitting of ConceptNet embedding model, we leverage it to generate vector representations of each post in the suicide dataset created from r/SuicideWatch and other relevant online mental health communities. Similarly, we create representations of clinical notes documented in EHRs. Note that our method to create representation is post-level and clinical notes-level, not user-level. Psychiatrists treat a siloed community of patients suffering from mental health disorders, which restricts diversification. A strategic comparison of clinical notes in EHRs with population-level social media markers could enable the psychiatrists to develop better contextual questions in diagnostic interviews and elucidate disease epidemiology for better patient engagement. The representations of the two sources of content were clustered independently using a non-parametric clustering algorithm, OPTICS (Ordering Points To Identify the Clustering Structure)[9]. Our selection of OPTICS over approaches such as DBScan, K-Means, Gaussian Mixture Models, is based on the clustering algorithm's ability to create diverse (at least equal to the number of SRFs) clusters, where each cluster most-likely cohesively represents an SRF. We calculate the similarity between the representation of the centroid of the cluster and the SRFs. The SRF, with the highest similarity with the centroid, is the estimated label of the cluster. We followed this process to label clusters created from suicide dataset and clinical notes in EHRs.

Results and Discussion

On the clusters labeled with a set of SRFs, this study discusses the commonalities and differences in the expressions of suicide-risk from patients and users in WCM EHRs and r/SuicideWatch respectively. For instance, *depressive feelings, psychological disorders, drug abuse, and suicide ideations* are the common SRFs communicated on both platforms. However, both the platform differs from each other concerning following SRFs: *depressive symptoms* and *suicide around individual* is revealed only from clinical notes; *bullying behavior, self-harm, impulsivity*, and *family violence and discord* significantly manifests in r/SuicideWatch communications only. We ranked the SRFs independently for each platform. In clinical notes, most frequent SRFs are depressive feeling (24%), psychological disorders (21.1%), drug abuse (18.2%), depressive symptoms (14.9%), suicide around individual (12.6%), and suicide ideations (9.1%) (see Figure 4a). On analyzing the clusters derived from the EHR data, we observed mentions of gun ownership contextualizes bullying behavior and suicide around individual (see Figure 3a).

In r/SuicideWatch posts, gun ownership (17.4%), self-harm (14.6%), bullying behavior (13.2%), drug abuse (13%), depressive feelings (11.6%), suicide ideation (10.7%), psychological disorders (10%), impulsivity (9.6%) are frequently discussed (see Figure 3b and Figure 4b). A semantic analysis of the posts on r/SuicideWatch showed family violence and discord as the reason for impulsivity, leading to a suicide attempt (see Figure 4b). Further, suicidality measured through co-occurrence of drug abuse and bullying behavior often showed a high frequency of terms mapped to family violence, then depressive feelings.



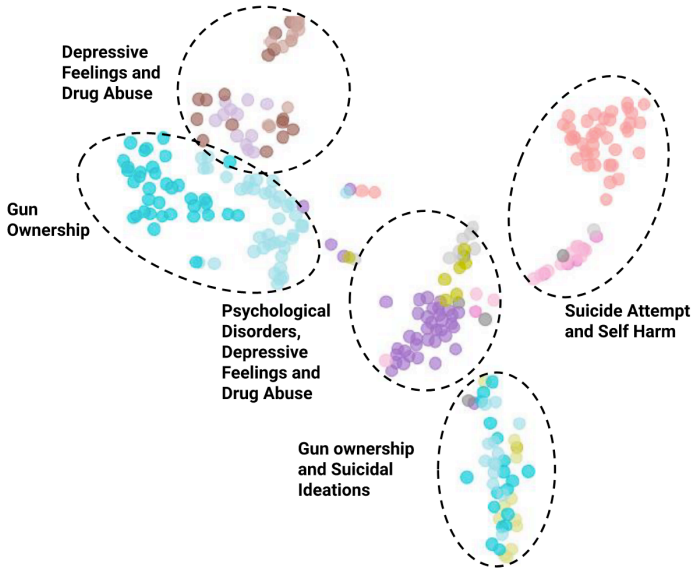
(a) Top three SRFs elicited from clustering clinical notes in WCM EHR dataset. SRF such as Suicide around individual is often mentioned with gun ownership and bullying behavior.

(b) Top five SRFs elicited from clustering r/SuicideWatch posts. Depressive feelings, bullying behavior, and drug abuse are often mentioned together. Suicide ideations are expressed by users mentioning signs of depressive feelings and threatening to possess gun.

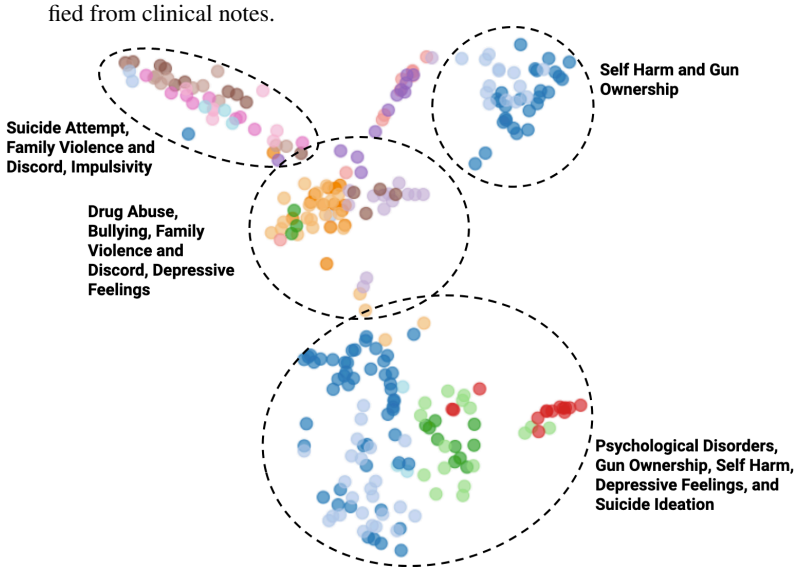
Figure 3: Description of top clusters showing interrelation between different SRFs identified in r/SuicideWatch and clinical notes.

However, we seldom saw the manifestation of SRF, "family violence and discord" in the documented clinical notes. Clinical settings steered away from discussing family violence and impulsivity, significant SRFs mentioned on r/SuicideWatch. This may indicate that individuals who act impulsively, leading to the risk of committing suicide, are less inclined to discuss such behavior privately with their MHPs. On the contrary, Depressive feelings were prominent SRF followed by psychological disorders and drug abuse in clinical settings. Figure 4a shows that semantic clusters of suicidal ideations and gun ownership were formed from the representations of clinical notes, which is consistent with the clusters derived from r/SuicideWatch and is illustrated in Figure 4b. Further, written communications from users on r/SuicideWatch mention suicide ideations when describing psychological disorders and guns as threats to self-harm. These findings are relatable with clinical insights derived from C-SSRS[21].

Both the platforms show content from users (or patients) discussing gun ownership and bullying behaviors when expressing potential suicide risk. In contrast, r/SuicideWatch clusters were more revealing by co-locating other SRFs such as drug abuse and self-harm. Considering the findings from the analysis of the suicide risk-related content this study specify a supplementary and complementary relationship between r/SuicideWatch and EHRs. Besides, while our clinically grounded approach demonstrated the feasibility to intervene among those at suicide risk consensually as-



(a) SRF clusters of clinical notes. Out of 14, 7 SRFs were identified from clinical notes.



(b) SRF clusters of r/SuicideWatch posts. Out of 14, 10 SRFs were identified from the posts made by users on r/SuicideWatch.

Figure 4: Clusters of SRFs identified from clinical notes and r/SuicideWatch posts.

sociating their social media profile with EHR data, we plan to elucidate certain computational and practical limitations in our future research.

Limitations

The findings described in the study should be interpreted in the context of some limitations concerning (a) the semantic lexicon, (b) the suicide dataset, and (c) the WCM EHR dataset. The abstraction of posts made on r/SuicideWatch using the semantic lexicon comes with a limitation regarding its completeness. The lexicon utilized in this study is a composition of concepts in PHQ-9, C-SSRS, DAO, and SNOMED-CT, which was semantically appropriate for this study. But it can be improved with slang terms and moderation from domain experts. However, care must be taken while extending the lexicon as it brings ambiguity which might falsely determine SRFs. The suicide dataset prepared by strategically accumulating the content from MH-subreddits ignored some subreddits directly related to

r/SuicideWatch but are sparse. For example, r/euthanasia (assisted suicide) and r/suicideprevention, are other discussion forums on suicide risk that were not included in this study.

In addition to the list of SRFs provided by Jashinsky et al.[5], we found additional topics such as "relationship issues", "brain damage", "physiological stressors", "cant afford rent, debt, failure", and "unemployment" that are significant contributors to suicide ideations but could not be mapped to the existing list and require a clinically relevant SRF label. For now, we considered these stressors as "Other Important SRFs," and in this study, we did not provide a comparison between the two platforms based on this category. This is because it is a mixed category in terms of SRFs; hence a vector representation would be semantically misleading, causing false inferences. Likewise, we formed another category, termed as "Accessory," which contains phrases having mention of a material or substance that assisted suicide. For instance "self-inflicted injury by suffocation by plastic bag" (a plastic bag is an accessory), "suicide or self injury by jumping from bridge" (the bridge is a navigational concept [SNOMEDCT: 242843002]), "suicide or self injury by caustic substance", "attempt suicide by car exhaust (event is an accessory)", "indirect self harm due to mechanical threat" (trapped in a car trunk, refrigerator, etc.). Like "Other Important SRFs", "Accessory" is a mixed category, we could not assess the commonalities and disparities between r/SuicideWatch and WCM EHR clinical notes. As future work, we will explore these incohesive categories from a clinical perspective and further strengthen our study with demographic and spatial information. Additionally, we will explore ways to differentiate between suicide completers and suicide attempters.

Conclusion

Suicide risk factors can be determined and used for suicide prevention at an early stage; however, poorly documented clinical notes curtail MHPs from devising intervention strategies. Further, EHRs shed some light on a patient's current and anew psychopathology status, but fall to cultivate a broader understanding of the mental health conditions. Recently, people have investigated social media, mainly Reddit, to gather insightful population-level markers for assisting MHPs. However, ambiguous and sparse content in Reddit and EHRs requires structured hierarchical knowledge for apprehension and effective decision making. In this work, we investigated the commonality and disparity in the conversations specific to SRFs from users on r/SuicideWatch and patients in the clinical setting. In the process, we created an SRF-specific lexicon for semi-automatically identifying medical concepts on r/SuicideWatch and clinical notes for contextualization and semantic clustering of SRFs. We observed a few similarities between the SRFs discussed within the private EHR data versus an anonymized, public setting, SW. Simultaneously, many dissimilarities were observed across the datasets suggesting future studies should focus on linking clinical and non-clinical data at an individual level to get a comprehensive view of an individual's suicide risk. The post-level and user-level annotated dataset created from r/SuicideWatch will be made publicly available upon acceptance of the study. Further, the source code developed to conduct this study will be made online on github for reproducibility.

Acknowledgement

This work was funded in part by NIH grants R01MH105384, R01MH119177, and P50MH113838. Any opinions, findings, and conclusions/recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NIH.

References

[1] James Tang, Yizhen Yu, Holly C. Wilcox, Chun Kang, Kun Wang, Cunku Wang, Yu Wu, and Ruoling Chen. Global Risks of Suicidal Behaviours and Being Bullied and Their Association in Adolescents: School-Based Health Survey in 83 Countries. *SSRN Electronic Journal*, 2020.

[2] Identifying suicidal behavior among adolescents using administrative claims data. *Pharmacoepidemiology and Drug Safety*, 22(7):769–775, 2013.

[3] Barak-Corren Y., Castro V.M., Javitt S., Hoffnagle A.G., Dai Y., Perlis R.H., Nock M.K., Smoller J.W., and Reis B.Y. Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, 174(2):154–162, 2017.

[4] Manas Gaur, Amit Sheth, Ugur Kursuncu, Raminta Daniulaityte, Jyotishman Pathak, Amanuel Alambo, and Krishnaprasad Thirunarayan. "Let me tell you about your mental health!" Contextualized classification of reddit posts to DSM-5 for web-based intervention. *International Conference on Information and Knowledge Management, Proceedings*, pages 753–762, 2018.

[5] Jared Jashinsky, Scott H. Burton, Carl L. Hanson, Josh West, Christophe Giraud-Carrier, Michael D. Barnes, and Trenton Argyle. Tracking suicide risk factors through Twitter in the US. *Crisis*, 35(1):51–59, 2014.

[6] Sharath Chandra Guntuku, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 2017.

[7] Alvarez-Jimenez M, Bendall S, Lederman R, Wadley G, Chinnery G, Vargas S, Larkin M, Killackey E, McGorry PD, and Gleeson JF. On the HORYZON: moderated online social therapy for long-term recovery in first episode psychosis. *Schizophrenia Research*, 143(1):143–149, 2013.

[8] Raina M. Merchant, David A. Asch, Patrick Crutchley, Lyle H. Ungar, Sharath C. Guntuku, Johannes C. Eichstaedt, Shawndra Hill, Kevin Padrez, Robert J. Smith, and H. Andrew Schwartz. Evaluating the predictability of medical conditions from social media posts. *PLoS ONE*, 14(6), 2019.

[9] Jianhong Luo, Jingcheng Du, Cui Tao, Hua Xu, and Yaoyun Zhang. Exploring temporal suicidal behavior patterns on social media: Insight from Twitter analytics. *Health Informatics Journal*, 26(2):738–752, 2020.

[10] Sumithra Velupillai, Gergö Hadlaczky, Enrique Baca-Garcia, Genevieve M. Gorrell, Nomi Werbeloff, Dong Nguyen, Rashmi Patel, Daniel Leightley, Johnny Downs, Matthew Hotopf, and Rina Dutta. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. *Frontiers in Psychiatry*, 10(FEB), 2019.

[11] Shades of Knowledge-Infused Learning for Enhancing Deep Learning. *IEEE Internet Computing*, 23(6):54–63, 2019.

[12] Arunima Roy, Katerina Nikolitch, Rachel McGinn, Safiya Jinah, William Klement, and Zachary A. Kaminsky. A machine learning approach predicts future risk to suicidal ideation from social media data. *npj Digital Medicine*, 3(1), 2020.

[13] Manas Gaur, Ugur Kursuncu, Amit Sheth, Ruwan Wickramarachchi, and Shweta Yadav. Knowledge-infused Deep Learning. pages 309–310, 2020.

[14] Qijin Cheng PhD, Tim M H Li PhD, Chi-Leung Kwok PhD, Tingshao Zhu PhD, and Paul S F Yip PhD. Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. *Journal of Medical Internet Research*, 19(7), 2017.

[15] Derek Howard, Marta Maslej, Justin Lee, Jacob Ritchie, Geoffrey Woollard, and Leon French. Transfer learning for risk classification of social media posts: Model evaluation study. *arXiv*, 2019.

[16] Manas Gaur, Keyur Faldu, and Amit Sheth. Semantics of the Black-Box: Can knowledge graphs help make deep learning systems more interpretable and explainable? 2020.

[17] T Mikolov, I Sutskever, K Chen, G Corrado, and J Dean. Distributed representations of words and phrases and their compositionality. In: Conference on Advances in Neural Information Processing Systems. *Distributed Representations of Words and Phrases and Their Compositionality*, pages 3111–3119, 2013.

[18] Suneel Kumar Kingrani, Mark Levene, and Dell Zhang. Estimating the number of clusters using diversity. *Artificial Intelligence Research*, 7(1):15, 2017.

[19] Manas Gaur, Ugur Kursuncu, Amit Sheth, Amanuel Alambo, Krishnaprasad Thirunarayan, Randon S. Welton, Joy Prakash Sain, Ramakanth Kavuluru, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, pages 514–525, 2019.

[20] Amanuel Alambo, Manas Gaur, and Krishnaprasad Thirunarayan. Depressive, drug abusive, or informative: Knowledge-aware study of news exposure during COVID-19 outbreak. *CEUR Workshop Proceedings*, 2657:17–22, 2020.

[21] Zimri S. Yaseen, Mariah Hawes, Shira Barzilay, and Igor Galynker. Predictive Validity of Proposed Diagnostic Criteria for the Suicide Crisis Syndrome: An Acute Presuicidal State. *Suicide and Life-Threatening Behavior*, 49(4):1124–1135, 2019.

Severity Prediction for COVID-19 Patients via Recurrent Neural Networks

Junghwan Lee, MA[1], Casey Ta, PhD[1], Jae Hyun Kim, PhD[1], Cong Liu, PhD[1*], Chunhua Weng, PhD[1*] (*: equal contribution)
[1]Department of Biomedical Informatics, Columbia University, New York, N.Y.

Abstract

The novel coronavirus disease-2019 (COVID-19) pandemic has threatened the health of tens of millions of people worldwide and imposed heavy burden on global healthcare systems. In this paper, we propose a model to predict whether a patient infected with COVID-19 will develop severe outcomes based only on the patient's historical electronic health records (EHR) prior to hospital admission using recurrent neural networks. The model predicts risk score that represents the probability for a patient to progress into severe status (mechanical ventilation, tracheostomy, or death) after being infected with COVID-19. The model achieved 0.846 area under the receiver operating characteristic curve in predicting patients' outcomes averaged over 5-fold cross validation. While many of the existing models use features obtained after diagnosis of COVID-19, our proposed model only utilizes a patient's historical EHR to enable proactive risk management at the time of hospital admission.

INTRODUCTION

The novel coronavirus disease-2019 (COVID-19) has threatened the health of tens of millions of people over the world and imposed heavy burden on global healthcare systems. To fight against the pandemic and mitigate the burden, numerous efforts have been made by scientists to develop risk prediction models for COVID-19 patients. Prognostic models, among the important risk prediction models, has been developed to predict risks of mortality[1-3] and progression to severe status[4-6] for COVID-19 patients. Commonly used predictors for those COVID-19 prognostic models include comorbidities, age, sex, lab test results (e.g., lymphocyte count, C reactive protein, and creatinine), and radiologic imaging features[7]. The existing models, however, spanning from Cox proportional hazards models to state-of-the-art machine learning and deep learning models, heavily rely on features obtained after hospital admission or diagnosis of COVID-19 for post-diagnosis prognosis[7].

Recurrent neural networks (RNN) have been widely used in modeling sequential phenomena such as speech and language due to its strengths of capturing hidden relationships between the sequential data[8]. There have been several studies in the healthcare domain that used RNN to predict future medical events or the risk of certain diseases, leveraging the sequential nature of electronic health records (EHR). For example, Lipton et al.[9] and Choi et al.[10] both used RNN for predicting future medical events based on the historical EHR data. Choi et al.[11] published related work using RNN for predicting the risk of heart failure based on the patient's historical EHR data.

In this study, we applied RNN on a patient's historical EHR data to predict the patient's risk of developing severe outcomes from COVID-19, including mechanical ventilation, tracheostomy, or death. The prediction represents the probability for a patient to progress into a severe status after being infected with COVID-19. One major advantage of this method is that the model does not require any data after the diagnosis of COVID-19 (e.g., lab test results and vital signs), so that it can predict the risk of developing severe outcomes from COVID-19 for a patient before or at the time of hospital admission. This advantage allows proactive risk management by the clinical care team and resource allocation in advance, which can be critical for health policy makers and hospital administrators.

METHODS

COVID-19 Cohort Description

New York City has been one of the epicenters of the COVID-19 pandemic. NewYork Presbyterian Hospital/Columbia University Irving Medical Center (NYP/CUIMC) has treated a large cohort of COVID-19 patients since the onset of the pandemic. For this work, we obtained all EHR data for the patients infected with COVID-19 updated until May 31, 2020 from NYP/CUIMC's Observational Medical Outcomes Partnership (OMOP) database, which contains 30 years' worth of comprehensive EHR data for about 6.5 million patients. This work received institutional review board approval (AAAR3954) with a waiver for informed consent.

The COVID-19 cohort was identified as patients 18 years or older who were hospitalized and tested positive for SARS-CoV-2 within 21 days before or during their hospitalization. The patients must have at least one visit record prior to March 1, 2020 and with at least one condition (i.e. diagnosis) concept. We obtained all condition concepts in historical inpatient and outpatient visits prior to the hospital admission due to infection of COVID-19 for the identified patients in the cohort in temporal order. In total, 5,774 unique condition concepts were identified from all patients in the cohort. Demographic information (i.e. sex and age at the most recent hospital admission) of the patients were also obtained. Characteristics of the COVID-19 cohort are shown in **Table 1**. We classified patients in the COVID-19 cohort into two groups: severe vs. moderate. Severe patients were identified as the patients who had at least one of the following outcomes during hospitalization: mechanical ventilation, tracheostomy, or death; these events correspond to a severity score of ≥ 6 in the World Health Organization ordinal scale for clinical improvement[12]. Moderate patients refer to the patients who were either discharged without developing severe outcomes during hospitalization or were still hospitalized but without any signal of the severe outcomes.

Table 1. Characteristics of the COVID-19 cohort. Senior patients: aged ≥ 65 at the most recent hospital admission. SD: standard deviation.

	Severe patients	Moderate patients
Total # of patients	546	1,828
# of patients with either mechanical ventilation or tracheostomy	15	-
# of death	531	-
# of senior patients (%)	455 (83.3%)	878 (48.0%)
male patients (%)	322 (59.0%)	891 (48.7%)
Avg. age of the patients (SD)	76.79 (12.92)	61.34 (18.29)
Median # of visits per patient (25 percentile, 75 percentile)	16.0 (4.0, 46.0)	12.0 (3.0, 38.0)

Problem Definition

For notation, we denote vectors with italic bold lower-case (e.g. $\boldsymbol{h_1}$, $\boldsymbol{x_1}$), matrices with italic bold upper-case (e.g. $\boldsymbol{W_{FC}}$), and scalars with italic lower-case (e.g. $\hat{y}$). For notational convenience, we assume that the input for the model is a single patient.

For each patient in the cohort, all historical inpatient and outpatient visits were extracted in the form of multi-hot encoded vector $\boldsymbol{x_i}$ for $i = 1, \ldots, p$, where p is the number of total visit that the patient made before the hospital admission due to COVID-19. Inpatient visits included emergency room visits or hospitalizations via emergency room. The multi-hot encoded vector $\boldsymbol{x_i} \in [0,1]^k$ represents the i-th visit of the patient, where k denotes the number of unique medical concepts observed in the cohort. $\boldsymbol{x_i^l}$ is 1 if the l-th medical concept was observed in the patient's i-th visit and 0 otherwise. Our goal is to predict a patient's risk of developing severe outcomes based on the patient's historical EHR data. The predicted risk score ranges between 0 and 1 and represents the estimated probability for the patient to progress into a severe outcome from COVID-19.

Model Architecture

The proposed RNN model to predict the risk score is depicted in **Figure 1**. At each timestamp i, the model receives a patient's visit x_i and the previous hidden state h_{i-1} as input and outputs hidden state h_i for $i = 1, \ldots, p$, where p is the number of total visit that the patient made. We used Gated Recurrent Units[13] (GRU) for the RNN model in this work. Although Long Short Term Memory[14] (LSTM) is the most widely used RNN cell among all other RNN variants and generally outperforms GRU on large datasets[15,16], GRU show comparable or better performance on tasks with relatively small datasets with fewer parameters[17]. Preliminaries of GRU are available in the supplementary material.

For efficient training of the model, we used an embedding layer that transforms the multi-hot encoded input $\boldsymbol{x_i}$ into a low-dimensional embedding (described below). The hidden state at the last timestamp is concatenated with the patient's demographic information vector and subsequently fed into a fully connected layer with hyperbolic tangent activation. Finally, an output layer that contains a single neuron with sigmoid activation (i.e. logistic regression layer)

is applied over the output of the fully connected layer to generate the risk score of the patient as defined in Eq(1) and Eq(2):

$$\boldsymbol{o_{FC}} = tanh(\boldsymbol{W_{FC}}[\boldsymbol{h_p}, \boldsymbol{d}] + \boldsymbol{b_{FC}}) \qquad Eq(1)$$

$$\hat{y} = \sigma(\boldsymbol{W_{LR}}\boldsymbol{o_{FC}} + b_{LR}) \qquad Eq(2)$$

where $\boldsymbol{W_{FC}}$, $\boldsymbol{W_{LR}}$, $\boldsymbol{b_{FC}}$, b_{LR}, $\boldsymbol{o_{FC}}$, $\boldsymbol{h_p}$, $\boldsymbol{d}$ and $\hat{y}$ denote the weight matrix of the fully connected layer, weight matrix of the logistic regression layer, bias of the fully connected layer, bias of the logistic regression layer, the output vector of the fully connected layer, the hidden state at the last timestamp p, demographic information vector, and the predicted risk score of the patient respectively. $[\cdot,\cdot]$ denotes vector concatenation, *tanh*() denotes the hyperbolic tangent activation function, and $\sigma()$ denotes the sigmoid activation function. A patient's demographic information vector is a simple concatenation of one-hot encoded sex (i.e. [1, 0] for male and [0, 1] for female) and min-max normalized age of the patient.

The true label y for each patient was determined based on outcome status of the patient as observed in the CUIMC database: we assigned 1 for severe patients and 0 for moderate patients. Since severe and moderate cases were imbalanced in the dataset, we used weighted cross entropy loss, defined as Eq(3):

$$L = -\sum_{j=1}^{N}(wy^{(j)}\log\hat{y}^{(j)} + (1-w)(1-y^{(j)})\log(1-\hat{y}^{(j)})) \qquad Eq(3)$$

where $y^{(j)}$, $\hat{y}^{(j)}$, N, and w are the true label for the j-th patient, the predicted risk score for the j-th patient, the total number of patients in the batch, and weight for the cross entropy. We used 0.75 for the weight of the cross entropy considering the ratio of the severe and moderate patients in the cohort to provide more weight on accurately predicting severe cases (i.e. more focus on sensitivity).

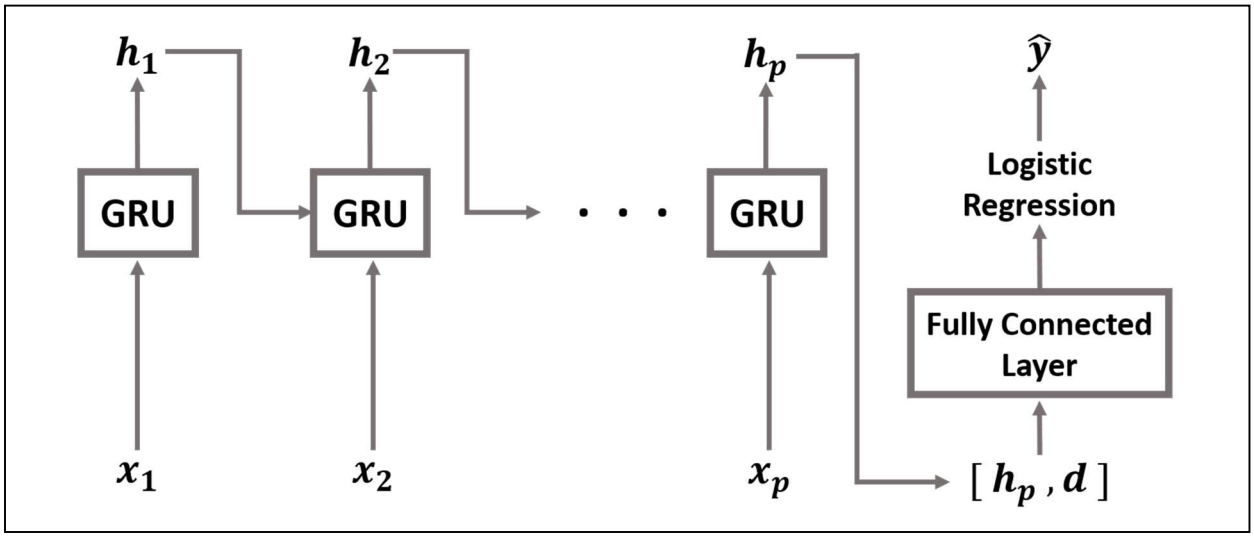


Figure 1. The architecture of the proposed recurrent neural network model. GRU: Gated Recurrent Unit.

RESULTS

Experiment Setup

To evaluate the performance of the RNN model, we compared the average area under the receiver operating characteristic curve (AUC) based on 5-fold cross validation with two other baselines – logistic regression and multilayer perceptron (MLP). The entire dataset was divided into 5 chunks: 3, 1, and 1 chunk(s) were allocated to the training set, validation set, and test set respectively (i.e. 60% training, 20% validation, 20% test split). Different combinations of chunks were allocated to the training set, validation set, and test set at every fold, thus the model was trained, validated, and tested on different datasets at every fold. All models were trained with a maximum of 50 epochs at every fold and the model achieved the highest AUC on the validation set was finally used for test set evaluation. We reported the average and standard error of AUCs of all 5 folds based on the test set.

We used an embedding layer to transform the multi-hot encoded input $\boldsymbol{x_i}$ into a low-dimensional embedding. We experimented with two different initializations of the embedding layer in the RNN model: (1) the embedding layer initialized with a random normal distribution; (2) the embedding layer initialized with pre-trained embedding. Random normal distribution with mean 0 and standard deviation 0.01 was chosen for initialization since it showed better performance than many other baselines in word embedding tasks[18]. We pre-trained an embedding using GloVe[19] on the co-occurrence matrix obtained from the cohort for 100 epochs. The pre-trained embedding captures the relationships between the medical concepts since GloVe utilizes the global co-occurrence matrix of concepts for its training, where the co-occurrence matrix is calculated based on the concept co-occurrence in every patients' visit. In the literature, the dimensionality of the embedding is generally set between 100-500 for medical concept vocabularies with sizes from a few hundred to tens of thousands of concepts[11,20], therefore we set the dimensionality of the pre-trained embedding and randomly initialized embedding to 128. The embedding layer was fine-tuned jointly with the prediction task of the model.

Since mini-batch training shows good generalization performance when the size of the data is relatively small[21], we used a small batch of size 2 in the training. We also empirically found that prediction performance of the model decreased with larger batch size. To prevent the model from overfitting, L_2 weight decay with regularization coefficient of 0.001 was applied to weights of the fully connected layer in the RNN model. We tried dropout[22] to non-recurrent connection of the RNN model and found that dropout did not improve the performance of the model, therefore we did not use dropout.

Baselines

Logistic regression

A simple logistic regression model was used for the first baseline with three different types of input: aggregated multi-hot encoded vector, aggregated embedding, and aggregated pre-trained embedding. For each patient, aggregated multi-hot encoded vector is summation of input x_i at all timestamps, after which is clipped with maximum value to 1. Aggregated embedding and aggregated pre-trained embedding were generated by passing the aggregated multi-hot encoded vector through randomly initialized embedding layer or pre-trained embedding layer respectively. Those two embedding layers were initialized using the same scheme as the RNN model. Aggregation of the input can be understood as the summation for each concept observed across visits in a patient's history. All aggregated inputs were normalized to zero mean and unit variance for numeric stability during training. L_2 weight decay with regularization coefficient of 0.001 was applied to weights in the model to reduce overfitting.

Multilayer perceptron

Multilayer perceptron (MLP) with a single hidden layer was used for another baseline. A fully connected layer with hyperbolic tangent activation was used for the hidden layer and the output layer contains a single neuron with sigmoid activation. The three different types of inputs (aggregated multi-hot encoded vector, aggregated embedding, and aggregated pre-trained embedding) were used with the same settings as described above. The number of hidden units in the hidden layer was set to 128. L_2 weight decay with regularization coefficient of 0.001 was applied to weights in the model to reduce overfitting.

Implementation Details

We used Tensorflow 2.0.0[23] to implement the RNN model and all baselines. Adam[24] was used for optimization in training for all models. A machine equipped with 2 × Intel Xeon Silver 4110 CPUs and 192GB RAM was used. Hyperparameters and some important details of training are provided in the supplementary material. The source codes to implement all models are publicly available at https://github.com/Jayaos/rnn-covid.

Prediction Performance of the Risk Score

We calculated the average AUC of 5-fold cross validation to evaluate the prediction performance of the risk score generated by the models (**Figure 2**). Overall, the RNN model with pre-trained embedding achieved the highest average AUC (0.846). The RNN model also showed higher average AUC than the baselines when comparing the same embedding layer initialization schemes.

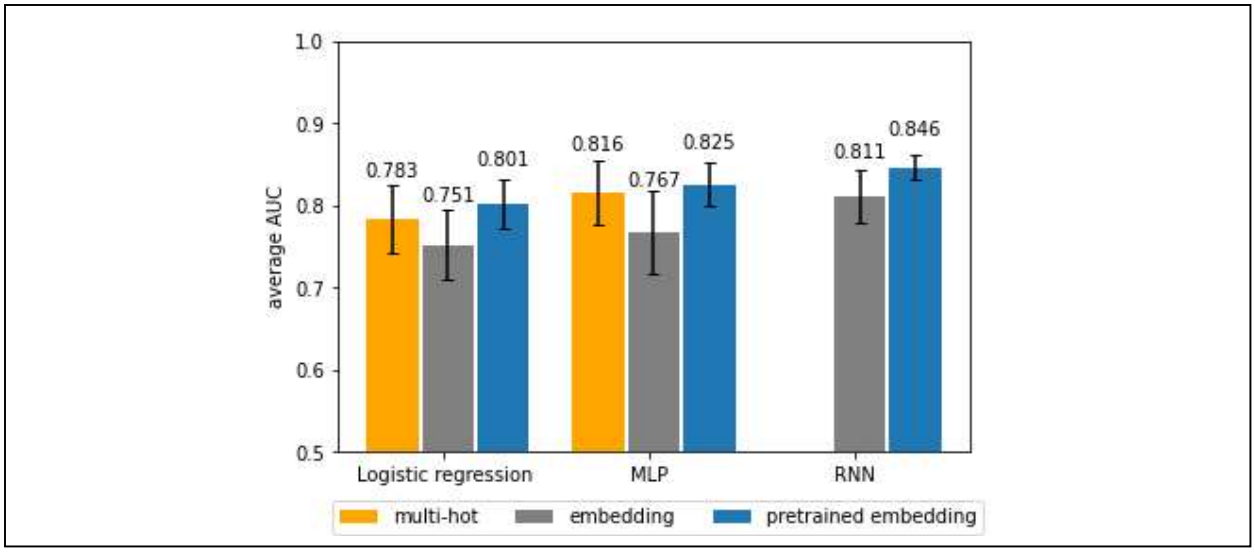


Figure 2. Average 5-fold cross validation AUC of all models. The values of average AUC and standard error of AUC for each model are provided in the supplementary material.

Prediction Time

For the RNN model with pre-trained embedding, which showed the best performance, approximately 0.017 seconds were required to make a prediction for a single patient. We measured the time by averaging the time that the model took to make predictions on the entire test set using the same machine as described in the Implementation Details section.

Analysis of the Risk Score

We analyzed the risk score generated by the RNN model with basic characteristics of the patients to understand how the risk score is affected by patient characteristics. Age and historical visit count were selected as baselines since they were expected to serve as proxies of a patient's general health status. We used the best performing model among the five models (RNN with pre-trained embedding) in 5-fold cross validation to obtain the risk score of the patients in the test set of the corresponding fold. **Figure 3** shows the scatterplot between the risk score and (**a**) outcome status (**b**) age (min-max normalized age), and (**c**) historical visit count of the patients. The regression coefficient were +0.136 ($p < 0.01$) in (**a**), +0.376 ($p < 0.01$) in (**b**), and +0.0002 ($p < 0.01$) in (**c**). **Figure 4** shows the ROC curve of the risk score, age, and historical visit count in predicting the outcome status of the patients.

Visualization of Patients

The output vector of the fully connected layer in the RNN model is expected to contain information about the patient that is necessary for predicting the risk of developing severe outcomes from COVID-19. We analyzed the patients by visualizing the output vectors of patients on 2-dimensional space using uniform manifold approximation and projection (UMAP)[25]. We trained the RNN model with pre-trained embedding on the entire dataset for 30 epochs and generated output vectors for patients by using the trained model on the entire data. **Figure 5a** shows the scatterplot of the output vectors of all patients in the dataset. **Figure 5b** and **5c** shows the scatterplot of the output vectors of severe patients color labeled by sex (**5b**) and age (**5c**). To further explore the pattern of output vectors for severe patients, we color-labeled them on the 2-dimensional space based on common comorbidities of the cohort. Two common comorbidities of COVID-19 patients in CUIMC, renal failure and type 2 diabetes mellitus (T2DM), were selected[26]. **Figure 6a** and **6b** shows scatterplots of the output vectors of severe patients color-labeled based on the observation of T2DM and renal failure respectively. Scatterplots of the output vectors of male and female severe patients separately color-labeled based on the observation of T2DM and renal failure are shown in **Figure 6c-6f**.

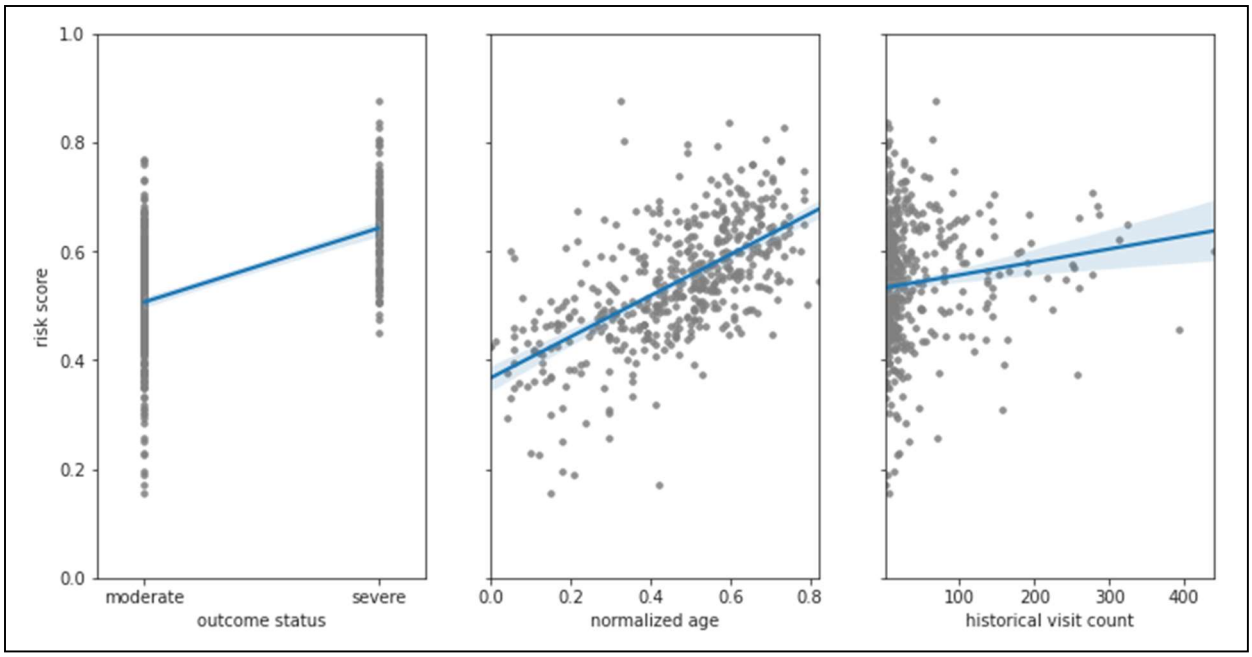


Figure 3. Scatterplot of (**a**) the outcome status and the risk score, (**b**) normalized age and the risk score, and (**c**) historical visit count and the risk score with the regression line. The gray-colored dots represent patients and shaded region around the regression line represents confidence interval.

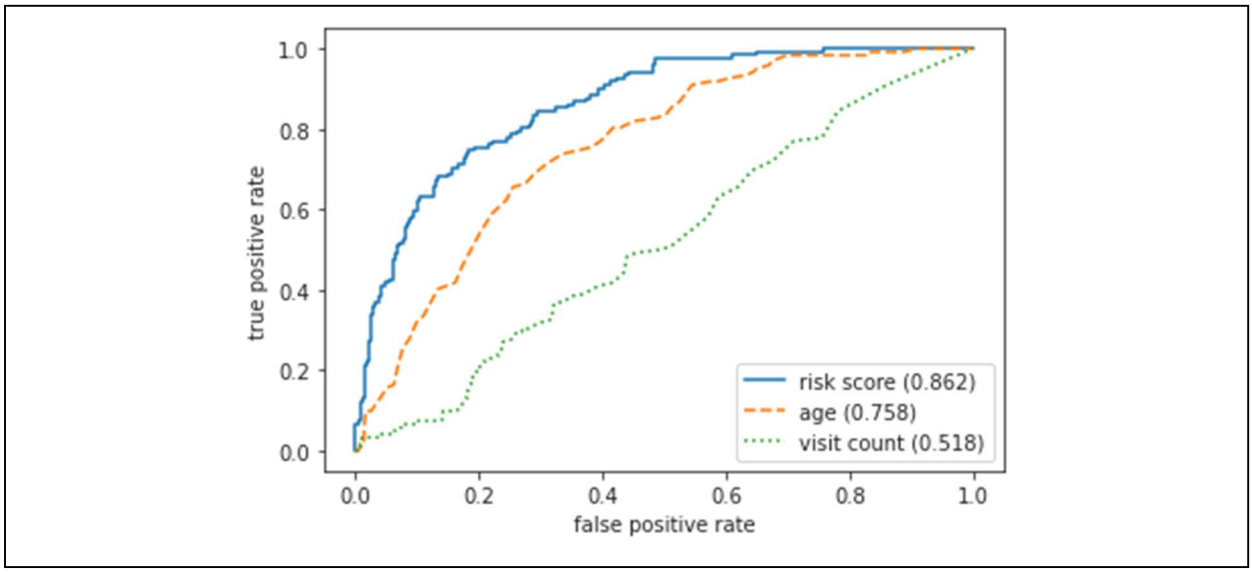


Figure 4. Receiver operating characteristic (ROC) curve of the risk score, age, and historical visit count in predicting the outcome status of the patients. Area under each ROC curve is denoted in the legend.

DISCUSSIONS

In this study, we proposed an RNN model to predict the risk of developing severe outcomes for COVID-19 patients by utilizing historical EHR data of the patients. The best average AUC was achieved by the RNN model with pre-trained embedding. However, it is worth noting that the difference between average AUC of the RNN model and other baselines are not significant considering the standard error although simple paired t-test confirmed statistically meaningful the difference between the average AUC of the RNN model and other baselines in each initialization scheme. Relatively high standard error is perhaps due to the small size of the dataset. We also found that using randomly initialized embedding in logistic regression and MLP underperforms the models using multi-hot representation as input while using pre-trained embedding improved the performance in all models. This is perhaps because the aggregation across patients' visits causes information loss for the randomly initialized embedding and the data set was not sufficiently sized to allow the embedding layer to be properly trained starting from random initialization. Additionally, the pre-trained embedding may be suboptimal because we only used the data from the COVID-19 cohort to pre-train the embedding. We expect that the performance will further improve if we use a larger data set to pre-train the embedding.

Although we used a relatively large data set compared to existing COVID-19 studies, which mostly have a few hundred cases[7], the 2,374 cases in our data set is still considered very small for training deep neural network models that contain a large number of parameters to learn. While the model will be able to learn better with more data, obtaining a large data set, however, is not easy for a single institution due to the limited number of patients (and we certainly hope the number of COVID19 patients will not further increase in our institution). We believe obtaining a larger size of data across different institutions and nations or using other disease cohorts as proxy cohorts will resolve this limitation. One advantage of our approach is that our analysis used a standardized clinical data format, the OMOP Common Data Model. The source code for this analysis can be easily shared with others who have similarly formatted clinical data for evidence aggregation.

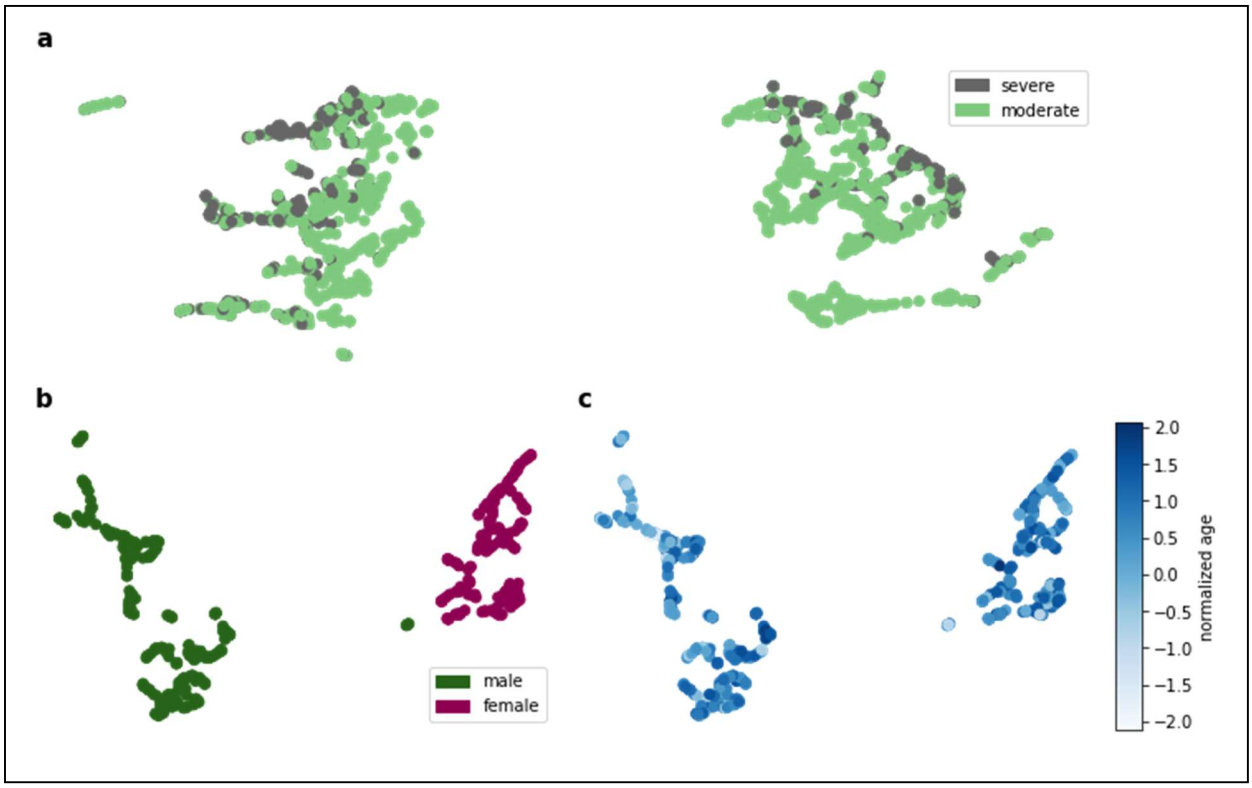


Figure 5. Scatterplots of the output vectors of patients in the COVID-19 cohort. All patients are shown in (**a)** with color representing severity status. Only severe patients are shown in (**b)** and (**c)**, with color representing sex in (**b)** and normalized (i.e. normalized with mean and standard deviation) age in (**c)**.

While higher accuracies (0.73-0.99) were reported in other studies, the intended use of these models were often not clearly described[7]. The RNN model we propose is intended to aid decision making at the time of or before hospital admission due to COVID-19, since only historical EHR data were needed in the model. In addition, the RNN model can be applied to the general population that is not confirmed COVID-19 positive to identify people at high risk of developing potential severe outcomes if infected by COVID-19. The RNN model can readily be applied to the situations above with much larger datasets or in a real-time setting since it can compute the risk score of the patient in a small amount of time.

We demonstrated the effectiveness of the risk score predicted by the RNN model by analyzing it with basic characteristics (i.e., age and total historical visit count) of the patients. From **Figure 3a,** we can confirm that the risk score is correlated with the patient developing severe outcomes from COVID-19. We found that there exists a statistically significant positive relationship between age and the risk score of the patients in **Figure 3b**, which indicates that age itself is an important factor to predict the outcome status of patients. We also expected that the number of hospital visits in a patient's medical history would reflect the patient's general health status and therefore a positive relationship would exist between the total historical visit count and the risk score. **Figure 3c** shows, however, that the relationship between the historical visit count and the risk score of the patients is not strong. The risk score predicted by the RNN model outperforms the other two baselines in predicting outcome status of the patients as shown in **Figure 4**.

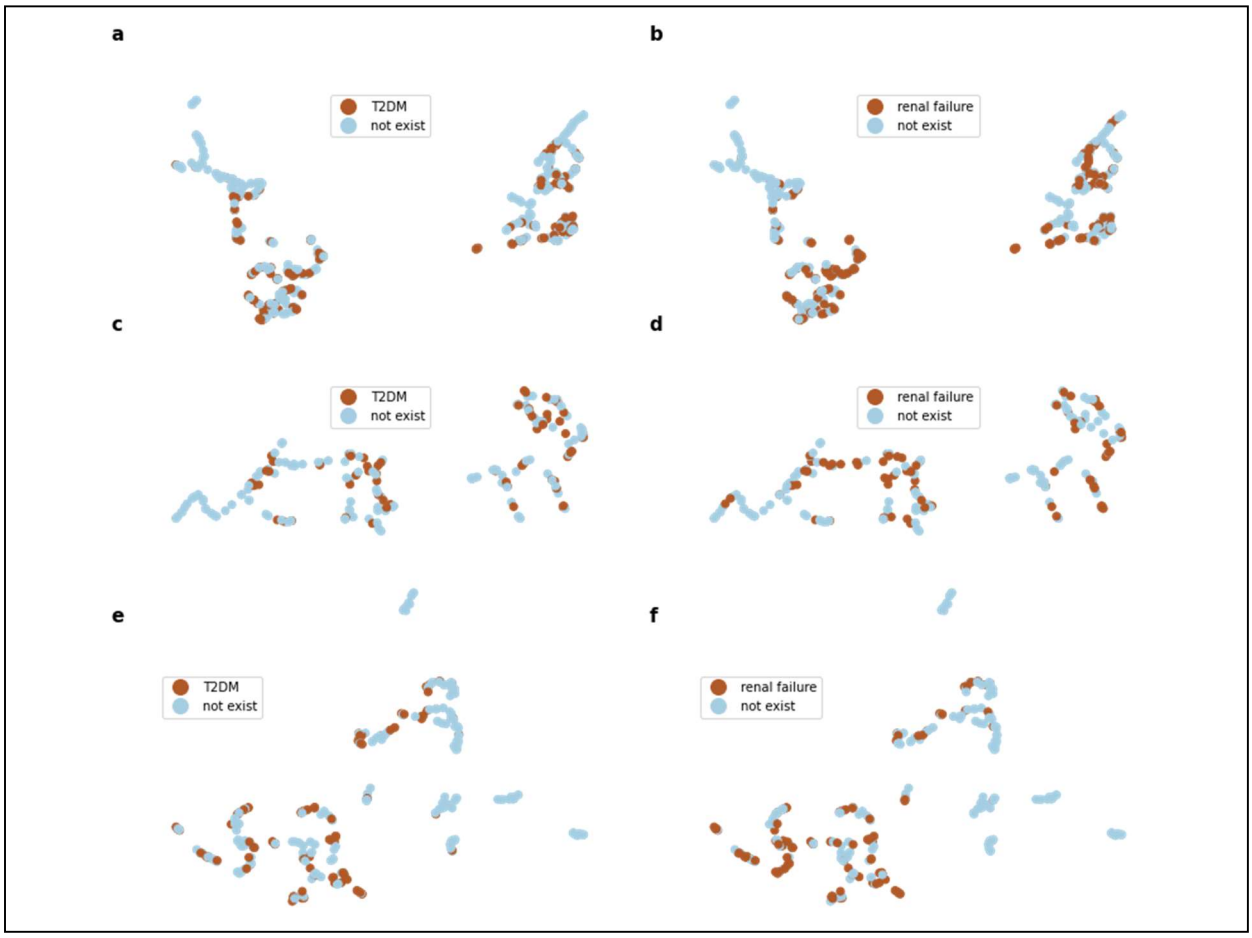


Figure 6. Scatterplots of the output vectors of severe COVID-19 patients, with color representing the observation of (**a**) type 2 diabetes mellitus (T2DM) and (**b**) renal failure. (**c**) and (**d**) are scatterplots of the output vectors of male severe COVID-19 patients, with color representing the observation of T2DM and renal failure respectively. (**e**) and (**f**) are scatterplots of the output vectors of female severe COVID-19 patients, with color representing the observation of T2DM and renal failure respectively.

From **Figure 5a**, we can see visible clusters of the severe COVID-19 patients. Male and female severe patients were divided into two clusters in **Figure 5b**. Age, however, does not show clearly distinguishable patterns in the clusters from **Figure 5c**. While we cannot confirm clear clusters based on the existence of T2DM or renal failure, we can see that the patients separate into distinct clusters throughout **Figure 6**. Since the patient vectors were generated based on the patients' observed conditions across visits, these clusters could reflect common comorbidities among severe COVID-19 patients. Additionally, the presence of visible clusters within the scatterplots of the male and female severe patient groups suggests that there exist multiple subgroups of severe COVID-19 patients with distinct characteristics, which shows the potential possibility of subtyping COVID-19 patients. We believe that further efforts to uncover detailed characteristics of the clusters are warranted for subtyping COVID-19 patients.

A drawback of the RNN model is that the model lacks interpretability. The model interpretability is critically important for the model utilizes medical data since interpretable model output can deliver new insights to the problem. For example, we can compare the impact of individual concept on developing severe outcome of COVID-19 by analyzing the weights in logistic regression model with multi-hot vector input. Although the RNN model showed better performance than other models, this gain is at the cost of interpretability. We would like to address this limitation by developing interpretable model without compromising on accuracy in the future study.

Our study shared some common limitations with the existing predictive models for COVID-19 patients. Wynants et al. performed a review of existing predictive models for COVID-19 patients and reported that most of the models have high risk of bias when evaluated with PROBAST (prediction model risk of bias assessment tool)[7,27]. They found that two common causes of risk of bias in predictive models for COVID-19 were lack of external validation and selection bias. Since the COVID-19 cohort in this study includes patients whose clinical course of care has not yet completed and who may still potentially develop a severe outcome, there is a chance that discharged patients without any signal of severe status during hospitalization at NYP/CUIMC will later develop a severe outcome outside of NYP/CUIMC. Future work will include developing an RNN model to predict various states of a patient being infected with COVID-19 rather than simply predicting the risk score. We also plan to modify the RNN model for time-to-event analysis to appropriately handle censored data.

Additionally, the model was not validated with an external cohort. This limitation is mainly caused by medical data exchange issues across different medical institutes, which limits the sharing of medical data across institutions. Since the RNN model is based on a dataset implemented with OMOP common data model, we expect that applying the model to another institution using the common data model will be easily conducted. For example, Burn et al., has performed deep phenotyping on more than 30,000 patients hospitalized with COVID-19 patients in Asian, Europe and American countries using OHDSI network dataset[28]. Future work includes experimenting with and validating the RNN model across different institutions in various countries using the OHDSI network dataset.

CONCLUSION

We proposed a predictive model using recurrent neural networks to predict the risk of developing severe outcomes for COVID-19 patients. The proposed RNN model outperforms logistic regression and multi-layer perceptron models in predicting severe outcome status of COVID-19 patients. We also demonstrated the effectiveness of the risk score by analyzing the risk score generated by the RNN model with the basic characteristics of the patients. Future work includes experimenting with the model with a larger dataset and validating the model with an external dataset, adding interpretability to the model, as well as further improving the RNN model using more concepts from other domains (e.g., drug, measurements, and procedure) and using time-to-event analysis, which also can address the censored patient issue.

ACKNOWLEDGEMENT

This work was supported by The National Library of Medicine grant R01LM012895-03S1 and The National Center for Advancing Translational Science grant 1OT2TR003434-01. Supplementary material is available at https://github.com/Jayaos/rnn-covid. The authors thank the anonymous reviewers for their valuable feedback that helped us to make significant improvements to this study.

References

1. Zhang H, Shi T, Wu X, et al. Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. 2020.
2. Lu J, Hu S, Fan R, et al. ACP risk grade: a simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China. 2020.
3. Xie J, Hungerford D, Chen H, et al. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. 2020.
4. Liang W, Yao J, Chen A, et al. Early triage of critically ill COVID-19 patients using deep learning. *Nature communications.* 2020;11(1):1-7.
5. Huang H, Cai S, Li Y, et al. Prognostic factors for COVID-19 pneumonia progression to severe symptom based on the earlier clinical features: a retrospective analysis. *medRxiv.* 2020.
6. Carr E, Bendayan R, Bean D, et al. Supplementing the National Early Warning Score (NEWS2) for anticipating early deterioration among patients with COVID-19 infection. *medRxiv.* 2020.
7. Wynants L, Van Calster B, Bonten MM, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *bmj.* 2020;369.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. *nature.* 2015;521(7553):436-444.
9. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:151103677.* 2015.
10. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. Paper presented at: Machine Learning for Healthcare Conference2016.
11. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association.* 2017;24(2):361-370.
12. World Health Organization. COVID-19 Therapeutic Trial Synopsis. https://www.who.int/publications/i/item/covid-19-therapeutic-trial-synopsis. Accessed August 26 2020.
13. Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:14061078.* 2014.
14. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation.* 1997;9(8):1735-1780.
15. Britz D, Goldie A, Luong M-T, Le Q. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:170303906.* 2017.
16. Weiss G, Goldberg Y, Yahav E. On the practical computational power of finite precision RNNs for language recognition. *arXiv preprint arXiv:180504908.* 2018.
17. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:14123555.* 2014.
18. Kocmi T, Bojar O. An exploration of word embedding initialization in deep-learning tasks. *arXiv preprint arXiv:171109160.* 2017.
19. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. Paper presented at: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)2014.
20. Lee J, Liu C, Kim JH, et al. Comparative Effectiveness of Medical Concept Embedding for Feature Engineering in Phenotyping. *medRxiv.* 2020:2020.2007.2014.20151274.
21. Masters D, Luschi C. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:180407612.* 2018.
22. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research.* 2014;15(1):1929-1958.
23. Abadi M, Barham P, Chen J, et al. Tensorflow: A system for large-scale machine learning. Paper presented at: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)2016.
24. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980.* 2014.
25. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426.* 2018.
26. Argenziano MG, Bruce SL, Slater CL, et al. Characterization and clinical course of 1000 patients with coronavirus disease 2019 in New York: retrospective case series. *bmj.* 2020;369.
27. Wolff RF, Moons KG, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine.* 2019;170(1):51-58.
28. Burn E, You SC, Sena A, et al. Deep phenotyping of 34,128 patients hospitalised with COVID-19 and a comparison with 81,596 influenza patients in America, Europe and Asia: an international network study. *medRxiv.* 2020.

Privacy-preserving Sequential Pattern Mining in distributed EHRs for Predicting Cardiovascular Disease

Eric W. Lee, MS[1], Li Xiong, PhD[1], Vicki Stover Hertzberg, PhD[2], Roy L. Simpson, RN, DNP[2], Joyce C. Ho, PhD[1]
[1]**Department of Computer Science, Emory University, Atlanta, GA**
[2]**Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, GA**

Abstract

From electronic health records (EHRs), the relationship between patients' conditions, treatments, and outcomes can be discovered and used in various healthcare research tasks such as risk prediction. In practice, EHRs can be stored in one or more data warehouses, and mining from distributed data sources becomes challenging. Another challenge arises from privacy laws because patient data cannot be used without some patient privacy guarantees. Thus, in this paper, we propose a privacy-preserving framework using sequential pattern mining in distributed data sources. Our framework extracts patterns from each source and shares patterns with other sources to discover discriminative and representative patterns that can be used for risk prediction while preserving privacy. We demonstrate our framework using a case study of predicting Cardiovascular Disease in patients with type 2 diabetes and show the effectiveness of our framework with several sources and by applying differential privacy mechanisms.

Introduction

The rapid growth of electronic health records (EHRs) provides rich information about patients' conditions, treatments, and outcomes[1]. EHRs have been widely used in various healthcare researches such as risk predictions[2,3] and phenotyping[4,5]. Yet, an often overlooked aspect of mining EHRs is the temporal nature of the data. As the data contains the previous and current status of patients, EHRs can be viewed as a sequential database chronologically ordered by date and time. Thus, sequential pattern mining (SPM) can be applied to EHRs to discover interesting, useful, and unexpected patterns that can be used by a predictive model to forecast the patient's future disease status from the current and previous patient conditions.

Existing works performing SPM of EHRs have demonstrated its potential predictive power. For example, Wright *et al*.[6] mined sequential patterns of diabetes medication prescriptions to predict the next medication to be prescribed. Ghosh *et al*.[7] proposed to apply SPM to stream bed monitors in ICUs for predicting acute hypotension in critical care patients. Lee and Ho[2] used a sequence of diagnosis codes in clinical records to predict chronic heart failure using SPM. However, the existing SPM-based methods assume all the EHRs are stored in a central repository or database. In practice, each healthcare system may have one or more clinical data warehouses to store all the patient data. Thus, one of the key challenges towards developing robust models that generalize across multiple systems is to mine data that are distributed across multiple locations or sources.

While mining distributed data itself can be challenging, a further complication arises from privacy laws. A key challenge related to large-scale analysis of EHRs is that the privacy of human subjects should be protected. Therefore, patient data in its original form cannot be transmitted without privacy guarantees that protect against some privacy attacks such as learning information about individuals from data release. To alleviate this issue, differential privacy[8] (DP) has emerged as one of the strongest privacy guarantees for statistical data release of sources such as EHRs.

In this paper, we propose a privacy-preserving SPM-based framework for mining EHRs across multiple sources. Unlike existing SPM methods that work only for EHRs stored at a central (single) source, we propose to extract discriminative or representative patterns separately at each source, share the patterns in a DP-preserving manner to a centralized location, and use the patterns for future risk prediction. A major benefit of our framework is that it can guarantee patient privacy for each source separately and still achieve approximately the same overall predictive performance of the model as the central model. We demonstrate our framework using a case study of predicting cardiovascular disease in patients with type 2 diabetes. The experimental results illustrate the effectiveness and flexibility of our framework using different DP mechanisms with several sources.

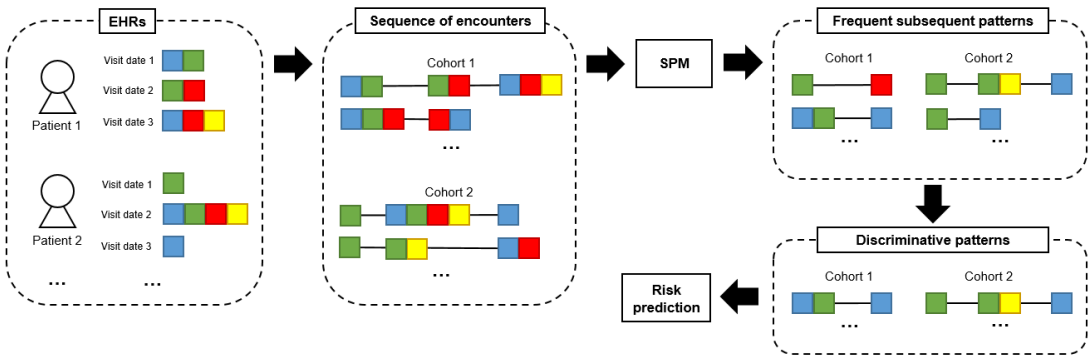


Figure 1: An illustration of the SPM-based framework in risk prediction. Each colored box in the figure denotes a single diagnosis code. There exist only one sequence of encounters for each patient. Note that for both frequent subsequent patterns and discriminative patterns, not all the patterns are shown in the figure.

A Case Study of Cardiovascular Disease in Patients with Type 2 Diabetes. Approximately 8.2% of the US population suffered from diabetes in 2018[9]. Moreover, diabetes can lead to other health complications and often results in heavy economic burden[10]. One common complication for patients with diabetes is cardiovascular disease (CVD) which refers to a number of heart-related conditions including heart disease, stroke, and heart failure. CVD incurs heavy health and economic burdens with a projected medical cost of $358 billion in 2015[11]. There is also a strong correlation between diabetes and CVD. The mortality risk of CVD among people with diabetes is high; 65% (age 65 or older) die because of heart disease and 16% die of stroke[12]. As the healthcare expenditure and resources are high in patients with diabetes and CVD, early intervention in CVD patients can lead to favorable health outcomes[13]. Thus, we demonstrate our method to predict whether a patient with diabetes will develop CVD in the future.

Background

Sequential Pattern Mining. In the field of data mining, pattern mining is broadly used to discover interesting, useful, and unexpected patterns in the database[14]. When the ordering of the events is important, SPM is proposed as a prominent solution. The goal of SPM is to find the set of all frequent subsequent patterns in the sequence database that satisfies a user-specified threshold θ. Here, θ indicates the frequency of the subsequent pattern (also known as the support count) that appears in the database.

SPM can be applied to EHRs as it contains sequences of medical evidences and actions. As an example, in each patient's visit (or encounter), ICD-9 or ICD-10 diagnosis codes are recorded. A sequence of encounters can be represented by listing each patient's encounters in chronological order by the visit dates. Once, the sequence of encounters is constructed, any SPM algorithm can be applied to extract frequent subsequent patterns.

SPM-based Framework in Risk Prediction. To mine useful patterns that can be used for risk prediction, it is often helpful to find representative patterns that distinguish patients exposed to the disease (i.e., cases) from patients not exposed to the disease (i.e., controls). Lee and Ho[2] refer to these representative patterns as *discriminative patterns*. In other words, a discriminative pattern is one that appears in one cohort but not in the other. These patterns can then be used as a feature representation for risk prediction. Figure 1 illustrates the process.

To obtain discriminative patterns, Lee and Ho[2] proposed the application of SPM to extract all frequent subsequent patterns from the sequence of encounters of each cohort that satisfies the user-specified support count. A lower support count is used to extract more patterns to improve patient representation. However, the extracted patterns may be common patterns that exist in the other cohort. To discard these common patterns, patterns existing in the other cohort are filtered out to obtain *discriminative patterns*. However, this may be too restrictive as many of the discovered patterns exist in both cohorts[2]. Thus, a threshold, τ, was proposed to allow some patterns to exist in the other cohort but require a higher support count that satisfies τ. Suppose the frequent subsequent pattern p in cohort c_1 has a support count of 10. If $\tau = 2$ and p has a support count of 5 in the other cohort c_2, then p is a discriminative pattern for cohort c_1. This process is done for both cohorts to obtain discriminative patterns for each cohort.

Differential Privacy. To protect the privacy of a human subject, Differential Privacy[8] (DP) was proposed. Under DP, the main goal is to learn useful information from the EHRs while nothing is learned about the patient. In other words, although any patient's record is arbitrarily changed, the output of an algorithm should be approximately the same. The formal definition of DP is as follows.

Definition 1. (ϵ-Differential privacy) *A privacy algorithm K satisfies ϵ-differential privacy if and only if for all neighboring databases D and D' differing on at most one record and for any possible output $S \subseteq Range(K)$,*

$$Pr[K(D) \subseteq S] \leq exp(\epsilon) \times Pr[K(D') \subseteq S] \tag{1}$$

From the Equation (1), ϵ is a privacy budget which is a metric to determine how strict the privacy is. A smaller value of ϵoffers better privacy protection. Common DP mechanisms to achieve ϵ differential privacy is the Laplace mechanism[8] and Exponential mechanism[15]. Laplace mechanism achieves ϵ differential privacy by adding a random noise sampled from the Laplace distribution to a statistical measure, and the Exponential mechanism uses exponential distribution for adding noise.

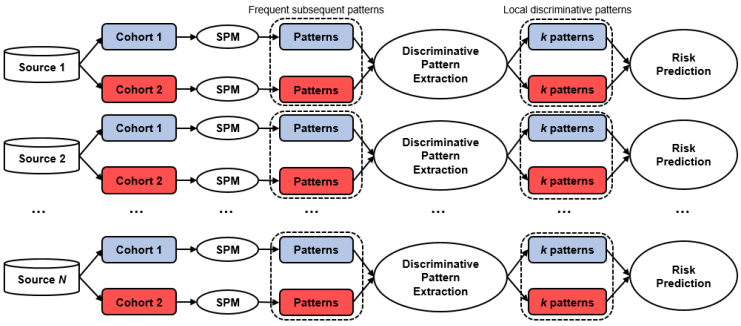
Differential Privacy in Multiple Distributed Sources. Many DP mechanisms are proposed for a traditional centralized source where all data is located in a single source[8,15,16]. There are cases when the data is distributed across multiple sources (decentralized sources). For such cases, federated learning is used to train the model across multiple decentralized sources without the sharing of raw data[17]. And for sensitive data such as EHRs, privacy is an important issue when using federated learning. Thus, many frameworks are proposed for federated learning with DP[18–20] which collaboratively train the model while preserving privacy. For example, Truex *et al*.[19] proposed to combine DP and secure multiparty computation (SMC) in a federated learning system to address the risk inference during the model learning process, and Choudhury *et al*.[20] proposed to apply DP in distributed EHRs for prediction of adverse drug reaction and mortality rate. However, for the SPM-based federated learning framework, we propose to apply DP mechanisms to support counts of subsequent patterns in multiple decentralized data sources for privacy guarantee.

Methods

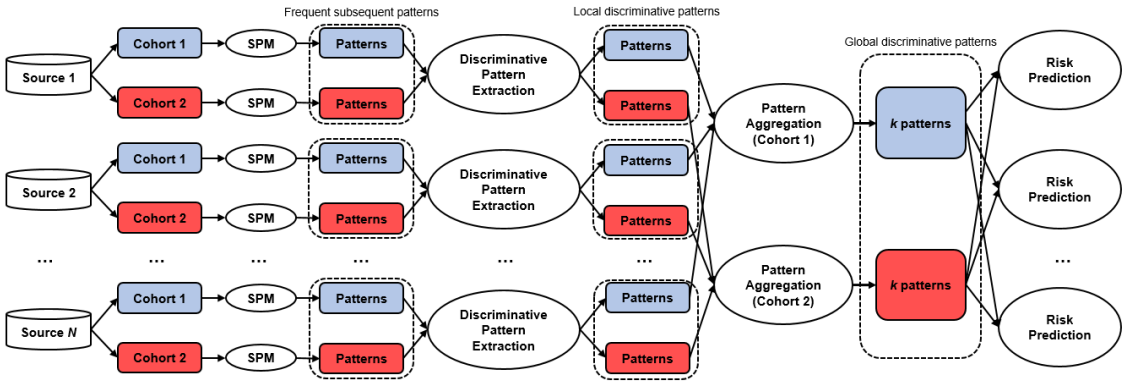
In this section, we propose our SPM-based framework for distributed EHRs. We first show a framework that trains the model individually for each data source (no aggregation framework), then we propose a framework that collaboratively trains the model by sharing the patterns from each data source while preserving the privacy (aggregation framework).

SPM-based Framework with No Aggregation. For the SPM-based framework with no aggregation, we introduce a federated learning-based technique that does not share any patient information across different data sources. Given distributed data sources, we apply the SPM-based framework and train the risk prediction classifier for each source separately. It is important to note that we call the obtained discriminative patterns as "local" discriminative patterns as it is obtained within a single data source. Once the local discriminative patterns are extracted from each cohort, we construct a feature representation based on these patterns. From the obtained patterns, we use the top k discriminative patterns based on their support counts from each cohort and use it as our feature representation for risk prediction. We use the existence of the pattern as a feature for the machine learning classification model. Suppose the top 2 discriminative patterns are extracted, p_1 and p_2 from cohort c_1, and p_3 and p_4 for cohort c_2. We construct a feature representation as $[p_1, p_2, p_3, p_4]$, and set the feature to be the existence of the pattern. For example, if one patient contains patterns p_1 and p_4, then the feature of this patient will be $[1, 0, 0, 1]$. This representation is depicted in Figure 2(a). As illustrated in the figure, all processing is performed "locally" for each data source, and none of the pattern information is shared between other data sources. In addition, each classifier is trained locally using only the patients from the same data source. One limitation of this framework is that discriminative patterns from other data sources are not used which can be more important. The disease prevalence can vary depending on location and populations, thus, to learn a stronger classifier, it may be necessary to use information from other data sources.

Privacy-preserving SPM-based Framework with Aggregation. To alleviate the limitation, we propose a SPM-based framework with aggregation that uses a centralized server to aggregate local discriminative patterns from all data sources and selects the top k global discriminative patterns which are used to train each classifier for the data source. Similar to the SPM-based framework with no aggregation, we apply SPM to each data source, extract frequent subsequent patterns, and obtain local discriminative patterns for each cohort. However, unlike the SPM-based framework



(a) SPM-based framework with no aggregation



(b) SPM-based framework with aggregation

Figure 2: A framework overview of SPM-based frameworks without and with aggregation. The blue box denotes patterns associated with cohort 1, and the red box denotes patterns associated with cohort 2.

with no aggregation, we use "global" discriminative patterns as a feature representation instead of "local" patterns. We aggregate all local discriminative patterns from each data source into a single list. In other words, there will be two lists of global discriminative patterns for each cohort (*i.e.*, one for the case and one for control). When aggregating the patterns from a centralized location for each cohort, we take the union of patterns and do the summation of the support counts of each pattern. From this aggregated list of patterns, we select the top k discriminative patterns and use them as the feature representation. These top k patterns are then shared back to the original source to train classifiers of each source. It is important to note that although all the local classifiers share the same global discriminative patterns, only the patients from one data source are used to train each classifier. The illustration of the framework is shown in Figure 2(b).

Unlike the SPM-based framework with no aggregation, this framework aggregates local discriminative patterns from all data sources into a centralized server. As patient information is shared across data sources, privacy protection becomes necessary. Therefore, once the frequent subsequent patterns are extracted from each cohort, we apply a DP mechanism such as Laplace mechanism[8] or Exponential mechanism[15] to the support counts of each pattern. The noisy support counts will be used to extract local discriminative patterns from each data source and also used when aggregating local patterns from all data sources into a centralized server. This may lead each cohort to extract a different set of local discriminative patterns compared to the framework without DP because it will make some patterns that were not frequent to become frequent and vice versa under a specific threshold. In this way, although the framework is using global discriminative patterns, privacy can be preserved by not sharing the exact patient support count information from the data source. Later, we demonstrate that the results of applying DP mechanisms will only cause a marginal decrease in predictive performance compared to the results without applying DP.

Experiment Settings

Dataset. We use Project NeLL™(Nursing electronic Learning Laboratory), a database that contains de-identified electronic health records from more than 1 million patients seen at Emory Healthcare from 2012 to 2018. It contains over 8 million unique records, including structured text (*e.g.*, lab values) and unstructured text (e.g., clinical notes, radiology reports). Patients with type 2 diabetes are identified using the ICD-9 code of '250.*' or the ICD-10 code of 'E11.*'. Note that only patients with the admitting, discharge, or final diagnosis of type 2 diabetes are used, thus ensuring that these patients are more likely to suffer from the disease. These patients who then develop cardiovascular disease (CVD) are identified using the ICD-9 codes of '428.*' or '414.*' or the ICD-10 codes of 'I50.*' or 'I25.*' (those relating to chronic heart failure and coronary heart disease). Only patients who developed CVD after diabetes are considered (i.e., any patient that had pre-existing CVD prior to diabetes is not considered in our cohort) and only the ICD-9 codes before the CVD is recorded are used. Also, any patients who have only one visit are excluded as there aren't sufficient events to model.

From the patient records, we use demographic variables such as gender, age, and race. Each encounter of patients is listed chronologically based on their visit dates. For the purpose of this study, we focus only on the discharge diagnosis codes associated with each encounter. Instead of fine-grained ICD-9 codes, Clinical Classifications Software (CCS) codes[21], a categorization scheme for the International Classification of Diseases, is used to group ICD-9 into broader categories to yield better interpretability of the patterns. For each visit, there can be multiple CCS codes. Moreover, each patient has a different number of visits with the length of the sequence of encounters varying from 2 to 546 with an average sequence length of 12.65.

Case-Control Cohort Study. Given the imbalanced ratio of CVD patients to non-CVD patients, we designed a case-control study to identify useful sequences of events. Without this process, the extracted patterns will be dominated by non-CVD patients. Thus, we matched non-CVD patients to CVD patients in a ratio of $4:1$. Patients are grouped based on the age when diabetes was first diagnosed as well as their ethnicity. The top 4 nearest patients based on Euclidean distance from the non-CVD patients are then matched to the CVD patients such that each non-CVD patient is from the same race and of a similar age as a CVD patient. Therefore there will be at most 4 non-CVD patients for every CVD patient. The resulting dataset contains 2,112 patients with CVD and 10,464 non-CVD patients, representing 34% of patients from the original dataset.

Experimental Design. To construct the sequences of encounters, we only consider encounters after the date of diabetes were developed. We also adopt the FuzzyGap sequence representation[2] to construct the sequence of encounters. This representation is constructed by setting a user-specified boundary range – encounters within the boundary between two intervals will be added to both intervals. We set the interval to 1 month, where encounters within the same month are recorded into a single encounter and 12 days for the boundary range. We note that FuzzyGap also captures gap-sensitive frequent patterns such as $[\{\text{CCS codes}\}, \{\}, \{\text{CCS codes}\}]$ which allows an empty encounter between two encounters each with a set of CCS codes. After processing the sequence of encounters to be in FuzzyGap sequence representation, we end up having 2,112 CVD patients and 7,998 non-CVD patients after excluding patients who have only one interval.

To evaluate the efficiency of our framework, we split the dataset into several partitions (or data sources). We explore 4 different partition settings: 1, 2, 4, and 8. Partition setting with 1 represents the single data source. For simplicity, we will denote these settings as *n=1*, *n=2*, *n=4*, and *n=8* for the partition settings 1, 2, 4, and 8, respectively. For each partition setting (*e.g.*, *n=4*), we evaluate 5 different random partitions. And for every random partition, there is a train-test split with a ratio of 70% and 30% respectively, and this is done 3 times by randomly selecting patients for train-test splitting. In total, we are running 15 experiments for each partition settings. For every train-test split, we have 7,078 patients in the train set and 3,032 patients in the test set.

While there are several fast and memory-efficient SPM algorithms such as FAST[22], CM-SPADE[23], and CloFast[24], our preliminary experiments using these algorithms implemented in the SPMF library[25] ran out of memory or only could obtain patterns with high support count (on a machine with 100GB of RAM). Thus, we discovered patterns by performing a sequential pairwise comparison between two patients as used in FuzzyGap[2]. As Lee and Ho[2] discussed previously, the predictive power is similar between other SPM algorithms and pairwise comparison.

To select the best top k, we first evaluate various top k settings (from 40 to 700) in the *n=1* setting without any DP mechanisms. Note that top k means k discriminative patterns from each class, hence, $2 \times k$ patterns are used as the feature representation. For extracting the discriminative patterns, we use the filtering threshold, $\tau = 2$, which allows some patterns to exist in the other cohort. Once we select the best top k from *n=1*, we use the same top k throughout the remaining experiments.

To evaluate the impact of DP mechanisms on our SPM-based framework with aggregation, we apply three DP mechanisms, Laplace mechanism[8], Exponential mechanism[15] and SVT[16]. For consistency, the privacy budget, $\epsilon = 0.1$ is used as 0.1 is a small value for ϵ and it provides strong privacy protection.

Evaluation Metrics. We evaluate the risk prediction task using the F1 score and area under the receiver operating curve (AUC). In addition to evaluation using the predictive task, we also evaluate our framework based on the recoverability of the discriminative patterns with a single data source. We use precision and recall in the information retrieval context which is defined as below.

$$precision = \frac{|relevant \cap retrieved|}{|retrieved|} \tag{2}$$

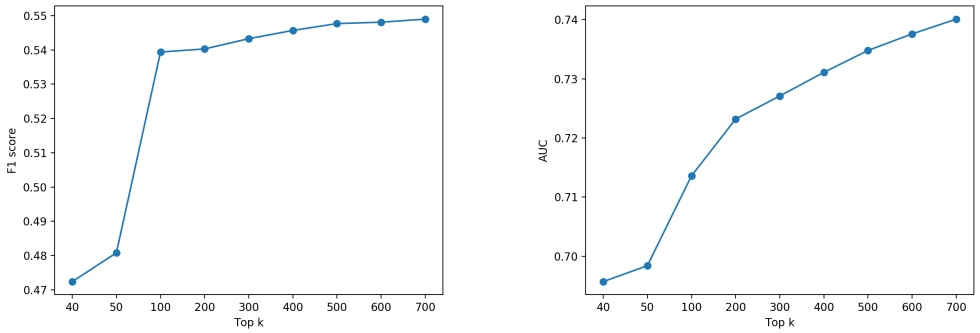
$$recall = \frac{|relevant \cap retrieved|}{|relevant|} \tag{3}$$

For both metrics, relevant refers to the discriminative patterns extracted from the single data source, and retrieved denotes the discriminative patterns extracted from the distributed sources. Only the top k discriminative patterns from each class are compared and those patterns are used as the feature representation.

Empirical Results

In this section, we use the term "*no aggregation*" for the SPM-based framework with no aggregation and no privacy, "*no privacy*" for SPM-based framework with aggregation and no privacy, and "*Laplace*", "*Exponential*", and "*SVT*" for privacy-preserving SPM-based framework with aggregation using different DP mechanisms. Note that the reported results are the average of all 3 trials.

Selecting Top k. When applying the *no aggregation* framework to *n=1* setting, many patterns are extracted. For example, in our case-control study, we discovered 924 discriminative patterns for CVD while 27,360 discriminative patterns are discovered for non-CVD patterns. Since the patterns are used for risk prediction, direct usage of all patterns may not yield desirable results due to the potential overfitting of the downstream predictive models. Thus, we use the top k discriminative patterns from each class, resulting in $2 * k$ patterns used as the feature representation.

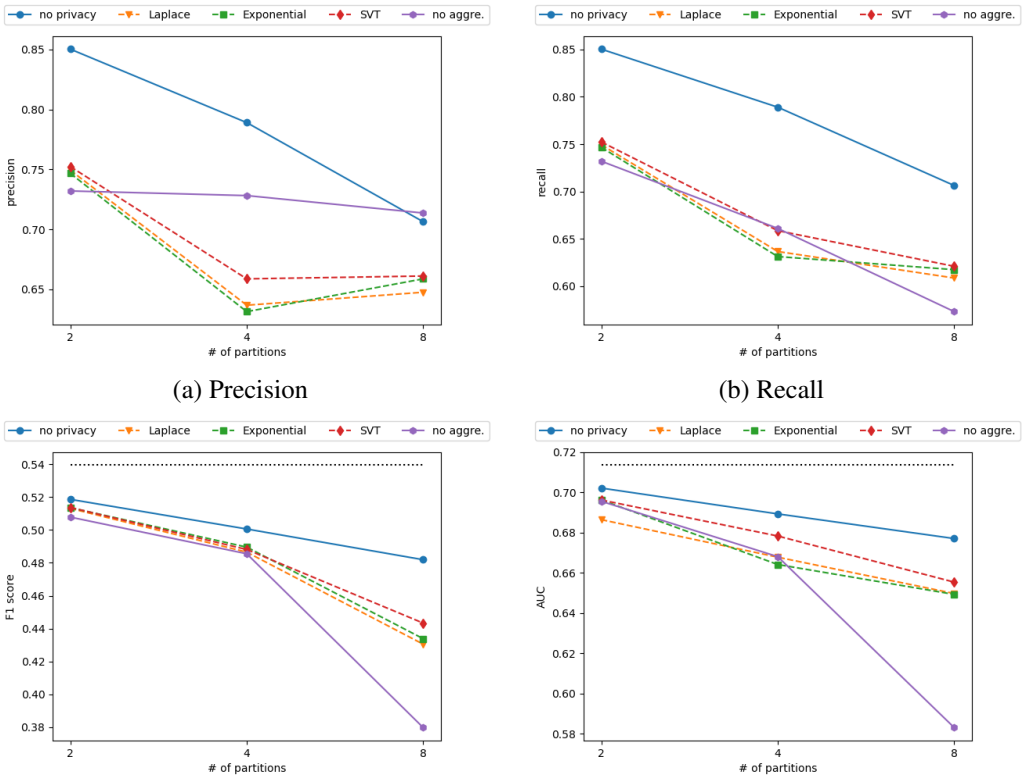


(a) F1 score (b) AUC

Figure 3: The F1 score and AUC results of risk prediction in *n=1* setting. The results are using the SPM-based framework with no aggregation with various top k values without any DP mechanisms. Note that top k means selecting top k discriminative patterns from each class, hence in total, we are using $2 * k$ patterns as the feature representation.

Table 1: The number of discriminative patterns discovered. The reported numbers are the average of partition settings out of 5 different random partitions with 3 different train-test split.

Model	2 partitions		4 partitions		8 partitions	
	CVD patterns	non-CVD patterns	CVD patterns	non-CVD patterns	CVD patterns	non-CVD patterns
No aggregation	100	100	85	100	29	100
No privacy	100	100	100	100	100	100
Laplace	100	100	98	100	82	100
Exponential	100	100	98	100	86	100
SVT	100	100	100	100	90	100



(a) Precision (b) Recall

(c) F1 score (d) AUC

Figure 4: The results show the impact of applying various DP mechanisms in various partition settings. The results are reported on the average of 5 trials of different partition settings, and 3 different train-test split. For recoverability, precision and recall are used and the results are compared with *n=1* setting (w/o DP) to check the percentage of the discriminative patterns being found from each partition. F1 score and AUC is used to evaluate the predictive power. The black dotted line in (c) and (d) are the results of *n=1* setting. All the results are using $k = 100$.

To select the best top k used throughout the remaining experiments, we evaluated k between 40 to 700. Figure 3 illustrates the results on the F1 score and AUC using various top k without any DP mechanisms on *no aggregation* framework with 1 partition. The top k in the x-axis denotes k number of discriminative patterns from each class, thus 100 means, in total, 200 patterns are used as a feature representation. From $k = 50$ to $k = 100$, both results improve rapidly while after $k = 100$, F1 score improvement becomes marginal and AUC gradually increases. This means that not all discriminative patterns are useful for risk prediction. Thus, throughout the remainder of the experiments, we fix k to be 100.

Impact of Partitions. Figure 4 summarizes the recoverability and predictive power of the resulting patterns for each

partition setting. To compute the precision and recall, we use the top 100 discriminative patterns from each partition setting (excluding *n=1*) as 'the 'retrieved" set in the Equation (2) and (3), while the top 100 discriminative patterns from *n=1* are used as "relevant".

As shown in Figure 4(a), precision stays constant for *no aggregation*. To better understand this, the number of discriminative patterns obtained by each partition setting is shown in Table 1. From the *no aggregation* row in the table, we observe that less than 100 discriminative patterns are returned when the number of partitions exceeds 2. Moreover, the number of CVD discriminative patterns decreases as the partition increases. As the "retrieved" patterns in the denominator become smaller in Equation (2), the precision should increase. However, the precision for *no aggregation* stays constant, and this indicates that the recoverability of CVD patterns is low and the precision of *no aggregation* is more related to non-CVD patterns because the number of CVD discriminative patterns decreases while the number of non-CVD discriminative patterns stays the same as the number of partition increases. On the other hand, for *no privacy*, the precision decreases as the number of partition increases, and as shown in Table 1, all the partition settings return the same number of discriminative patterns. This indicates that partitioning results in the loss of some important patterns (patterns with high support count in *n=1* setting). From Figure 4(b), we observe that the recall decreases as the number of partition increases for both *no aggregation* and *no privacy*. Similar to the precision of *no privacy*, the size of $|relevant \cap retrieved|$ is decreasing by failing to extract important patterns, thus recall decreases.

Both F1 score and AUC decrease as the number of partition increases as shown in Figures 4(c) and (d) which follow a similar trend as recall. This shows the importance of discriminative patterns and indicates that as the number of partition increases, more important discriminative patterns are being lost. This is especially true for *no aggregation*, as it is not using global discriminative patterns. The low recoverability of CVD discriminative patterns for *no aggregation* causes the predictive power to decrease. From the Figures 4(c) and (d), a larger number of partitions results in a marginal decrease in terms of predictive power for *no privacy*, and also shows that there is a marginal decrease compared to *n=1*. However, for *no aggregation*, predictive power drastically decreases as the number of partition increases, and this shows the importance of using global discriminative patterns by aggregation.

Impact of Differential Privacy. We kept the same partition settings (*i.e.*, 2, 4, 8) and evaluated the impact of various DP mechanisms. Figure 4 shows the results of applying various DP mechanisms with $\epsilon = 0.1$ and are denoted as *Laplace*, *Exponential*, and *SVT*. Figure 4(a) shows that all privacy-preserving frameworks show a similar trend but different from *no privacy*. From Table 1, we can see that the number of discriminative patterns decreases as the number of partition increases for all three DP mechanisms. And the increment of precision from *n=4* to *n=8* is a result of the number of discriminative patterns (or "retrieved" patterns) in the denominator becoming smaller in Equation (2). The decrease of precision from *n=2* to *n=4* for all DP mechanisms occurs as not all important global discriminative patterns are obtained from aggregation because the number of discriminative patterns returned is close to or equal to 100 as shown in Table 1. For recall shown in Figure 4(b), it follows a similar trend as *no privacy* which was explained previously.

The predictive performance in terms of F1 score and AUC are shown in Figure 4(c) and (d) respectively. They show that *no privacy* outperforms all other frameworks and suggests the importance of discriminative patterns. In other words, as the framework can discover more important discriminative patterns, the predictive power increases. For all DP mechanisms, they show a similar trend in predictive power which is decreasing as the number of partition increases. By comparing the results with *no privacy* and DP mechanisms, it shows that there is a trade-off by having a privacy guarantee, however, there is a less sharp loss than *no aggregation*. For the AUC of *n=4*, *no aggregation* has a higher score than *Laplace* and *Exponential*. This suggests that the framework using *Laplace* and *Exponential* mechanisms discover less important discriminative patterns which result in a lower score than *no aggregation*. In other words, by applying the *Laplace* and *Exponential* mechanisms, important discriminative patterns are discarded and discriminative patterns with low support counts are returned. And for *SVT*, although it has a similar recall with *no aggregation*, it uses 15 more CVD discriminative patterns than *no aggregation*, thus, having higher AUC. For *n=2* setting, as shown in Figure 4(d), *Exponential*, *SVT*, and *no aggregation* has similar AUC score while *Laplace* has slightly lower AUC. This again emphasizes the importance of discriminative patterns as all 4 frameworks have an equal number of discriminative patterns. The difference between all performance except *no aggregation* is marginal across the number of partitions compared to the *n=1* setting. Overall, our results suggest that our framework has minor trade-offs for preserving

privacy with predictive performance.

Discussion and Conclusions

In this paper, we propose a privacy-preserving SPM-based framework with aggregation and show the effectiveness of our method. In a large-scale analysis of EHRs, protecting patients' information is an important task, and we have shown that our privacy-preserving framework has almost similar predictive power with the framework without using DP. One limitation of the work is the usage of single real-world EHRs. As disease prevalence can vary depending on location and populations, it is important to use heterogeneous populations because single EHRs typically reflect more homogeneous populations. Another limitation is the even size of the partitions. In practice, EHRs can be stored in one or more data sources but not evenly distributed. When extracting discriminative patterns, local discriminative patterns extracted from a larger data source could dominate other local patterns, resulting in their emergence as global discriminative patterns. One possible extension is to use a weight-based aggregation to prevent one set of local discriminative patterns from dominating others. The last limitation is using only the ICD-9 codes of the patient. One possible extension is to use more information such as procedure codes or prescriptions. However, using multiple information will require more computational resources for pattern extraction. Nevertheless, our framework shows promising results, thus we leave this as future work.

In conclusion, we presented the privacy-preserving SPM-based framework with aggregation for predicting CVD risk in sequences of encounters. To demonstrate the efficiency, we compared the three frameworks (no aggregation, aggregation but no privacy, aggregation with privacy) and show the importance of discriminative patterns. Our experimental results suggest that there are minor trade-offs by applying DP mechanisms. Overall, the prediction results show the effectiveness of the framework with and without applying DP using the extracted discriminative patterns.

Acknowledgements

This work was supported by National Science Foundation awards IIS-1838200 and CNS-1952192; National Institute of Health awards 1K01LM012924-01, R01LM013323-01, and R01GM118609; and CTSA award UL1TR002378.

References

1. Bai T, Zhang S, Egleston BL, Vucetic S. Interpretable representation learning for healthcare via capturing disease progression through time. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018. p. 43–51.

2. Lee EW, Ho JC. FuzzyGap: Sequential Pattern Mining for Predicting Chronic Heart Failure in Clinical Pathways. AMIA Summits on Translational Science Proceedings. 2019;2019:222.

3. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Medical care. 2010;p. S106–S113.

4. Henderson J, He H, Malin BA, Denny JC, Kho AN, Ghosh J, et al. Phenotyping through Semi-Supervised Tensor Factorization (PSST). In: AMIA Annual Symposium Proceedings. vol. 2018. American Medical Informatics Association; 2018. p. 564.

5. Warren JL, Harlan LC, Fahey A, Virnig BA, Freeman JL, Klabunde CN, et al. Utility of the SEER-Medicare data to identify chemotherapy use. Medical care. 2002;40(8):IV–55.

6. Wright AP, Wright AT, McCoy AB, Sittig DF. The use of sequential pattern mining to predict next prescribed medications. Journal of biomedical informatics. 2015;53:73–80.

7. Ghosh S, Feng M, Nguyen H, Li J. Risk prediction for acute hypotensive patients by using gap constrained sequential contrast patterns. In: AMIA annual symposium proceedings. vol. 2014. American Medical Informatics Association; 2014. p. 1748.

8. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. Springer; 2006. p. 265–284.

9. Centers for Disease Control and Prevention. National diabetes statistics report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020.

10. Dieren Sv, Beulens JWJ, Schouw YTvd, Grobbee DE, Neal B. The global burden of diabetes and its complications: an emerging pandemic. European Journal of Cardiovascular Prevention & Rehabilitation. 2010 May;17(1_suppl):s3–s8.

11. Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, et al. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation. 2011 Mar;123(8):933–944.

12. Association AH. Cardiovascular disease & diabetes; 2017. http://www.heart.org/HEARTORG/Conditions/More/Diabetes/WhyDiabetesMatters/Cardiovascular-Disease-Diabetes_UCM_313865_Article.jsp#.WckMddOGMUE.

13. Feldman DI, Valero-Elizondo J, Salami JA, Rana JS, Ogunmoroti O, Osondu CU, et al. Favorable cardiovascular risk factor profile is associated with lower healthcare expenditure and resource utilization among adults with diabetes mellitus free of established cardiovascular disease: 2012 Medical Expenditure Panel Survey (MEPS). Atherosclerosis. 2017 Mar;258:79–83.

14. Fournier-Viger P, Lin JCW, Kiran RU, Koh YS, Thomas R. A survey of sequential pattern mining. Data Science and Pattern Recognition. 2017;1(1):54–77.

15. McSherry F, Talwar K. Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). IEEE; 2007. p. 94–103.

16. Lyu M, Su D, Li N. Understanding the sparse vector technique for differential privacy. arXiv preprint arXiv:160301699. 2016;.

17. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST). 2019;10(2):1–19.

18. Geyer RC, Klein T, Nabi M. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:171207557. 2017;.

19. Truex S, Baracaldo N, Anwar A, Steinke T, Ludwig H, Zhang R, et al. A hybrid approach to privacy-preserving federated learning. In: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security; 2019. p. 1–11.

20. Choudhury O, Gkoulalas-Divanis A, Salonidis T, Sylla I, Park Y, Hsu G, et al. Differential privacy-enabled federated learning for sensitive health data. arXiv preprint arXiv:191002578. 2019;.

21. Geraci JM, Ashton CM, Kuykendall DH, Johnson ML, Wu L. International Classification of Diseases, 9th Revision, Clinical Modification codes in discharge abstracts are poor measures of complication occurrence in medical inpatients. Medical care. 1997;p. 589–602.

22. Fournier-Viger P, Gomariz A, Campos M, Thomas R. Fast vertical mining of sequential patterns using co-occurrence information. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2014. p. 40–52.

23. Salvemini E, Fumarola F, Malerba D, Han J. Fast sequence mining based on sparse id-lists. In: International Symposium on Methodologies for Intelligent Systems. Springer; 2011. p. 316–325.

24. Fumarola F, Lanotte PF, Ceci M, Malerba D. CloFAST: closed sequential pattern mining using sparse and vertical id-lists. Knowledge and Information Systems. 2016;48(2):429–463.

25. Fournier-Viger P, Lin JCW, Gomariz A, Gueniche T, Soltani A, Deng Z, et al. The SPMF open-source data mining library version 2. In: Joint European conference on machine learning and knowledge discovery in databases. Springer; 2016. p. 36–40.

A Comparison between Human and NLP-based Annotation of Clinical Trial Eligibility Criteria Text Using The OMOP Common Data Model

Xinhang Li[1#], Hao Liu[1#], Fabrício Kury[1], Chi Yuan[1], Alex Butler[1], Yingcheng Sun[1], Anna Ostropolets[1], Hua Xu[2], Chunhua Weng[1] *(#: equal-contribution first authors)*
[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA;
[2]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, USA

Abstract

Human annotations are the established gold standard for evaluating natural language processing (NLP) methods. The goals of this study are to quantify and qualify the disagreement between human and NLP. We developed an NLP system for annotating clinical trial eligibility criteria text and constructed a manually annotated corpus, both following the OMOP Common Data Model (CDM). We analyzed the discrepancies between the human and NLP annotations and their causes (e.g., ambiguities in concept categorization and tacit decisions on inclusion of qualifiers and temporal attributes during concept annotation). This study initially reported complexities in clinical trial eligibility criteria text that complicate NLP and the limitations of the OMOP CDM. The disagreement between and human and NLP annotations may be generalizable. We discuss implications for NLP evaluation.

Introduction

Named entity recognition (NER)—the process of automatically recognizing named entities and assigning appropriate semantic categories—is a fundamental task of natural language processing (NLP) [1] and has spurred the development of biomedical NLP systems [2, 3]. Because of its importance, the evaluation of NER has been an active field of research [4]. In literature, most of the evaluation of NER research focused on comparative evaluation of performance of commonly used NER systems [5-7], proposing new evaluation metrics [8-11]. However, the inconsistencies in NER evaluations, preventing objective cross-system comparisons, are underexplored.

Biomedical terminologies, also referred as ontologies, are rich sources of biomedical domain knowledge. Therefore, an ontology can be employed to validate whether a predicted entity is correct or not. For an entity, if its exact term or its synonym(s) exist in a reference ontology, the probability of recognizing them correctly is high. Thus, the involvement of ontologies to evaluate biomedical NER tools promise to facilitate error analysis. A possible drawback of evaluating clinical NER using a single ontology is that the same concept can be phrased or categorized differently across ontologies and cause discrepancies in concept normalization. For example, the concept *Breast cancer* can be represented by a post-coordinated term *Neoplasm with "body location"* being *"breast"* in SNOMED [12] but by a pre-coordinated term *Breast Neoplasms* in MeSH [13]. In CLEF [14], the annotations were mapped to concepts in the Unified Medical Language System (UMLS) [15]. CliCR [16] — a dataset of annotated clinical case reports — also used UMLS to obtain alternative phrase forms (synonyms, abbreviations and acronyms) for any recognized entity.

A common data model harmonizes concepts across various biomedical ontologies. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [17] is such a standard common data model that unifies 81 frequently used vocabularies in biomedical domain and is adopted by the OHDSI network [18]. We ***hypothesize*** that a corpus annotated based on the OMOP CDM can minimize discrepancies in human annotation and NLP-based concept normalization during NER evaluation. With the availability of a large-scale manually annotated medical corpus conforming to the OMOP CDM, this study compared the human and NLP-based annotations. We experimented with entity boundary relaxation and categorical relaxation. This in-depth analysis identifies challenges originating from the complexities in the eligibility criteria text and the limitations in the OMOP CDM, and makes recommendations regarding leveraging ontologies or common data models to facilitate clinical NER evaluations.

Background and Related Work

Only a handful of studies have conducted comparative evaluation of the recognition and classification performance of commonly used NER systems [5-7]. A few studies proposed new evaluation metrics to more precisely appraise NER performance [8-11], or compared the existent evaluation strategies [2]. Although most of the evaluations adopt the standard quantitative metrics such as precision and recall, the processes of computing such metrics vary significantly and deep understanding of NER errors and their root causes is still lacking [6]. Some measure precision and recall at

the word or token level, while others calculate these metrics at the concept level. Some use "exact match", which requires that a candidate entity can only be counted as a correct recognition if both its text spans and its class label fully agrees with an annotated entity [2], while others count partial match. Besides, we believe a proper evaluation of a clinical NER system should also leverage domain expertise [19] due to the complexity and imparity across entity categories in clinical corpora. For example, "*HIV positive*" can be categorized as a measurement entity ("*HIV*") with a value entity ("*positive*") or a condition entity ("*HIV positive*"). In clinical NER, classification of an entity's category, such as condition, measurement, or drug, can vary depending on the language context [2]. Therefore, during evaluation, effective leverage of domain knowledge is indispensable to recognize the right concept.

The Annotated Corpus for clinical NER

In a newly published corpus [20], we manually annotated the clinical trial eligibility criteria text extracted from 1,000 randomly selected clinical trials from Clinicaltrials.gov, and named the corpus Chia. The Chia dataset reached an 81% Kappa of inter-annotator agreement, which was calculated on annotations from randomly selected 50 clinical trials.

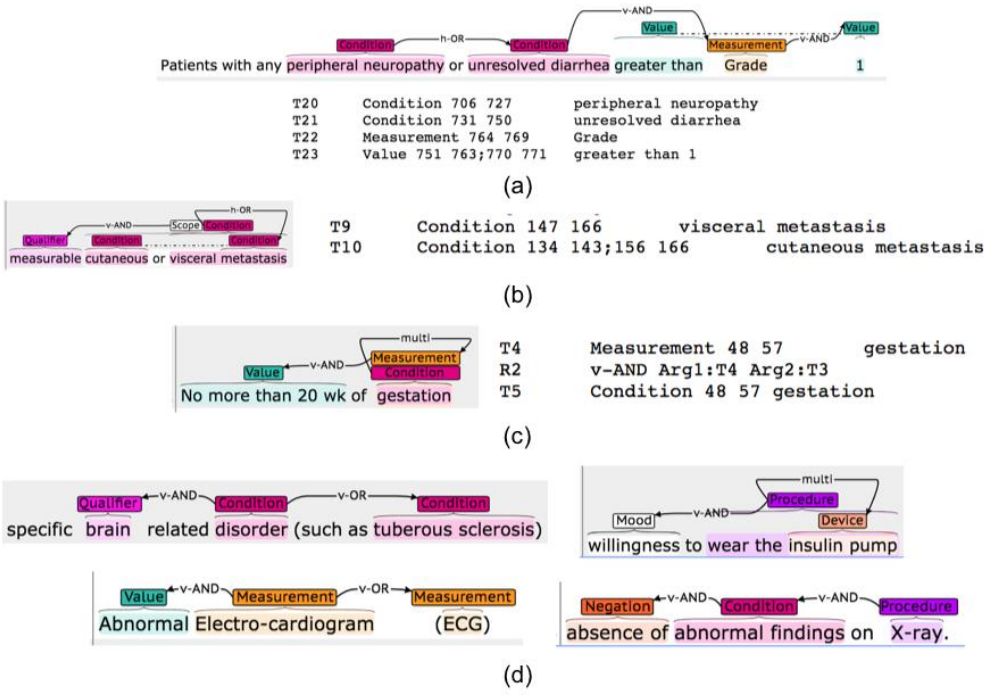


Figure 1. Various annotation examples from Chia and example C2Q predictions for some of them (highlighted by the green rectangles). **(a)**. Top: visualized annotation interface suitable for human review. Bottom: text file storing the annotation data suitable for machine processing. **(b)**. An example of a coordinated entity. **(c)**. An example where one same piece of text corresponds to two distinct annotated entities. **(d)** Examples of relationships annotated in Chia.

Alex *et al*. [21] categorized overlapping entities into three types: (1) entities containing one or more shorter embedded entities (e.g., "*wear the insulin pump*" and "*insulin pump*" shown in Figure 1(d)); (2) entities with more than one entity category (e.g., "*gestation*" can be categorized as Condition and Measurement as shown in Figure 1(c)); (3) coordination ellipsis ("*cutaneous metastasis*" and "*visceral metastasis*" in Figure 1(b)). Unlike most other flat corpora that exclude nested or overlapping entities [21], Chia uses a non-flat annotation scheme to accommodate these. The GENIA corpus supports overlapping entities [22] but focuses on biological entities, such as DNA, RNA, and protein.

The NER system: Criteria2Query

The NER system used in this study is Criteria2Query (C2Q) [23], which translates free-text eligibility criteria to OMOP CDM-based cohort queries. Its online demo is available at http://www.ohdsi.org/web/criteria2query/. Its output are computable queries in JSON format that can be directly fed into ATLAS [24] to define a patient cohort. In this study, we focused the comparison on the NER module, which recognizes eight entity types defined in the OMOP CDM, including CONDITION, DEMOGRAPHIC, DRUG, MEASUREMENT, OBSERVATION, PROCEDURE, VALUE, and TEMPORAL. An entity's type, indicating which category an entity is classified to, are denoted in

uppercase (e.g., CONDITION). Noted that as C2Q is constantly being updated, the latest online version of C2Q may cover more entity types that were not be available for inclusion in the evaluation at the time of this study.

Material and Methods

In this study, we refer to a recognized entity as a prediction and a manually annotated entity a reference (as ground truth). If the prediction and the reference are exactly same, we call it an exact match. Besides exact match, there are partial matches. For example, in text "intercostal post-herpetic neuralgia", if the NER system recognizes "neuralgia" as opposed to "post-herpetic neuralgia," there is a partial match. Researchers have developed a variety of rules that relax boundary matching criteria to different degrees, including *Left match*, *Right match*, *Partial match*, *Approximate match*, *Name part/fragment match*, *Core-term match*, etc. [2]. When there is no exact match, a discrepancy can occur at the syntactic level, where the human annotation and NLP annotation disagrees on the entity's boundaries, or at the semantic level, where the human annotation and NLP annotation disagrees on the semantics (e.g., concepts or categories). The latter requires adjudication by domain experts. We compared the output of Criteria2Query with Chia's annotations in terms of entity span and type (category). An entity's span is composed of one or more words or phrases.

We defined the following matches for analysis of the disagreement between human and NLP-assisted annotations:

- "relaxed match"—the prediction's span overlaps with the reference's span, which include
    - "exact match"—the prediction's span exactly overlaps with the reference's span.
    - "extra match"—the prediction's span strictly contains the reference's span.
    - "partial match"—the prediction's span is strictly contained by the reference's span.
- "spurious match"—the prediction's span has no overlaps with any reference's span.
- "missing match"—no prediction's span overlaps with the reference's span.

If there is a match, we further compared the agreement on concept categories as one of the following:

- "correct"—the prediction's category agrees with the reference's category.
- "incorrect"—otherwise.
- "N/A"—not applicable.

We counted the frequency of each combination of disagreement scenarios. We also grouped the entities by its annotated categories and check the disagreement type distributions in each. These can be conducted completely systematically without human inspection of the texts. The questions of our main interest include:

- Which type of disagreement between human and NLP-based annotations are frequent?
- What is the distribution of the disagreement types?
- Between which categories are mis-categorization frequent?

The disagreement type depends on how a reference aligns with a prediction. During the NER evaluation, a prediction and a reference align if their spans overlap. It is possible that a prediction aligns with multiple references, or vice versa. Hence, disagreement can occur for one prediction if different references are selected as alignments. This multi-matching issue is commonly observed in many NER evaluation and are particularly frequent for evaluations against non-flat annotations [4]. To simplify our error analysis, we designed a rule-based method to align the predictions and references with the highest overlapping ratio in their spans. After the alignment, each prediction was matched to at most one reference (if aligned with nothing, the case is a "missing"), and was assigned a unique disagreement type.

We designed the following 3-step workflow for our disagreement analysis: 1) we first manually inspected the disagreement instances and examples; 2) we abstracted a disagreement pattern from similar instances and formulated the definition for the pattern; and 3) according to the definition, we programmatically fetched all the incidents belonging to this pattern and determined if it represents substantial portion of all the disagreement.

Results

Data Statistics

A total of 1000 trials in Chia were used for the evaluation after excluding 66 trials due to errors in Chia annotations or technical barriers in running C2Q due to server instability issues. From the eligibility criteria text of those 934 trials, there were 28,671 distinct references from Chia, while C2Q generated 19,974 distinct predictions, which fell into the aforementioned eight entity categories. Figure 2 shows entities annotated in Chia and predicted by C2Q for each category. We saw CONDITION was the dominant category in Chia with the highest frequency, while the frequencies of the other five categories (DEMOGRAPHIC, DRUG, MEASUREMENT, VALUE, TEMPORAL) shared the same

order of magnitude. The number of entities predicted by C2Q for each category were roughly proportional to the corresponding category of Chia's annotations except categories of OBSERVATION and PROCEDURE.

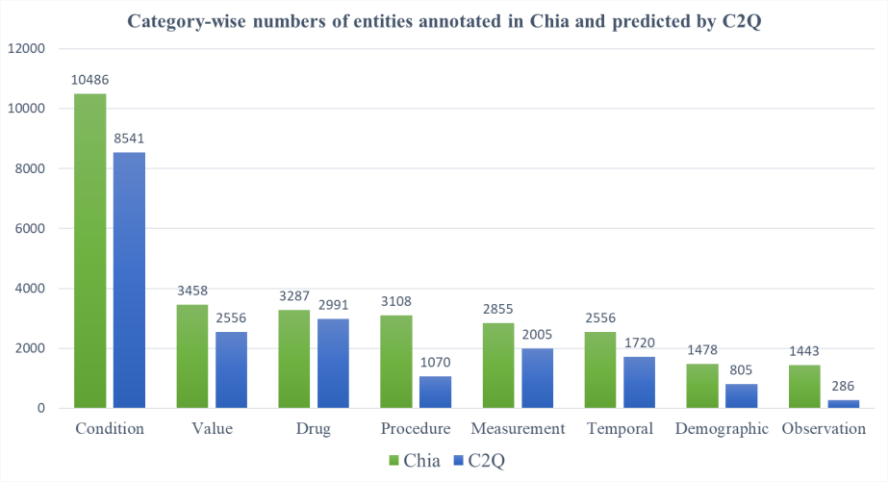


Figure 2. Category-wise numbers of entities annotated in Chia and predicted by Criteria2Query, respectively.

In-depth human- and NLP-based annotation disagreement analysis

We identified six disagreement patterns and reported the number of incidents for each pattern in Table 1. These patterns are indictors of the inconsistences between human annotations and machine learning-based annotations, not necessarily NER errors due to the imperfection of human annotations. Among these patterns, the most prevalent pattern includes 1532 disagreement instances, where C2Q predicted an entity together with its descriptive qualifier, temporal or value attributes. The pattern with the least number of cases (120) is mis-categorization between CONDITION and OBSERVATION. Each disagreement pattern along with its example(s) and error analysis are elaborated later.

Table 1. Prevalence of the disagreement pattern.

No.	Type of Match	Disagreement patterns	Frequency
1	Extra Match (N=3300)	Recognizing qualifier, temporal and value attributes together with an entity	1532 (46.4%)
2		Recognizing a coordinated elliptical expression as an entity	591 (17.9%)
3		Recognizing parentheses together with an entity	338 (10.2%)
4	Missing (N=10433)	Missing predictions around multi-labeled or nested entities	601 (5.8%)
5	Partial Match (N=1366)	Omitting the reference point of a TEMPORAL entity	392 (28.7%)
6	Exact Match (N=5217)	Mis-categorization between CONDITION and OBSERVATION	120 (2.3%)

Table 2 shows the entity count for each boundary matching criteria and correct predictions. C2Q predicted 12,584 "exact" matches with Chia's entities with 11,501 correct predictions. In addition, 3,801 "extra" matches and 1,853 "partial" matches were identified with boundary relaxing. With our relaxed matching criteria, we found additional 4,666 correct predictions (28.9% of total correct predications) to obtain 16,167 correct predictions.

Table 2. Comparison between human and Criteria2Query annotations

Type of Match	Match w Chia	No Match w Chia	<NA>	Total predictions by Criteri2Qery
exact	11501 (91.4%)	1083 (8.6%)	0	12584
extra	3300 (86.8%)	501 (13.2%)	0	3801
partial	1366 (73.7%)	487 (26.3%)	0	1853
spurious	0	0	1736	1736
missing	0	0	10433	10433
Total	16167	2071		

For entities in the "exact" matching, we further computed their categorization contingency table (Table 3). From the two tables we can see C2Q achieved a good overall prediction accuracy in categorization. We observed that 74 (42.5% =74/174) observation entities are mis-categorized into CONDITION, and 214 (22.1%=214/969) procedure entities are mis-categorized into DRUG. Possible reasons for this observation are discussed later.

Table 3. The contingency table of categorization over "exact" predictions.

		Criteria2Query Annotations							
		Condition	Demographic	Drug	Measurement	Observation	Procedure	Value	Temporal
Chia	Condition	4956 (96.2%)	0	35 (0.7%)	80 (1.6%)	46 (0.9%)	33 (0.6%)	1 (0.02%)	0
	Demographic	9 (1.2%)	722 (96.4%)	3 (0.4%)	7 (0.9%)	6 (0.8%)	0	0	2 (0.3%)
	Drug	64 (3.3%)	0	1819 (94.8%)	28 (1.5%)	0	8 (0.4%)	0	0
	Measurement	67 (5.8%)	1 (0.09%)	36 (3.1%)	1033 (89.9%)	7 (0.6%)	4 (0.3%)	1 (0.09%)	0
	Observation	74 (42.5%)	0	8 (4.6%)	7 (4.0%)	81 (46.6%)	4 (2.3%)	0	0
	Procedure	125 (12.9%)	0	214 (22.1%)	70 (7.2%)	9 (0.9%)	550 (56.8%)	0	1 (0.1%)
	Value	8 (0.4%)	3 (0.2%)	0	3 (0.2%)	0	0	1693 (94.4%)	87 (4.8%)
	Temporal	0	0	0	0	0	1 (0.15%)	31 (4.6%)	647 (95.3%)

We elaborate on the six disagreement patterns. For each pattern, one or more examples are analyzed to delineate it. Then a more systematic definition of the pattern was used to programmatically fetch all the instances following such pattern. Then the representation of this patterns among all the disagreement cases are discussed. Note that these disagreement patterns are not necessarily mutually exclusive, many disagreement instances being the compound of two or more patterns.

1). Recognizing the qualifier together with an entity

Figure 3. Example of recognizing the qualifier together with an entity (**Left**: Chia; **Right**: C2Q).

We observed that entities classified into "extra" was much more frequent than "partial." This is due to C2Q frequently included an entity's qualifiers as part of the real entity. Figure 3 (b) shows an example where C2Q predicted the "severe respiratory disease" as a whole entity while in Chia (Figure 3(a)) the "severe" is annotated as the qualifier of the "respiratory disease". We introduced a disagreement type called *extra_qualifier* to indicate cases where the prediction is "exact" or "partial" to the reference if dropping the "extra" qualifier(s). We found 1049 instances could be classified into this pattern, making this the most frequent pattern among "extra" recognitions.

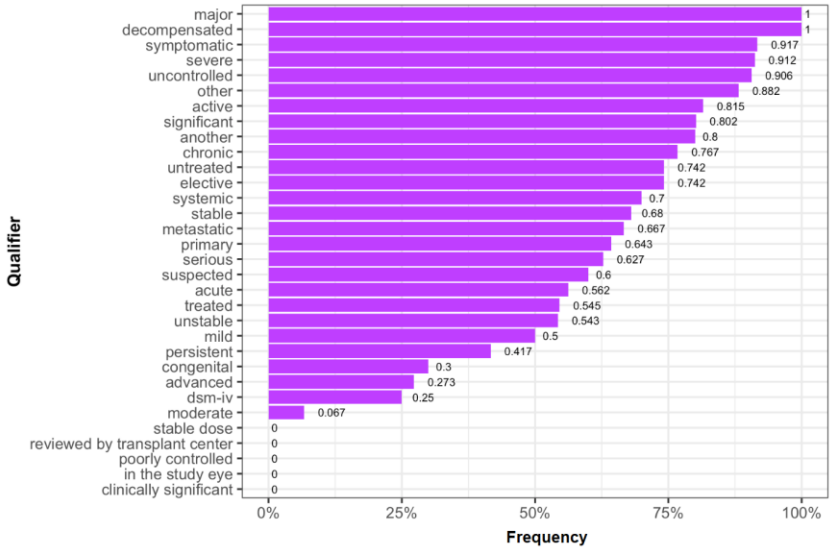


Figure 4. Frequency of qualifiers being parsed together with the target entity.

Two follow-up questions we then had were: (1) What qualifier terms were prone to be recognized as part of the target entity being qualified? (2) Was the probability of a qualifier to be recognized as part of its target term dependent on the target term itself? To address the first questions, we computed each qualifier term's relative frequency if a qualifier was parsed together with its target entity by C2Q for at least 10 times. Figure 4 shows the top portion of the results. Four qualifiers (i.e., *major*, *decompensated*, *symptomatic*, *severe*) were recognized together with their target entity over 90% of occurrences while some other qualifiers (*clinically significant*, *in the study eye*, etc.) rarely were. To answer the second question, we checked that for a qualifier that was recognized together with entities for at least 80% of cases, whether it is dominant by certain target entities. Interestingly, none of them had a dominant target entity, implying that they were considered part of the entity regardless of the real target entity.

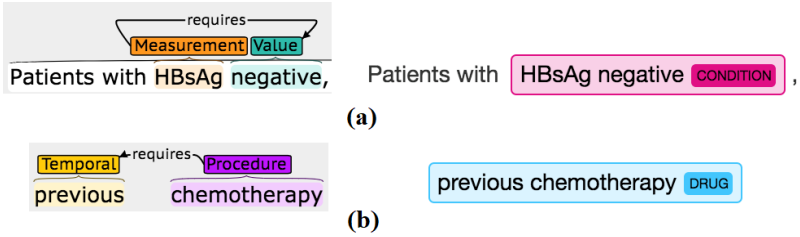


Figure 5. Example of recognizing the value or temporal together with an entity (**Left**: Chia; **Right**: C2Q).

Recognizing an entity's value (Figure 5(a)) or temporal (Figure 5(b)) as part of an entity also caused a substantial number of "extra" disagreement instances. Systematically, we defined the disagreement type *extra_value* and *extra_temporal* to be the cases where the prediction is "extra" to the reference but can be updated to "exact" or "partial" if its value(s) or temporal(s) are dropped, respectively. A total of 329 (9.9%) "extra" cases were assigned to *extra_value*. Similarly, 154 (4.7%) "extra" cases were classified to *extra_temporal*.

2). Recognizing a coordinated elliptical expression as an entity

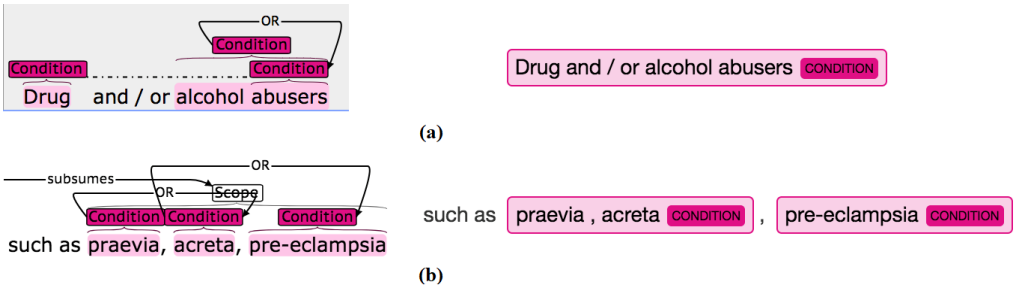


Figure 6. Example of recognizing a logic group in its entirety as an entity (**Left**: Chia; **Right**: C2Q).

Another major cause of "extra" was the recognition of a coordination ellipsis as a single entity. For example, in Figure 6 (a), C2Q recognized "Drug and / or alcohol abusers" as a single entity, which was annotated as two separate entities "Drug abusers" and "alcohol abuser" connected by "or" in Chia. In Figure 6 (b), C2Q recognized "praevia, acreta" as a single entity while the "praevia" and "acreta" were two separate entities in Chia. Systematically, we defined this disagreement type as *extra_logic* to represent cases where: 1) a prediction is matched to more than one references; 2) the prediction is "extra" to each of the reference; 3) the prediction's text contains " and ", " or ", or ", " (note the spaces around the keywords) or "/" (excluding the cases where "/" is used for unit or mathematical formula). This pattern can be further subdivided into two types: the complete recognition of a simple logic group where the entities in the group are mutually independent (Figure 6 (b)); the complete recognition of a coordinated ellipsis group where the entities in the group share a piece of text (Figure 6 (a)). We found Criteria2Query mistaken 591 coordination ellipsis expressions for one semantic unit (17.9% of all "extra match"), implying the significance of this phenomenon in criteria text and the need for dedicated technology to address it. Yuan *et al*. contributed a related method [25].

3). Recognizing parentheses together with an entity

Figure 7 shows four examples of inconsistency between Chia and Criteria2Query around parentheses. Systematically, we defined the disagreement type *extra_parenthesis* to be the case where a prediction is "extra" to the "reference" and the prediction text contains "(" or ")". We found 338 (10.2%) "extra match" disagreement could be assigned to *extra_parenthesis*, and they can be further divided into four scenario: 1) recognition combining the entity and its

acronym in parentheses (N=153 instances, e.g. Figure 7(a)); 2) recognition combining an entity followed by an open parenthesis plus the first term (N=119 instances, e.g. Figure 7(b)); 3) recognition combining an ending parenthesis in addition to the last entity inside the parentheses (N=63 instances, e.g. Figure 7(c)); 4) recognition combining multiple close and open parentheses were also observed (N=3 instances, e.g. Figure 7(d)).

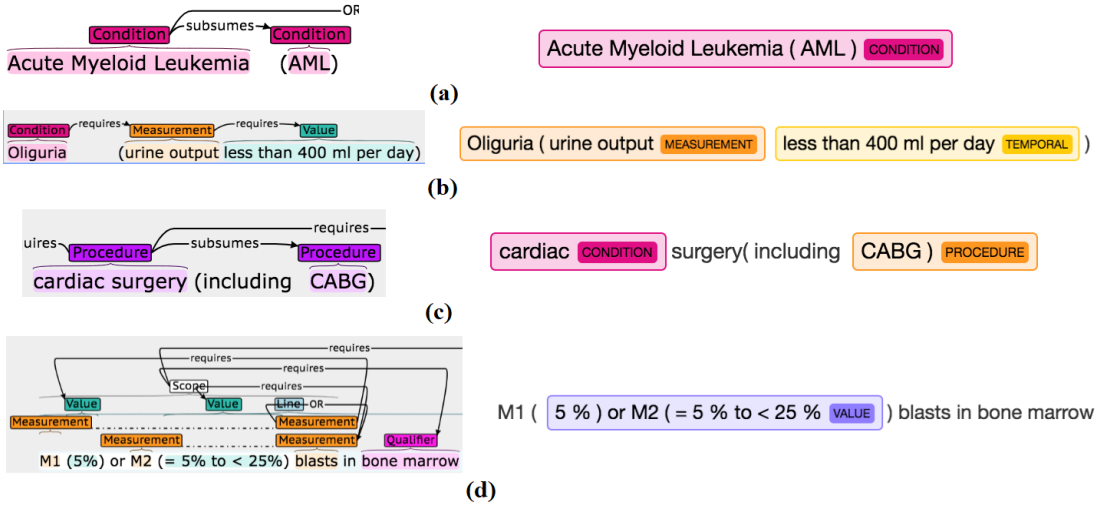


Figure 7. Example of recognizing parentheses together with an entity (**Left**: Chia; **Right**: C2Q).

4). Missing predictions around multi-labeled or nested entities

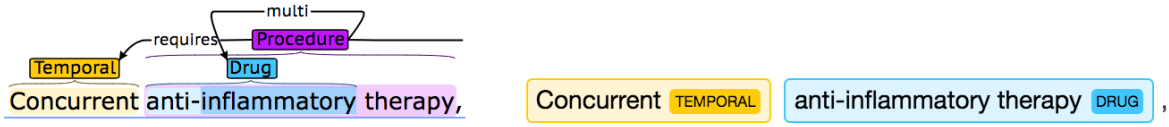


Figure 8. Example of missing predictions around multi-labeled or nested entities (**Left**: Chia; **Right**: C2Q).

We have stressed the existent of multi-labeled or nested entities in Chia. However, C2Q made flat predictions, i.e. it made no predictions for overlapping spans. Hence it inevitably missed some entities that are part of non-flatness annotations. For example, the "anti-inflammatory" in Figure 8 is a DRUG entity nested in the PROCEDURE entity "anti-inflammatory therapy", while C2Q could only make one prediction for such text. Systematically, we defined an disagreement pattern *missing_multi/nested* to be cases where: a prediction aligns with multiple references and its span overlaps with each of their spans. We found this pattern caused 601 (5.8%) of "missing" disagreement.

5). Omitting the reference point of a TEMPORAL entity

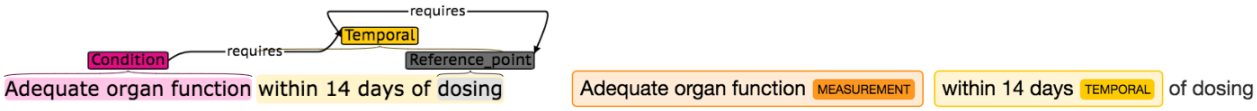


Figure 9. Example of omitting the reference point of a TEMPORAL entity (**Left**: Chia; **Right**: C2Q).

In general, disagreement was less frequent when using "partial match" than "extra match". However, for TEMPORAL entities, "partial" disagreement manifested more frequently than the "extra" disagreement. The reason is that C2Q omitted the reference point of recognitions for TEMPORAL entities. An example is shown in Figure 9. The most often used keywords for the reference time points were "after", "before", "of", "from", "since", "past", "preceding", "pre" and "post". Thereby, systematically, we defined this disagreement type as *partial_reference_point*, to represent prediction of TEMPORAL entities with "partial" disagreement due to omitting the reference time points indicated by the keywords listed above. We found this pattern was responsible for 392 (28.7%) disagreed "partial matches".

6). Mis-categorization between CONDITION and OBSERVATION

From Table 3 we observe that 74 of the observation entities were mis-categorized as CONDITION and 46 condition entities were annotated as OBSERVATION. For example, in Figure 10(a) the "sexually active" was annotated as

OBSERVATION in Chia but was predicted as CONDITION by C2Q in Figure 10(b). However, we noticed Chia contains inconsistent annotations of the same term. For example, in Chia "sexually active" was annotated as CONDITION in Figure 10(c). We found that 45 terms were miscategorized from OBSERVATION to CONDITION at least once, and 15 terms were miscategorized from CONDITION to OBSERVATION at least once.

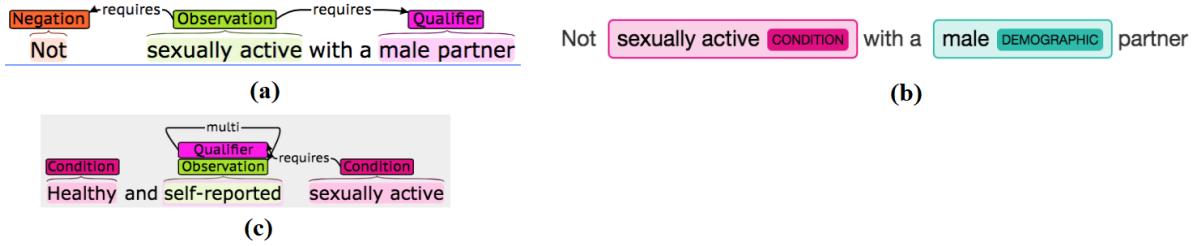


Figure 10. Example of mis-categorization between CONDITION and OBSERVATION.

Discussion

In this study, we compared human and NLP annotations of clinical trial eligibility criteria text that both followed the OMOP CDM, an increasingly popular and widely adopted clinical data standard by the clinical research informatics community as a rich clinical data model. Therefore, the disagreement patterns identified in this study may offer generalizability implications for many NER tasks of interest with similar setup. The disagreement patterns help unveil the complex semantic expressions and sophisticated logic used in clinical trial eligibility criteria text, such as the frequently combined conditions in coordination ellipsis (pattern 2) and the importance of reserving temporal reference point for TEMPORAL entities (pattern 5). Patterns 1 and 6 provide insights to the ambiguity of the OMOP CDM, while other patterns (pattern 2, 4, and 5) manifested themselves as the semantic/logic complexity in eligibility criteria text that may require tailoring the ML-based NER system accordingly. Meanwhile, explicit definition of principles for annotating descriptive modifiers and temporal or value attributes for named entities is important for achieving the comparative effectiveness of NER evaluations.

Standard NER systems evaluated using corpora from challenges such as MUC, CONLL or ACEU employed "exact match". However, strict exact-boundary match may not always reflect the true performance of an NER system. For example, a human annotator annotated "SARS-CoV-2 infection" from an eligibility criterion "progressive disease suggestive of ongoing SARS-CoV-2 infection." A clinical NER system makes the prediction of "progressive disease suggestive of ongoing <CONDITION>SARS-CoV-2</CONDITION> infection." We noticed a boundary mismatch disagreement where "infection" was not included in the NER system's extraction. However, this disagreement may not be an error of NLP systems if "SARS-CoV-2" as a condition is adequate for downstream tasks such as concept normalization, document indexing, or relationship extraction. The *extra_qualifier*, *extra_value* and *extra_temporal* patterns, refer to disagreement where descriptive adjectives were annotated as parts of following entities. In reality, even annotators are oftentimes confused and make subjective decisions on whether descriptive adjectives such as "severe" or "secondary" should be considered part of entity names based on the clinical context. Therefore, these three patterns should not be judged as errors bluntly because some cases may be acceptable in the context of applications. From the perspective of OMOP CDM vocabulary, some pre-coordinated terms are included. For example, "severe cytopenia" is in the terminology while "severe arrhythmia" is not. Pre-coordination vs. post-coordination of the terms has been actively discussed and poses challenge to NER systems for predicting descriptive adjectives. Of course, specific rules and examples in annotation guidelines may help alleviate this issue; but ambiguity seems inevitable.

The open-ended definition of Observations in OMOP CDM leads to some concepts existing in both the CONDITION and OBSERVATION domains. This is due to the catch-all nature of the ambiguous OBSERVATION domain (https://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:observation) in the OMOP CDM: *"The* OBSERVATION table captures clinical facts about a Person obtained in the context of examination, questioning or a procedure. Any data that cannot be represented by any other domains, such as social and lifestyle facts, medical history, family history, etc. are recorded here." Fan and Friedman previously reported the abundant ambiguity in the "finding" semantic type in the UMLS also cause similar problems for semantic classification of named entities [26]. Moreover, the human annotations are not perfect. Chia has an inter-rater agreement of 81%, which imposes an accuracy ceiling for the machine learning algorithm for NER. Reasons for human errors could be complex, e.g., the ambiguities in the OMOP CDM, the lack of details in annotation guidelines, subjective decisions by annotators, or even human mistakes due to fatigue. Similarly, Xu *et al*. found that nearly half of the discrepancy between the system

and gold standard were due to errors in gold standard annotation [27]. Therefore, imperfections in human annotations is a common problem; yet its impact on machine learning-based NLP needs more careful evaluation.

For categories that are not clearly defined or not consistently exclusive, one can employ categorical relaxation to merge the two categories to reduce the ambiguity of NER evaluations [2]. This technique is often used to merge protein, DNA and RNA when no distinctions are required. In our study, 174 disagreement will be eliminated if categorical relaxation is employed to merge the evaluation of CONDITION with OBSERVATION. The aforementioned ambiguous representations in the OMOP CDM and inconsistent annotations in the evaluation corpus can diminished an NER's real performance. With Chia's inter-annotator agreement being around 81%, even if an NER system is trained with Chia corpus, the annotation inconsistencies and ambiguity exist in Chia will inevitably propagate to the machine learning based NER system. If the strict exact-boundary match is used as gold standard, the trained machine learning NER model could be penalized for overfitting the human annotations in Chia. It is another reason that for certain applications, exact match can weaken the reliability of an NER system's performance and customized relaxed matching criteria should be leveraged.

Limitations: One limitation of this study is that Criteria2Query was not able to resolve multiple or nested annotated entities at the time of the study. This limitation is due to current Criteria2Query's conditional random field (CRF) implementation cannot handle nested NER, where an entity can be contained in other entities [21]. Alex *et al*. [21] attempted to recognize the nested entities in GENIA corpus by specially pre-processing the annotation and saw an improvement over the baseline flat system. Byrne [28] used a multi-word token method and obtained a promising result in recognizing nested named entities in historical archive texts. Nevertheless, the work on nested NER has almost been entirely ignored until recently [29], and the technology is not so mature as that of regular flat NER. Another limitation of this study is the lack of integration of our method for recognizing coordination ellipsis—another case that the non-flatness concerns. The coordinated ellipses is particular hard [30] in NER. The coordinated ellipses, along with the simple logic group and the parenthesized insertion, are called composite mention in some research [31]. Buyko *et al*. [30] utilized CRF to resolve the coordinated ellipses in GENIA corpus and attained a good performance. Wei *et al*. [31] integrated machine learning and pattern identification to handle all types of the composite mentions. Yuan *et al.* [25] proposed a graph-based representation model to reconstruct concepts from coordinated elliptical expressions. This model is being incorporated into Criteria2Query to further improve the NER performance.

Conclusions

This study describes an in-depth analysis of NER annotation disagreement between human and NLP for clinical trial eligibility criteria by comparing human annotations with NLP annotations. We identified six types of disagreement between human and NLP annotations following the OMOP CDM, and highlighted the complexities in logic and temporal information, all requiring further improvement of NER methods. Our study also shows that relaxed match increased accuracy reporting by about 28.9% over exact match. We recommend reporting prevalent disagreement patterns in the context of application in addition to quantitative metrics to formulate a comprehensive NER assessment.

Acknowledgements

This research was sponsored by the National Library of Medicine grant R01LM009886 (PI: Weng) and the National Center for Advancing Translational Science grant U24TR001579 (PI: Harris).

Conflicts of Interest

Dr. Xu and The University of Texas Health Science Center at Houston have research-related financial interests in Melax Technologies, Inc.

References

[1] Marrero M, Urbano J, Sánchez-Cuadrado S, Morato J, Gómez-Berbís JM. Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces. 2013;35(5):482-9.

[2] Tsai RT-H, Wu S-H, Chou W-C, Lin Y-C, He D, Hsiang J, et al. Various criteria in the evaluation of biomedical named entity recognition. BMC bioinformatics. 2006;7(1):92.

[3] Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. Journal of biomedical informatics. 2015;58:11-8.

[4] Galibert O, Rosset S, Grouin C, Zweigenbaum P, Quintard L. Structured and extended named entity evaluation in automatic speech transcriptions. Proceedings of 5th International Joint Conference on Natural Language Processing; 2011.

[5] Atdağ S, Labatut V. A comparison of named entity recognition tools applied to biographical texts. 2nd International conference on systems and computer science; 2013: IEEE.
[6] Marrero M, Sánchez-Cuadrado S, Lara JM, Andreadakis G. Evaluation of named entity extraction systems. Advances in Computational Linguistics, Research in Computing Science. 2009;41:47-58.
[7] Jiang R, Banchs RE, Li H. Evaluating and combining name entity recognition systems. Proceedings of the Sixth Named Entity Workshop; 2016.
[8] Chinchor N. MUC-4 evaluation metrics. Proceedings of the 4th conference on Message understanding; McLean, Virginia: Association for Computational Linguistics; 1992. p. 22–9.
[9] Chinchor N, Sundheim BM. MUC-5 evaluation metrics. Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993; 1993.
[10] Jannet MB, Adda-Decker M, Galibert O, Kahn J, Rosset S. Eter: a new metric for the evaluation of hierarchical named entity recognition2014.
[11] Makhoul J, Kubala F, Schwartz R, Weischedel R. Performance measures for information extraction. Proceedings of DARPA broadcast news workshop; 1999: Herndon, VA.
[12] SNOMED. Available from: http://www.snomed.org/.
[13] Medical Subject Headings. Available from: https://www.nlm.nih.gov/mesh/meshhome.html.
[14] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Setzer A, et al. Semantic annotation of clinical text: The CLEF corpus. Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining; 2008.
[15] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl_1):D267-D70.
[16] Šuster S, Daelemans W. CliCR: a dataset of clinical case reports for machine reading comprehension. arXiv preprint arXiv:180309720. 2018.
[17] OMOP Common Data Model – OHDSI. Available from: https://www.ohdsi.org/data-standardization/the-common-data-model/.
[18] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics. 2015;216:574.
[19] Cimino JJ, Shortliffe EH. Biomedical Informatics: Computer Applications in Health Care and Biomedicine (Health Informatics): Springer-Verlag; 2006.
[20] Kury F, Butler A, Yuan C, Fu L-h, Sun Y, Liu H, et al. Chia, a large annotated corpus of clinical trial eligibility criteria. Scientific data. 2020.
[21] Alex B, Haddow B, Grover C. Recognising nested named entities in biomedical text. Biological, translational, and clinical language processing; 2007.
[22] Kim J-D, Ohta T, Tateisi Y, Tsujii Ji. GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19(suppl_1):i180-i2.
[23] Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association. 2019;26(4):294-305.
[24] ATLAS. Available from: http://www.ohdsi.org/web/atlas/#/home.
[25] Yuan C, Wang Y, Shang N, Li Z, Zhao R, Weng C. A graph-based method for reconstructing entities from coordination ellipsis in medical text. Journal of the American Medical Informatics Association. 2020.
[26] Fan J-W, Friedman C. Semantic classification of biomedical concepts using distributional similarity. Journal of the American Medical Informatics Association. 2007;14(4):467-77.
[27] Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Studies in health technology and informatics. 2004;107(Pt 1):565.
[28] Byrne K. Nested named entity recognition in historical archive text. International Conference on Semantic Computing (ICSC 2007); 2007: IEEE.
[29] Finkel JR, Manning CD. Nested named entity recognition. Proceedings of the 2009 conference on empirical methods in natural language processing; 2009.
[30] Buyko E, Tomanek K, Hahn U. 2007. Resolution of coordination ellipses in biological named entities using Conditional Random Fields. In PACLING 2007-Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics; 2007: Citeseer.
[31] Wei C-H, Leaman R, Lu Z. SimConcept: a hybrid approach for simplifying composite named entities in biomedical text. IEEE journal of biomedical and health informatics. 2015;19(4):1385-91.

Challenges to Global Standardization of Outcome Measures

Zoe Liao, B.S. Pharmacy[1], Yuri Quintana PhD[2,3]
[1]School of Pharmacy, Northeastern University, Boston, MA, USA; [2]Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Boston, MA, USA; [3]Harvard Medical School, Boston, MA, USA

Abstract

Global standardization of outcome measures for disease states can help researchers and healthcare providers compare healthcare institutions' and populations' health outcomes. Despite the creation of standardized outcome sets, clinical institutions' adoption of these sets is not common. A literature review shows that among the challenges to standardizing outcome measures include the difficulties of achieving consensus in the working groups creating these outcome sets, the tradeoffs made when selecting outcome measurement tools, and the high costs of implementing a new or different set of outcome measures. The duplication of effort to create these standard sets can also limit standardization, which could be minimized through increased transparency of how these standard sets are developed. We propose some approaches to improve how to create and implement standard sets to broaden their usability across institutions.

Introduction

As healthcare spending increases across the world, many nations are reforming their healthcare systems to prioritize measuring the value of care over the volume of care. The "value" of health care is defined as "the outcomes patients may experience relative to the cost of delivering these outcomes."[1] Scaling and continuously collecting these outcome data is now possible given the advancement of technological capabilities.[2] Moreover, the systematic use of information collected from patient-reported outcome measures has improved patient-provider communication and patient satisfaction with health care.[2,3] However, even for the same disease state, there is a wide variation in the outcomes recorded in electronic health records (EHRs), claims databases, patient registries, and prescription databases.[2] There is also a variation of outcomes measured in the design of clinical trials for the same disease.[4] On top of this, the actual measures used to evaluate the desired outcomes are heterogeneous. The current lack of global standardization of outcome measures hinders direct comparisons and meta-analyses of clinical trials. It detracts from global learning and goals of using longitudinal data for comparison of intervention effects.[1]

Achieving consensus on key considerations–which outcomes to include, how to measure them, and when to measure them–is a lengthy and challenging process. In 2010, the Core Outcome Measures in Effectiveness Trials (COMET) Initiative created a free, online database detailing the ongoing studies that aim to apply rigorous consensus methods to develop core outcome sets (COS).[4] Core outcome sets are defined as the minimum sets of outcomes that should be measured and reported in all clinical trials of a specific disease or for application in disease registries or clinical practice.[4]

Among the collaboratives that have been working to create these core outcome sets is the nonprofit International Consortium for Health Outcomes Measurement (ICHOM), which prioritizes the incorporation of patient-reported outcomes measures (PROMs) into COS they call "Standard Sets." These PROMs include symptoms, health-related quality of life, and satisfaction of care.[1,5,6] As of 2018, ICHOM Working Groups, which consist of clinicians, researchers, and patient representatives around the world, have developed Standard Sets that cover 54% of the global disease burden.[4] However, from their start in 2012, the only studies documenting the implementation or feasibility of using ICHOM Standard Sets have been conducted or funded by ICHOM themselves.[1,5] This may be due to competing efforts of other groups developing standard sets for the same diseases and the financial and logistical challenges for institutions to start implementing standard outcome sets. This paper will focus on the problems and potential solutions for different goals and compositions of the groups creating these outcome sets, the tradeoff required when choosing between the utility of an outcome measure and its feasibility for collection, the high financial costs of implementation, and the global applications.

Methods

A literature search was conducted in the PubMed database using health outcome measures, outcome assessments, and global health standards. The disease states searched were focused on cardiovascular disease, oncology, diseases common for the elderly, and mental or behavioral health. This was intended to explore any differences between healthcare areas with better infrastructure for care (e.g., cardiovascular disease, oncology) compared to areas with more fragmented care (e.g., mental health care, elder care). Both COS development and COS implementation studies were included to look for potential relationships between how core outcome sets are developed and how they are implemented.

Results

A literature search conducted on PubMed in June 2020 and results were screened for inclusion. Eight studies describing the standard sets for prostate cancer, dementia, heart failure, and hip and knee osteoarthritis, and behavioral health were included for analysis (Table 1).

Table 1. Summary of studies included in the review

Author (Year)	Disease State	Type of study: COS development, COS implementation, Both
Seligman et al. [1] (2018)	Cardiovascular diseases	Both (general overview and case study)
Meregaglia et al. [2] (2020)	Prostate cancer	Both (scoping review and case study)
Webster et al. [4] (2017)	Dementia	COS development (systematic review and consensus)
Ackerman et al. [5] (2018)	Osteoarthritis	COS implementation (feasibility)
Martin et al. [6] (2015)	Prostate cancer	COS development
McNamara et al. [7] (2015)	Coronary artery disease	COS development
Rajaram et al. [8] (2019)	Breast cancer	COS implementation (cross-sectional comparison)
Wing et al. [12] (1998)	Behavioral Health	COS development

Composition of Groups Creating Standard Sets

The major organizations developing core outcomes sets were the International Consortium for Health Outcomes Measurement (ICHOM) and the Core Outcome Measures in Effectiveness Trials (COMET) Initiative. ICHOM was founded in 2012 by Harvard Business school, The Boston Consulting Group, and The Karolinska Institute, through their funding and funding from various international sponsors.[5-9] Their goal for creating standardized, open-access sets of outcome measures is to include outcomes that matter to patients, as well as outcomes that can be tracked across different health systems and clinical registries. The members they seek in their working groups include both clinicians and non-clinicians around the world. The Core Outcome Measures in Effectiveness Trials (COMET) Initiative is an organization that provides methodological support to groups trying to develop core outcome sets.[2,10,11] The multidisciplinary organization grew from a 2010 meeting of researchers, regulators, and policymakers interested in developing core outcomes sets to improve the standards for data reporting and synthesis in clinical trials. Their publicly available database of ongoing COS development studies aims to promote collaboration among researchers as well as the application of the developed COS.[2]

The composition of the working groups creating these sets can impact the consensus of decisions made.[1] In a review of core outcome sets for prostate cancer, Meregaglia et al.[2] found that there were "notable gaps in reporting the 'stakeholders involved' and 'consensus process' adopted," and that "geographic representativeness of stakeholders was unbalanced in favor of Europe and North America." This scoping review applied the COS-STANdards for Development framework developed by COMET to assess COS development studies' quality systematically. Reviewers found that the ICHOM study protocol for the development of the Standard Set for Prostate Cancer did not report if the scoring process, the definition of consensus, and the criteria for including/adding/dropping outcomes, were determined a priori.[2,6] Furthermore, the Methods section of the same ICHOM study protocol does not disclose the extent of patient involvement or provide any background information on the patient representatives.[6] A similar

critique could be made for the study describing how ICHOM developed an outcome set for coronary artery disease patients.[7]

Discussion

Deciding Between Outcome Measures

A key aspect to consider in the development of COS is the tradeoff when choosing between the comprehensiveness of outcome measures and the feasibility of collecting such outcome measures. For ICHOM's prostate cancer Standard Set, Martin et al. [6] chose the Expanded Prostate Cancer Index Composite 26-question short form (EPIC-26) to measure PROMs instead of the validated EPIC-16 instrument. Although EPIC-16 was designed for easy implementation, ICHOM went with the lengthier EPIC-26 because it included a question on rectal bleeding, which they considered crucial since it could indicate late toxicity from radiation. ICHOM also did not have common instruments like the International Prostate Symptom Score and International Index of Erectile Function as part of their Standard Set due to the overlap with the EPIC-26 domains.

Another example of the difficulty of choosing between measurement tools can be found in developing a COS for dementia disease-modifying clinical trials.[4] Outcomes from ICHOM's Standard Set for dementia were considered for inclusion. The COS developers recommended either the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) or the Mini-Mental State Examination (MMSE) for measuring cognition due to the difficulty of choosing between cost and time. The ADAS-Cog is free but can only be administered by a trained tester and takes 45 minutes to administer. The MMSE can be administered by clinical staff with minimal extra training, but it is costly due to copyright. Ultimately, the decision of which measurement instrument to use was left to the COS user due to the dependence of feasibility and practicality on the user's resources and current workflow.

Care priorities can also look different among various socioeconomic and cultural groups. In a cross-sectional comparison study of patient-reported outcome measures among breast cancer survivors in Malaysia versus high-income countries,[9] investigators found that well-being, survival, and physical functioning were the most important PROMs for Malaysians and high-income country patients. However, Malaysian breast cancer survivors were less likely to rate social, emotional, cognitive, and sexual functioning as very important. Instead, they were more likely to prioritize symptoms and complications management. A gap analysis between the ICHOM Heart Failure Standard Set and a global selection of real-world data sources revealed that "data captured in data sources from North America and Europe more closely resembled the Standard Set, whereas data sources in Africa deviated the most."[1] Different countries also have different healthcare systems and technological capabilities. ICHOM considered this when developing their coronary artery disease COS, where they restricted longitudinal outcomes to those that could be captured as administrative data since countries without a single-payer health system could have trouble identifying events outside of the specific acute care episode in their registries or electronic health record databases.[7]

Cost of Implementation

The high costs of implementing COS serve as a barrier to the global standardization of outcomes. A feasibility study of implementing the ICHOM Standard Set for Hip and Knee Osteoarthritis in two hospitals in Australia calculated the costs of implementation and 17 months of data collection to be 94,955 AUD ($65,234 USD). This amount accounted for project coordinator time, IT support, ICHOM implementation support, equipment, and physiotherapist support for recruitment.[5] Costs will vary for institutions and disease states, depending on the available clinical, administrative, and IT support available as well as patient volume. Financial costs may increase as the time for data collection continues, since ICHOM recommends long-term or possibly lifetime data collection to properly assess PROMs,[6] but the cost per patient could decrease over time.[5] Rates of COS implementation would increase with financial support from research funders, trial registries, and policymakers.[2] For example, the UK National Institute of Health Research (NIHR) funded the COS development of the COS for disease modification of dementia so that future NIHR-funded trials will use that COS.[4]

Development Process

Studies outlining the development of the standard sets should provide transparency not only on the composition and backgrounds of the working group members but also on how and when the criteria for including outcome measures

were decided. The Methods section of various ICHOM study protocols do not disclose the extent of patient involvement or provide any background information on the patient representatives, limiting the generalizability of the developed outcome sets.[2,6,7] This is especially important, as patients or patient representatives from different socioeconomic, cultural, and educational backgrounds may have different perspectives on which outcomes are important to them. Furthermore, if the Standard Set developers change the criteria after conducting a Delphi survey,[2] this could introduce bias.

Different outcome sets for the same disease state are being created, with the main reason being cited as the difference in use case. Although the use case for ICHOM Standard Sets is intended for clinical practice, there is a potential for overlap between outcomes assessed for practice and research.[10] Groups developing COS for research purposes can incorporate PROMs from ICHOM Standard Sets. There is a greater chance of collaboration if all COS creators are transparent in their COS development protocols and register their studies on the COMET database to prevent the duplication of effort. This will make it easier and more evident to see if projects can be complementary. Collaboration is ideal for aligning standards across contexts, reducing the spread of limited resources for implementation.

Global Mental Health

In 1993, the United Kingdom sought to have some standardized behavioral health outcomes, and in 1998 the Health of the Nation Outcome Scales (HoNOS) were introduced[12]. HoNOS-1 is a 20-item instrument that covers four key areas of functioning of patients: behavior, impairment, symptoms, and social functioning. A study[13] in 2020 looked for evidence of its value or cost-effectiveness to consumers, clinicians, or administrators. Of the 260 studies reviewed, only one study reported positive outcomes, and none of them attempted to assess the cost of using the Health of the Nation Outcome Scale (HoNOS). The study investigated the effect of routine outcome measurement but concluded that it failed to result in the provision of evidence-based care. To date, the ability of HoNOS to improve the health and social functioning of mentally ill people has not been demonstrated. A very recent study[14] suggested that clinician sensitivity and bias may affect the use of the instrument. ICHOM has developed a set of outcome measures for anxiety[16], depression[17], and addiction[18], but it is too early to know if groups will adopt it. International collaboration for increasing mental health services and research capacity in Africa emphasizes the need for cooperation between institutions and training for the successful use of evidence-based knowledge.[15] Informatics training on terminology and data modelling will be needed to successfully collect and apply outcomes for health care improvement.

Global Application

The creation of standard sets for global use should also consider what outcome measures institutions are already using. For example, clinical practices already using common instruments like the International Prostate Symptom Score and International Index of Erectile Function, which ICHOM did not include in their Standard Set,[6] could face disruption in the longitudinal data collected if they switch to ICHOM's proposed measure of EPIC-26. The advantages and limitations of measurement tools proposed (i.e., high cost, not yet validated, only available in certain languages, etc.) should be described by COS developers. Standardized sets are more likely to be implemented if the measurement tools proposed to take less time and resources to use than measurement tools previously used at the institution.[1,5] ICHOM suggests that future work should make commonly used outcome measures more comparable to transition to a universal standard.[6] In the meantime, it could help potential users if Standard Sets specifically outlined which outcome domains overlap among measurement tools. Data in the same domains can be collected at the same time points to make future comparisons easier.

Since the goal is global standardization, COS developers should keep in mind that different countries will vary in resources and in the volume of changes that will need to be made to current data collection processes to standardize. Countries with the advantage of recent investments in registry infrastructure, such as Ireland, Canada, Australia, and the U.S.,[6] should report on their implementations of standardized outcome sets to improve them for countries with more limited resources to spare. Developers of core outcome sets should describe or suggest ways to collect the requested data, such as through administrative data or electronic health record databases or registries.

Conclusion

The global standardization of outcome measures, for both clinical trials and clinical practice, can allow institutions to learn from each other about which interventions are best for improving patient outcomes and reduce the cost of care

through the elimination of ineffective interventions. Many groups are developing core outcome sets with the goal for international implementation, but many barriers currently stand in the way. These barriers include duplication of effort by different groups due to a lack of transparency in the study protocols for COS development, the limitations and compromises of recommended outcome measures, and the financial challenges of implementing long-term COS data collection recommendations. Solutions that address one challenge can also help minimize other challenges. Collaboration among developers can identify overlapping outcome domains for research and clinical practice, reducing the number of measures institutions would have to implement for the same disease state. Transparency of study protocols would clarify which outcome measures are best suited for the institution looking to implement the COS, based on resources and patient population. Future reports on the implementation and performance of this COS should provide insights into improving COS quality. An increase in the number of successful deployments will hopefully convince hesitant institutions and countries to prioritize adopting a common COS until eventually the goal of standardizing health outcome measurements across the world is achieved. Training on medical informatics processes for data collection and representation will be needed to succefully collect and apply these outcome standards.

References

1. Seligman WH, Salt M, la Torre Rosas AD, Das-Gupta Z. Unlocking the potential of value-based health care by defining global standard sets of outcome measures that matter to patients with cardiovascular diseases. *Eur Heart J.* 2018;5(2):92-95. doi: 10.1093/ehjqcco/qcy056
2. Meregaglia M, Ciani O, Banks H, et al. A scoping review of core outcome sets and their 'mapping' onto real-world data using prostate cancer as a case study. *BMC Med Res Methodol.* 2020;20(1):41-41. doi: 10.1186/s12874-020-00928-w
3. Nelson EC, Eftimovska E, Lind C, Hager A, Wasson JH, Lindblad S. Patient reported outcome measures in practice. *BMJ*. 2015;350:g7818. Published 2015 Feb 10. doi:10.1136/bmj.g7818
4. Webster L, Groskreutz D, Grinbergs-Saull A, et al. Core outcome measures for interventions to prevent or slow the progress of dementia for people living with mild to moderate dementia: systematic review and consensus recommendations. *PLoS One*. 2017;12(6):e0179521-e0179521. doi:10.1371/journal.pone.0179521
5. Ackerman IN, Cavka B, Lippa J, Bucknill A. The feasibility of implementing the ICHOM standard set for hip and knee osteoarthritis: a mixed-methods evaluation in public and private hospital settings. *J Patient Rep Outcomes*. 2018;2:32-32. doi: 10.1186/s41687-018-0062-5
6. Martin NE, Massey L, Stowell C, et al. Defining a standard set of patient-centered outcomes for men with localized prostate cancer. *Eur Urol*. 2015;67(3):460-467. doi:10.1016/j.eururo.2014.08.075
7. McNamara RL, Spatz ES, Kelley TA, et al. Standardized outcome measurement for patients with coronary artery disease: consensus from the International Consortium for Health Outcomes Measurement (ICHOM). *J Am Heart Assoc*. 2015;4(5):e001767. Published 2015 May 19. doi:10.1161/JAHA.115.001767
8. ICHOM: International Consortium for Health Outcomes Measurement. ICHOM website. https://www.ichom.org/. Accessed August 27, 2020.
9. Rajaram N, Lim ZY, Song CV, et al. Patient-reported outcome measures among breast cancer survivors: A cross-sectional comparison between Malaysia and high-income countries. *Psychooncology*. 2019;28(1):147-153. doi:10.1002/pon.4924
10. MacLennan S, Williamson PR, Lam TB. Re: Neil E. Martin, Laura Massey, Caleb Stowell, et al. Defining a standard set of patient-centered outcomes for men with localized prostate cancer. *Eur Urol* 2015;67:460–7.
11. COMET Initiative: Core Outcome Measures in Effectiveness Trials. COMET Initiative website. http://www.comet-initiative.org/. Accessed August 27, 2020.
12. Wing JK, Beevor AS, Curtis RH, Park SB, Hadden S, Burns A. Health of the Nation Outcome Scales (HoNOS). Research and development. *Br J Psychiatry*. 1998 Jan;172:11-8. doi: 10.1192/bjp.172.1.11. PMID: 9534825.
13. Bender KG. The meagre outcomes of HoNOS. *Australas Psychiatry.* 2020 Apr;28(2):206-209. doi: 10.1177/1039856219875066. Epub 2019 Sep 30. PMID: 31564114.
14. Williams, J. Clinicians' ratings of the Health of the Nation Outcome Scales (HoNOS) show sensitivity/bias that may affect patients' progress: analyses of routine administrative data. Preprint available at medRxiv 2020.11.19.20234674; doi: https://doi.org/10.1101/2020.11.19.20234674
15. Gureje O, Seedat S, Kola L, Appiah-Poku J, Othieno C, Harris B, Makanjuola V, Price LN, Ayinde OO, Esan O. Partnership for mental health development in Sub-Saharan Africa (PaM-D): a collaborative

initiative for research and capacity building. *Epidemiol Psychiatr Sci*. 2019 Aug;28(4):389-396. doi: 10.1017/S2045796018000707. Epub 2018 Nov 27. PMID: 30479242; PMCID: PMC6536364.
16. International Consortium For Health Outcomes Measurement. ICHOM Standard Set for Depression & Anxiety Available at URL: https://www.ichom.org/portfolio/depression-anxiety/
17. International Consortium For Health Outcomes Measurement. ICHOM Standard Set for Children & Young People with Depression & Anxiety. Available at URL https://www.ichom.org/portfolio/anxiety-depression-ocd-and-ptsd-in-children-and-young-people/
18. International Consortium For Health Outcomes Measurement. ICHOM Standard Set for Addiction. Available at URL: https://www.ichom.org/portfolio/addiction/

Integration of NLP2FHIR Representation with Deep Learning Models for EHR Phenotyping: A Pilot Study on Obesity Datasets

Sijia Liu, PhD[1], Yuan Luo, PhD[2], Daniel Stone, BS[1], Nansu Zong, PhD[1], Andrew Wen, MS[1], Yue Yu, PhD[1], Luke V. Rasmussen, MS[2], Fei Wang, PhD[3], Jyotishman Pathak, PhD[3], Hongfang Liu, PhD[1], Guoqian Jiang, MD, PhD[1]
[1]Mayo Clinic, Rochester, MN; [2]Northwestern University, Chicago, IL; [3]Weill Cornell Medicine, New York, NY

Abstract

HL7 Fast Healthcare Interoperability Resources (FHIR) is one of the current data standards for enabling electronic healthcare information exchange. Previous studies have shown that FHIR is capable of modeling both structured and unstructured data from electronic health records (EHRs). However, the capability of FHIR in enabling clinical data analytics has not been well investigated. The objective of the study is to demonstrate how FHIR-based representation of unstructured EHR data can be ported to deep learning models for text classification in clinical phenotyping. We leverage and extend the NLP2FHIR clinical data normalization pipeline and conduct a case study with two obesity datasets. We tested several deep learning-based text classifiers such as convolutional neural networks, gated recurrent unit, and text graph convolutional networks on both raw text and NLP2FHIR inputs. We found that the combination of NLP2FHIR input and text graph convolutional networks has the highest F1 score. Therefore, FHIR-based deep learning methods has the potential to be leveraged in supporting EHR phenotyping, making the phenotyping algorithms more portable across EHR systems and institutions.

Introduction

Electronic health record (EHR) data is being increasingly used for conducting clinical and translational research. Large scale research networks such as the electronic Medical Records and Genomics (eMERGE) network[1], Pharmacogenomics Research Network (PGRN)[2], The National Patient-Centered Clinical Research Network (PCORnet)[3], and the UK BioBank[4] have enabled multi-institutional studies using EHR data[5-8].

However, the lack of interoperability of EHR systems is a challenge for healthcare institutions and clinical research centers. Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR)[9] is one of the current data standards for representation of EHR data, and has been adopted by major EHR vendors to enhance data and system interoperability among different EHR implementations. The overall goal of FHIR is to facilitate available, discoverable and interpretable data sharing across institutions. Many research communities and medical centers are supporting the advancement and development of FHIR standards, including i2b2[10], SMART on FHIR[11] and eMERGE[12].

Due to its advantages on implementation readiness and interoperability among different EHR systems, FHIR is increasingly being used for exchanging EHR data. On top of representing normalized structured data, NLP2FHIR[13] has been developed as a data normalization pipeline, which provides a reference implementation of the FHIR standard for modeling unstructured data. A follow-up study was done on computational phenotyping with FHIR-based EHR representation, which demonstrated that NLP2FHIR-based representation of EHR data can effectively identify phenotypes using the case study on patients with obesity and multiple comorbidities from discharge summaries[14, 15]. Machine learning models such as Decision Tree, Support Vector Machine and Random Forest were also tested for effectively identification of obesity and multiple comorbidities using semi-structured information from discharge summaries.

However, little work has been done in standards and clinical research informatics communities on adopting FHIR for deep learning models. In this study, we use existing deep learning methods including convolutional neural networks (CNN)[16], Gated Recurrent Unit (GRU)[17] and Text Graph Convolutional Network (GCN)[18] to demonstrate how FHIR-based data representation can be integrated into deep learning models. We leveraged the NLP2FHIR pipeline and deep learning models on a case study to predict obesity and its comorbidities in two different datasets. We found that the combination of NLP2FHIR input, which is a graph-based input format, and the text graph convolutional networks has the highest F1 score. It shows promises to effectively use NLP2FHIR outputs as an

input standard for deep learning methods in supporting EHR phenotyping, making the phenotyping algorithms more portable across data systems and institutions.

Related Work

Standard-based phenotype algorithms and execution workflow have been studied in the clinical research informatics community to allow implementations of clinical logic and value sets in a modular software architecture[19]. Various machine learning algorithms have been leveraged by the Phenotype Execution and Modeling Architecture (PhEMA) project to identify phenotypes and sub-phenotypes for a number of conditions including acute kidney injury, heart failure, major depression and Alzheimer's disease[20-22]. Rasmussen et al[23] also proposed a framework using a common data model (CDM), standardized representation of the phenotype algorithms logic, and technical solutions to facilitate federated execution of queries. It is envisioned to help guide future research in operationalizing phenotype algorithm portability at scale. Hripcsak et al. described the process of transferring the phenotypes of type 2 diabetes mellitus (T2DM) and attention deficit and hyperactivity disorder (ADHD) to the Observational Medical Outcomes Partnership (OMOP) CDM within the eMERGE network[24].

Standardized preprocessing pipelines for machine learning[25, 26] can enable fair comparisons among machine learning models on publicly available datasets such as MIMIC III[27]. To further standardize these datasets into the clinical interoperable standard of FHIR, one of the representative work is done by Rajkomar et al[28]. They represented EHR data using FHIR and demonstrated that FHIR is capable of medical event prediction when tested on de-identified structured and unstructured EHR data from two US academic medical centers. Sharma et al. have studied a phenotyping system to integrate both rule-based and statistical machine learning methods[29]. The system has leveraged OHDSI's OMOP CDM with Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs) to represent clinical NLP concepts as input features for machine learning based classifiers in phenotype identification systems. Hong et al. has demonstrated FHIR-based EHR phenotyping can be applied to semi-structured discharge summaries for multiple comorbidities identification[15]. The NLP2FHIR implementation contains several different NLP components leveraging existing information extraction systems including cTAKES[30], MedTagger[31], MedXN[32] and UMLS VTS[33]. On the same task, Yao et al. used word embeddings and entity embeddings on CNN adapted from rule-based systems[34], but the system did not leverage any standards or CDMs and hence have limited interoperability.

Materials and Methods

Materials

In this work, we selected two datasets for our analysis: the i2b2 2008 obesity dataset[35] and MIMIC III dataset[27].

The i2b2 2008 obesity dataset is a fully de-identified dataset consisting of discharge summaries. The dataset contains human-curated obesity status explicitly mentioned in the texts as well as 15 comorbidities consisting of asthma, atherosclerotic cardiovascular disease (CAD), congestive heart failure (CHF), depression, diabetes mellitus, hypertension, gastroesophageal reflux disease (GERD), gallstones, hypercholesterolemia, hypertriglyceridemia, obstructive sleep apnea (OSA), osteoarthritis (OA), peripheral vascular disease (PVD), and venous insufficiency[35]. For each comorbidity, there are 4 labels as the prediction target: present, absent, questionable or unmentioned. Both textual and intuitive judgments are provided in the dataset for each patient. While the judgment of textual is based on explicit mentions, the intuitive judgments are based on the annotators' judgment, and may lead to additional inference (e.g. the statement of weights to infer obesity).

The MIMIC III (Medical Information Mart for Intensive Care) is a publicly available dataset containing vital signs, medications, lab test results, observations and clinical notes of 53,423 adult admissions of critical care units. To build an obesity related comorbidity prediction dataset which is similar to the i2b2 dataset, we follow the experiment settings of Hong et al. to validate the design of portability[15]. The obesity and non-obesity groups are selected based on body mass index (BMI) for adult patients. Adults with a BMI value larger than 30 at admission with discharge summaries are categorized in the obesity group, while adults with a BMI value between 18.5 to 24.9 at admission are categorized as the control group. A total of 2000 discharge summaries are randomly selected among all available discharge summary notes, with 1000 each for case and control. 70% of the notes are selected as the training set (n=1400), and 30% of the notes are selected as the test set (n=600).

We present dataset characteristics in Table 1.

Table 1. Statistics of the i2b2 and MIMIC IIIs datasets

Dataset	Train	Test	Average # of words/doc	Vocabulary size	Average # concepts/doc	Extracted concepts
I2b2 2008	719	496	833.4	8633	106.6	1919
MIMIC III	1400	600	1491.1	13129	158.6	3220

Methods

The proposed workflow of FHIR-enabled text classification application for clinical phenotyping is illustrated in Figure 1. Given the original texts from the two datasets, a document is first tokenized into to a list of tokens as the input of the deep learning models. During the preprocessing, stop words and words appearing less than three times are removed for the purpose of better performance in the embedding training phase. Then, FHIR resources in JSON format produced by the NLP2FHIR pipeline are concatenated into token-like representations, which are categorized into different resources (Condition, Procedure, MedicationStatement, and FamilyMemberHistory) and grouped into FHIR Bundles. The NLP2FHIR representation is based on an existing system primarily validated on various data types[36]. Figure 2 shows an example of NLP2FHIR output of the sentence "*Ms. [Name] is a 64-year-old female with nonischemic cardiomyopathy and class II-III symptoms who presented with worsening volume overload*". There are 2 concepts ("*cardiomyopathy*" and "*worsening volume overload*") identified by cTAKES to be normalized to SNOMED CT codes. To make the extracted concept objects compatible with word-based input formats, the coding

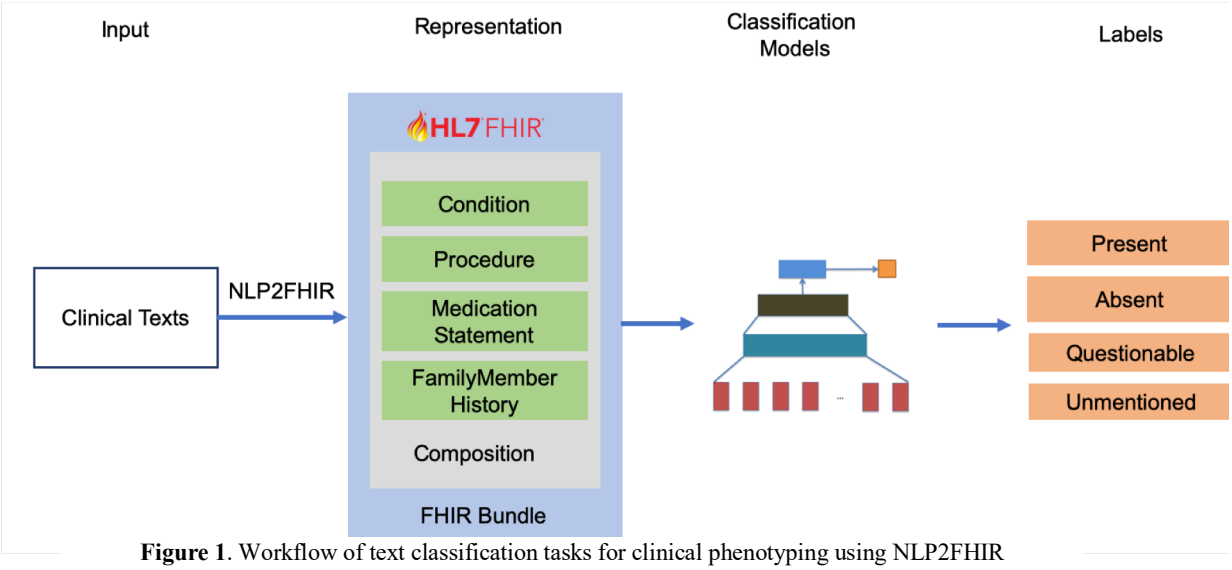


Figure 1. Workflow of text classification tasks for clinical phenotyping using NLP2FHIR

system URL (i.e. http://snomed.info/sct), the code, and the polarity from the "abatementString" field are concatenated into one "word" to represent their uniqueness. As an important factor for the learning and prediction for many machine learning models, although word orders are not supposed to be preserved by FHIR, it is preserved sequentially in the NLP2FHIR output naturally as the dictionary lookup generates sequential concepts as outputs.

After both the texts and NLP2FHIR representations are ready, the data are fed into machine learning/deep learning-based classifiers to classify the documents. We tested three different deep learning models in this study: CNN, GRU and Text GCN. The details of the models in this study are described as follows.

CNN

Convolutional neural networks are one of the earliest and most commonly used deep learning models in text classification tasks[16, 37]. Experiments in the biomedical domain have shown that CNN can achieve good performance without extensive model tuning. In a typical 1-dimensional CNN for text classification tasks, it can capture local contexts by leveraging a convolutional kernel (or filter) acting as a sliding window among tokens.

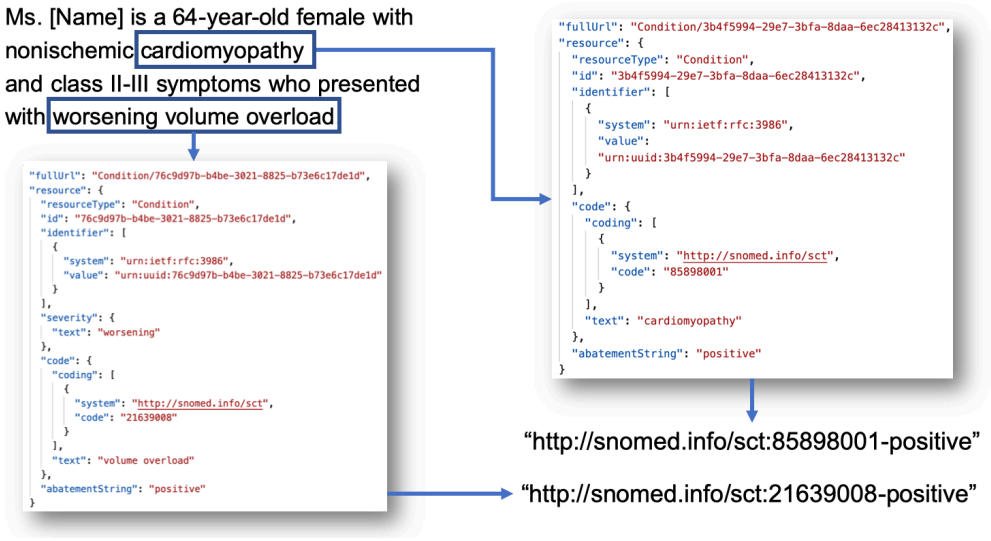


Figure 2. NLP2FHIR JSON representation of a sample clinical text to concept-based representation for deep learning models

In this study, we use a fixed length CNN where the length is a hyperparameter. If the input document is shorter than the expected length, the end of the sequence was padded with zeros. If the input document was longer than the expected length, the input document was truncated.

RNN/GRU

One challenge for CNN is that it does not capture long-term information when the contexts are not close. RNNs are capable to handle long-term patterns in sequential inputs, because the state of previous RNN units can be passed to the units behind them until the end of the sequence. There are multiple variations of RNN[38] which usually have better performance than vanilla RNN, including Long short-term memory (LSTM)[39] or Gated Recurrent Unit (GRU)[17]. Experiments showed that there is no consistency on which model would perform better in general. In our experiments, we selected GRU due to its faster convergence time, and it should not impact our conclusion as the performances of these two models are usually comparable with each other[17].

Text GCN

Kipf et al. proposed GCN [40], a graph neural network architecture for node classification. GCN is one of the methods to generalize neural networks to structured datasets. While CNNs or RNNs are typically good at modeling “array-like” input data, they will face challenges to model graphs as the data connections are more challenging to capture. Common characteristics like depths, degrees, density, or node connectivity cannot be easily modeled without adapting to graph specific models.

To make GCN work better for text classification tasks, Text GCN is proposed by Yao et al. as an extension of GCN[18]. Text GCN uses words and documents as nodes, and uses the trained embedding to classify document nodes into categories. The major differences between GCN and Text GCN is how the edges in the graph are represented.

There are 2 types of nodes in Text GCN: document nodes and token nodes. When a word appears in a document, an edge between the document node and the token node will be generated. Each element of adjacent matrix A which keeps the edge weights between two nodes (m and n) are defined as follows:

$$A_{m,n} = \begin{cases} \text{PMI}(m,n), & m,n \text{ are nodes (concept/word)} \\ \text{TF-IDF}_{m,n}, & m \text{ is document, } n \text{ is node (concept/word)} \\ 1, & m = n \\ 0, & \text{Otherwise} \end{cases}$$

The PMI (point-wise mutual information) given a word pair *m*, *n* can be calculated by:

$$PMI(m,n) = \log \frac{p(m,n)}{p(m)p(n)}$$

$$p(m,n) = \frac{\#C(m,n)}{\#C}$$

$$p(m) = \frac{\#C(m)}{\#C}$$

where $\#C(m)$ is the count of the sliding windows with the token *m* from the whole corpus, $\#C(m, n)$ is the count of sliding windows containing both *m* and *n*, and $\#C$ is the count of windows in the corpus. Only the positive PMI edges are added, since a negative PMI means there is little to no correlation of the words within the corpus. The constructed graph is then fed into a 2-layer GCN for training as proposed in Yao el al.[18] and Kipf and Welling[40], which allows messages passing among different nodes and layers.

The illustration of how the Text GCN graph can be adapted to the token and NLP2FHIR representations is shown in Figure 3. As a comparison, the phrase "acute renal failure" is normalized to a Condition concept with a SNOMED CT code 14669001.

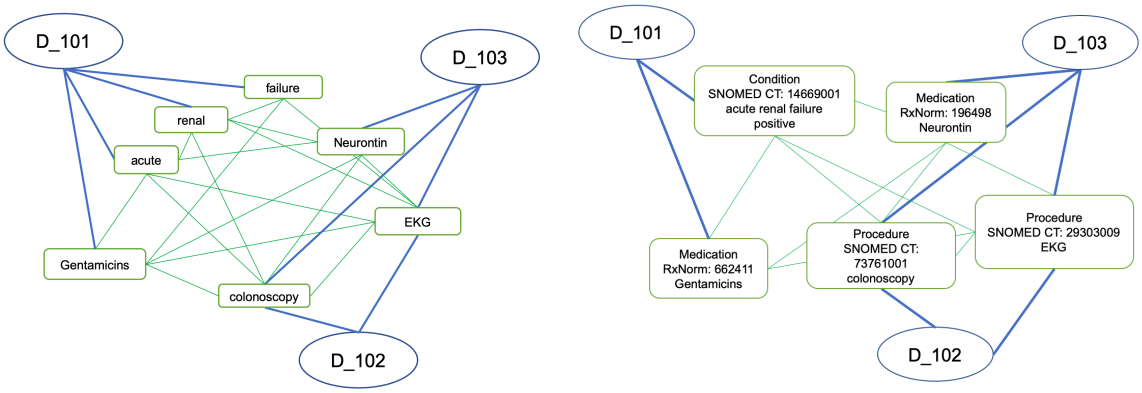


Figure 3. Token-based (left) and FHIR-concept-based (right) graphs for Text GCN text classification. The nodes denoted by circles are document nodes with document IDs, and the nodes denoted by round rectangles are token (left) or concept (right).

Results

In this section, we compare the performances of original tokens with the NLP2FHIR representations, used in different deep learning methods (CNN, GRU and Text GCN) in clinical text classification tasks. In the i2b2 obesity dataset, we use the official training and test set for evaluation, while the training and testing set of the MIMIC III dataset in this study is split by a ratio of 70% and 30% from the randomly selected notes described in the Materials section. All discharge summaries are flattened as lists of lower-case tokens with all the line-breaks removed before entering into the deep learning models.

The Text GCN implementation is adopted from Yao et al[41] and is implemented by scikit-learn[42] and TensorFlow[43]. The CNN and GRU implementations are based on Keras[44] using a TensorFlow backend. All the embedding layers are trained on the training set, and no pre-trained word embedding models are used in the experiments. The hyperparameters are tuned for different datasets separately. For the CNN model, we used 1 convolutional layer

before 1 fully connected layer with the number of filters as 200, maximum input length as 600, the embedding dimension as 50, and the convolution kernel size as 3. For GRU, the hidden dimension is set to 128. The number of epochs for both the CNN and GRU are 40, with an early stopping patience of 5 monitoring the validation loss. The validation set consists of 10% of the training data. The source code of the implementation can be found at https://github.com/BD2KOnFHIR/nlp2fhir-deep-learning.

For Text GCN, only text data is used, and for FHIR, the concepts consist of code and polarity (positive, negative). It is transformed to a multi-class document classification problem. The graph statistics of the i2b2 and MIMIC datasets are shown in Table 2. We used the default settings of 2 GCN layers as it shows better performance than the 1-layer model, and it is more likely to converge compared to 3 or more layer models in our early experiments.

Table 2. Statistics of the Text GCN graph for i2b2 and MIMIC datasets

Dataset	# of nodes	# of edges	Density
I2b2 2008	6134	79598	4.23 * 10e-3
MIMIC III	9204	168143	3.97 * 10e-3

The reported performances are the accuracy (equivalent to micro-precision, recall and F1-score), macro-averaged precision, recall and F1-score of the obesity and its 15 comorbidities and its 95% confidence intervals (95% CI) among different comorbidities. Table 3 and 4 show the mean macro-averaged precision, recall and F1-scores among obesity and different comorbidities. For the i2b2 dataset, we used textual gold labels instead of intuitive, which is more relevant to show how models understand the contexts without additional inferences by human experts.

Table 3. Performances of different experiment settings on the i2b2 dataset by accuracy, macro averaged precision, recall and F1-score with the 95% CI. The highest scores are on bold

Dataset	CNN	CNN- FHIR	LSTM	LSTM-FHIR	Text GCN	Text GCN - FHIR
Accuracy	0.737 ± 0.068	0.748 ± 0.068	0.652 ± 0.075	0.697 ± 0.075	0.707 ± 0.055	**0.795** ± 0.044
Precision	0.489 ± 0.080	0.493 ± 0.058	0.400 ± 0.074	0.520 ± 0.064	0.347 ± 0.057	**0.525** ± 0.049
Recall	0.504 ± 0.055	0.519 ± 0.055	0.457 ± 0.055	0.478 ± 0.056	0.500 ± 0.056	**0.523** ± 0.045
F1-score	0.495 ± 0.060	0.505 ± 0.056	0.415 ± 0.059	0.495 ± 0.061	0.410 ± 0.057	**0.524** ± 0.048

Table 4. Performances of different experiment settings the MIMIC III dataset by accuracy, macro averaged precision, recall and F1-score with the 95% CI. The highest scores are on bold

Dataset	CNN	CNN-FHIR	LSTM	LSTM-FHIR	Text GCN	Text GCN - FHIR
Accuracy	0.859 ± 0.080	0.857 ± 0.083	0.824 ± 0.083	0.873 ± 0.079	0.746 ± 0.077	**0.914** ± 0.070
Precision	**0.628** ± 0.052	0.607 ± 0.036	0.587 ± 0.041	0.614 ± 0.069	0.388 ± 0.076	0.616 ± 0.044
Recall	0.645 ± 0.053	0.638 ± 0.021	0.596 ± 0.001	0.643 ± 0.057	0.623 ± 0.037	**0.721** ± 0.058
F1-score	0.625 ± 0.052	0.616 ± 0.036	0.590 ± 0.025	0.622 ± 0.057	0.478 ± 0.047	**0.664** ± 0.050

From the experiments, we observe that NLP2FHIR representations provide better performances when used as input compared to the original texts. In most cases the use of FHIR representation have positive impacts on classification performance, with the CNN vs CNN-FHIR on MIMIC III dataset the exception in our experiments. CNN models are

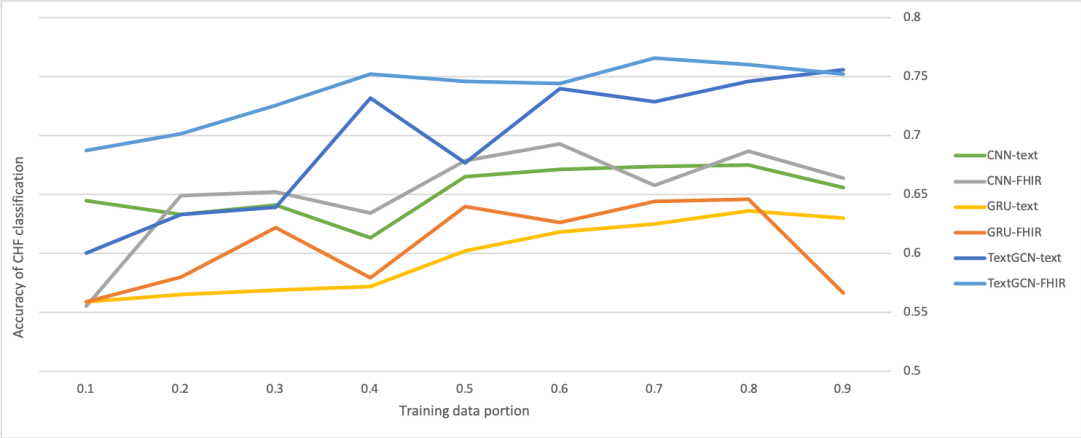


Figure 4. The impact of proportions of data into training on one of the comorbidity classifications (Congestive Heart Failure, CHF). The x-axis is the ratio of training data used from the all labeled data (training + testing), and the y-axis is the accuracy of CHF as an example of the comorbidities.

one of the strong baseline models for text classification on raw texts, with little to none preprocessing needed, in many studies. Therefore, applying information extraction pipeline (dictionary lookup) on texts may not lead to favorable performances on CNN models.

Text GCN, which is a graph-based algorithm, also outperforms other deep learning models. The main reason is that the data we tested are very sparse. Unlike other text classification tasks presented in the Text GCN experiments, such as movie reviews or abstracts, only a few tokens are related to the classification results. That results in difference in density of the graphs.

We also experimented with multiple settings with different portions of training data. The impacts of accuracy in one of the comorbidities (CHF) on the deep learning models are shown in Figure 4. We can observe increasing trends in general from left (fewer training samples) to right (more training samples), meaning increasing the amount of data into training while reducing the amount of data into testing. However, the trend is not obvious when the proportion is larger than 0.5, indicating the amount of training data can be considered sufficient to learn the hidden patterns in a fully annotated dataset. After the amount of data reaching the threshold, the model may at risk of overfitting that may have negative impact on the generalizability of the trained models due to the lack of generalizable test samples.

Discussion

In this paper, we designed and experimented with token-based and NLP2FHIR representations for text classification models. The tested models represent three different type of information for classification: CNN primarily classifies texts based on collections (max pooling) of local contexts, RNN on the actual sequences and Text GCN on graph structures of the tokens or concepts. The experimental results show that the sequence of normalized concept models from FHIR representation is better than the input data from the raw texts, or sequence of tokens, when applied to vanilla deep learning models without feature extraction and feature engineering. One potential reason for that is the contribution of the normalized representations that may be more informative comparing with sequences of tokens. With normalized concepts, the input sequence of the model is more condensed and standardized with potentially more edges in the graph. Another advantage of migrating data into the FHIR representation is the implementations and toolkits available through open-source FHIR development efforts. For the standardized implementation with improved portability and interoperability, deep learning applications can be deployable with minimal efforts across different datasets and systems.

One usage of the proposed standard-based design is to allow de-identified data sharing regarding protected health information (PHI). The FHIR elements will only contain higher-level concepts from clinical ontologies and knowledge bases. As general concepts (separated from any specific patient), they are intrinsically free of PHI as defined by Health Insurance Portability and Accountability Act (HIPAA)[45] that may make the data identifiable. The NLP2FHIR representation includes the annotated clinical mentions with normalized entities that are expected to be

PHI-free. This can further inspire more pilot studies on distributing NLP2FHIR representations without the original texts as a standard format to facilitate standard-based phenotype identification algorithms without sharing PHI or de-identification efforts.

There are also several limitations in this study. First, for the text classification problems, we only demonstrate a few models as a case study, and it is not an exhaustive evaluation to determine the best performing ML methods. As comparisons, our overall macro F1 score of 0.524 in the i2b2 dataset is higher than the decision tree on CUI performances (0.5121)[29] but lower than 0.6578 when a decision tree classifier is used including section information in Hong et al[15]. Likewise, other studies have demonstrated better performance than that in our experiments, although we note they were conducted based on different pre-processing steps and experimental settings. For instance, many top systems from the i2b2 challenge filtered the discharge summaries that are not relevant to the patients and developed keyword-based approaches to identify comorbidities[46-48], which contain hand-crafted rules and regular expressions that are not portable. However, our major goal in this study was to demonstrate the portability of deep learning models when applied to NLP2FHIR representations. Therefore, we did not work towards building corpus-specific dictionaries or rules, as such efforts and models tend to overfit to a specific task or corpus.

Second, the implementation and evaluation did not utilize any document or textual structures. The current structure represents the document structure such as sections or sentences, but in the experiments, we did not weigh in the structure due to the lack of sentence-level and section-level gold standard labels.

Third, the semantic based representations may not fully utilize syntax-based features that may be helpful for phenotype classification[49-51], because the sentence structural information is omitted by the concepts and thus are not retained in the FHIR based representations. This makes it challenging to apply NLP2FHIR outputs for contextual pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers)[52] and RoBERTa[53], which are intended to handle natural languages as neural language models rather than coded phenotypic representations. This can cause some contextual information eliminated before feeding into the neural models for the classification task.

Conclusion

NLP2FHIR outputs can be ported and integrated into deep learning methods. We found that the classification results of NLP2FHIR based methods outperformed the methods with original texts. We demonstrated that FHIR-based deep learning methods could be leveraged in supporting EHR phenotyping, making the phenotyping algorithms more portable across data systems and institutions. In the future, we will work on improving the performance by adding document structure such as sentences and sections into the document modeling.

Acknowledgment

Research reported in this publication was supported by National Institutes of Health under the awards FHIRCat (R56EB028101), BD2K (U01 HG009450), and PhEMA (R01 GM105688). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Reference

1. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genetics in Medicine. 2013;15(10):761-71.
2. Pharmacogenomics Research Network.
3. PCORnet: the National Patient-Centered Clinical Research Network.
4. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-9.
5. Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, Sacchi L. Secondary Use and Analysis of Big Data Collected for Patient Care. Yearb Med Inform. 2017;26(1):28-37.
6. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. Journal of biomedical informatics. 2014;52:28-35.
7. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annual Symposium proceedings AMIA Symposium. 2006;2006:1040-.

8. Benincasa G, Marfella R, Della Mura N, Schiano C, Napoli C. Strengths and Opportunities of Network Medicine in Cardiovascular Diseases. Circulation Journal. 2020;84(2):144-52.
9. FHIR Overview [Available from: https://www.hl7.org/fhir/overview.html.
10. i2b2: Informatics for Integrating Biology & the Bedside [Available from: https://www.i2b2.org/.
11. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. Journal of the American Medical Informatics Association. 2016;23(5):899-908.
12. Taylor CO, Lemke KW, Richards TM, Roe KD, He T, Arruda-Olson A, et al. Comorbidity Characterization Among eMERGE Institutions: A Pilot Evaluation with the Johns Hopkins Adjusted Clinical Groups® System. AMIA Jt Summits Transl Sci Proc. 2019;2019:145-52.
13. NLP2FHIR: A FHIR-based Clinical Data Normalization Pipeline and Its Applications [Available from: https://github.com/BD2KOnFHIR/NLP2FHIR.
14. Hong N, Wen A, Shen F, Sohn S, Wang C, Liu H, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open. 2019;2(4):570-9.
15. Hong N, Wen A, Stone DJ, Tsuji S, Kingsbury PR, Rasmussen LV, et al. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform. 2019;99:103310.
16. Kim Y, editor Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 oct; Doha, Qatar: Association for Computational Linguistics.
17. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014.
18. Yao L, Mao C, Luo Y, editors. Graph convolutional networks for text classification. Proceedings of the AAAI Conference on Artificial Intelligence; 2019.
19. Rasmussen LV, Kiefer RC, Mo H, Speltz P, Thompson WK, Jiang G, et al. A Modular Architecture for Electronic Health Record-Driven Phenotyping. AMIA Joint Summits on Translational Science proceedings AMIA Joint Summits on Translational Science. 2015;2015:147-51.
20. Xu Z, Feng Y, Li Y, Srivastava A, Adekkanattu P, Ancker JS, et al. Predictive Modeling of the Risk of Acute Kidney Injury in Critical Care: A Systematic Investigation of The Class Imbalance Problem. AMIA Jt Summits Transl Sci Proc. 2019;2019:809-18.
21. Xu Z, Luo Y, Adekkanattu P, Ancker JS, Jiang G, Kiefer RC, et al. Stratified Mortality Prediction of Patients with Acute Kidney Injury in Critical Care. Stud Health Technol Inform. 2019;264:462-6.
22. Xu Z, Chou J, Zhang XS, Luo Y, Isakova T, Adekkanattu P, et al. Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks. J Biomed Inform. 2020;102:103361.
23. Rasmussen LV, MS, Brandt PS, MSc, Jiang G, MD PhD, Kiefer RC, Pacheco JA, MS, Adekkanattu P, PhD, et al., editors. Considerations for Improving the Portability of Electronic Health Record-Based Phenotype Algorithms. AMIA Annual Symposium 2019; 2019; Washington DC, USA.
24. Hripcsak G, Shang N, Peissig PL, Rasmussen LV, Liu C, Benoit B, et al. Facilitating phenotype transfer using a common data model. Journal of Biomedical Informatics. 2019;96:103253.
25. Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. J Am Med Inform Assoc. 2020;27(12):1921-34.
26. Wang S, McDermott MBA, Chauhan G, Ghassemi M, Hughes MC, Naumann T. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. Proceedings of the ACM Conference on Health, Inference, and Learning; Toronto, Ontario, Canada: Association for Computing Machinery; 2020. p. 222–35.
27. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035.
28. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018;1(1):18.
29. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, et al. Developing a portable natural language processing based phenotyping system. BMC Medical Informatics and Decision Making. 2019;19(3):78.

30. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010;17(5):507-13.
31. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. Journal of the American Medical Informatics Association : JAMIA. 2011;18(5):580-7.
32. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. Journal of the American Medical Informatics Association : JAMIA. 2014;21(5):858-65.
33. UMLS Vocabulary and Terminology Service. 2018.
34. Yao L, Mao C, Luo Y. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Medical Informatics and Decision Making. 2019;19(3):71.
35. Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009;16(4):561-70.
36. Hong N, Wen A, Shen F, Sohn S, Liu S, Liu H, et al. Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data. AMIA Jt Summits Transl Sci Proc. 2018;2017:74-83.
37. Rios A, Kavuluru R. Convolutional Neural Networks for Biomedical Text Classification: Application in Indexing Biomedical Articles. ACM BCB. 2015;2015:258-67.
38. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-6.
39. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735-80.
40. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907. 2016.
41. Graph Convolutional Networks for Text Classification. AAAI 2019 [Available from: https://github.com/yao8839836/text_gcn.
42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12(Oct):2825-30.
43. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016.
44. Keras 2015 [Available from: https://keras.io/.
45. (OCR) OfCR. Summary of the HIPAA Security Rule 2013 [Available from: https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.
46. Yang H, Spasic I, Keane JA, Nenadic G. A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries. Journal of the American Medical Informatics Association. 2009;16(4):596-600.
47. Childs LC, Enelow R, Simonsen L, Heintzelman NH, Kowalski KM, Taylor RJ. Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. Journal of the American Medical Informatics Association : JAMIA. 2009;16(4):571-5.
48. Mishra NK, Cummo DM, Arnzen JJ, Bonander J. A rule-based approach for identifying obesity and its comorbidities in medical discharge summaries. Journal of the American Medical Informatics Association : JAMIA. 2009;16(4):576-9.
49. Komninos A, Manandhar S, editors. Dependency based embeddings for sentence classification tasks. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies; 2016.
50. Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. Journal of the American Medical Informatics Association. 2014;21(5):824-32.
51. Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, et al., editors. Abstract meaning representation for sembanking. Proceedings of the 7th linguistic annotation workshop and interoperability with discourse; 2013.
52. Devlin J, Chang M-W, Lee K, Toutanova K, editors. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019 jun; Minneapolis, Minnesota: Association for Computational Linguistics.
53. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv. 2019;abs/1907.11692.

Extracting Adverse Drug Events from Clinical Notes

Darshini Mahendran and Bridget T. McInnes, Ph.D.
Computer Science Department, Virginia Commonwealth University, Richmond, VA, USA

Abstract

Adverse drug events (ADEs) are unexpected incidents caused by the administration of a drug or medication. To identify and extract these events, we require information about not just the drug itself but attributes describing the drug (e.g., strength, dosage), the reason why the drug was initially prescribed, and any adverse reaction to the drug. This paper explores the relationship between a drug and its associated attributes using relation extraction techniques. We explore three approaches: a rule-based approach, a deep learning-based approach, and a contextualized language model-based approach. We evaluate our system on the n2c2-2018 ADE extraction dataset. Our experimental results demonstrate that the contextualized language model-based approach outperformed other models overall and obtain the state-of-the-art performance in ADE extraction with a Precision of 0.93, Recall of 0.96, and an F_1 score of 0.94; however, for certain relation types, the rule-based approach obtained a higher Precision and Recall than either learning approach.

1 Introduction

Adverse drug events (ADE) are unexpected incidents or accidents related to the administration of a drug and its attributes. It includes overdoses, allergic reactions, drug interactions, and medication errors. ADEs account for an estimated 30% of all hospital adverse events[1], and the cost of managing ADEs can be high because the clinical diagnosis of an ADE more often requires additional laboratory tests or procedures to investigate the cause of a patient's symptoms. An ADE may cause a prolonged length of stay in the hospital and increase the economic burden[2]. Some conditions can be caused by undiscovered ADEs, which can increase the costs and risks further and impact the patient economically and mentally. This can be prevented if we identify the potential reason and the information regarding the drug on time. Extracting relations from scientific publications and clinical narratives has always been challenging due to the complexity of language and domain-specific knowledge involved[3]. Processing the information from these clinical narratives, which records patient medication history, known allergies, reactions, and adverse events of the patient, allows a more thorough assessment of potential ADEs before they happen[1].

If we can extract all possible interactions of a drug and warn the patient when prescribing the drug, this would reduce the risks of an ADE taking place[1]. ADEs are a world-wide health-related concern, and therefore, a considerable amount of research and effort is dedicated to identifying possible ADEs in different instances[4]. Information on drugs, its attributes, and associated ADEs are stored in many databases; however, we require information about not just the drug itself but attributes describing the drug (e.g., strength, dosage), the reason why the drug was initially prescribed (e.g., reason), and the relation between the drug and its attributes.

However, manual extraction of ADEs is almost impossible[5] given the amount of data gathered every year, therefore, there is an urgent need for automated systems for ADE extraction. These data are more often unstructured and Natural Language Processing (NLP) techniques are utilized for this significant task to extract ADEs from the unstructured text. Relation Extraction (RE) is a sub-field of NLP whose goal is to detect and classify relations between entities in a text. In this work, we explore three RE approaches for ADE extraction: a rule-based approach utilizing co-location information, a deep learning-based approach utilizing Convolutional Neural Networks (CNNs), and a contextualized language model-based approach utilizing Bidirectional Encoder Representations from Transformers (BERT). We evaluate our system on the n2c2-2018 ADE extraction dataset. Our experimental results demonstrate that the contextualized language model-based approach outperformed other models overall and obtained state-of-the-art performance in ADE extraction with a Precision of 0.93, Recall of 0.96, and an F_1 score of 0.94; however, the rule-based system obtained a higher Precision and Recall for certain relation types.

The remainder of this paper is structured as follows. First, we discuss the previous works done in this area of research. Second, we describe the dataset we use to evaluate our system. Third, we describe our three approaches. Fourth, we present and analyze the results. Fifth, we conduct a comparison between our approaches and previous work. Finally,

we present the conclusions we derive from this work and what we plan to do in the future.

2 Related Work

Adverse drug event (ADE) extraction is gaining attention recently among the clinical NLP community. Many approaches have been explored and can be divided into four paradigms: 1) rule-based, 2) machine learning-based 3) deep learning-based, and 4) contextualized language model-based approaches.

Rule-based approaches. These systems use specified rules and patterns to extract the information from texts. Li, et al.[6], used a rule-based method to link drug names with their attributes. They used a string-based regular expression matching to match the drug names to a prescription list and then used the co-location information and RxNorm dictionary to determine whether they matched.[1].

Machine learning-based approaches. Traditional supervised machine learning systems utilize large amounts of the annotated corpus for training. Previous works have utilized learning algorithms such as Support Vector Machines (SVMs)[7] and Random Forests (RFs)[1]. Miller, et al.[7] used an SVM-based system and a neural system obtain for RE based on previous work on extracting temporal narrative container relations from sentences[8]. Yang, et al.[9] first applied heuristic rules to generate candidate pairs and then applied ML models to classify the relations. They divided the relations into different groups according to their cross-distance - defined as the number of sentence boundaries between the two entities and developed multiple classifiers to classify relations according to their cross-distance.

Deep learning-based approaches. These systems utilize multi-layer neural networks typically with featureless embedding representations such as word embeddings[10]. Recent work has explored using variations of Recurrent Neural Network (RNN) architectures. Xu, et al.[11] proposed a cascaded sequence labeling approach to recognize the entities and the relations simultaneously. Sorokin, et al.[12] proposed using a Long Short Term Memory (LSTM)-based encoder to jointly learn representations for all relations in a single sentence. Henry, et al.[1] summarizes the work of participants in the n2c2-2018 challenge who proposed an attention-based piecewise bidirectional (bi-) LSTM with standard features and unique candidate pair generation. Christopoulou, et al.[13] developed separate models for intra- and inter-sentence relation extraction and combined them using an ensemble method. The intra-sentence models use biLSTMs with attention mechanisms to capture dependencies between multiple related pairs in the same sentence. For the inter-sentence relations, they used the transformer network to improve performance for longer sequences. Research on RNNs and its variants has been studied however there are few exploring CNN architectures. Therefore, in this work, we explore CNN based architectures for relation extraction.

Contextualized language model-based approaches. Pre-trained contextualized language models have been shown to increase the performance for several NLP tasks. Wei, et al.[14] and Alimova, et al.[15] applied pre-trained language models of BERT to the ADE RE task. Wei, et al.[14] developed two BERT-based methods: Fine-Tuned BERT (FT-BERT) and Feature Combined BERT (FC-BERT) to determine relation categories for these candidate pairs. For the FT-BERT models, they represent a candidate relation pair in an input sentence by replacing the entity with its semantic type, and they added a linear classification layer on the top of The BERT model to predict the labels. For the FC-BERT, they represented the entities using the entity tags. Alimova, et al.[15] proposed a machine learning model with a novel set of knowledge-based and BioSentVec embedding[16] features. For comparison, they utilized three BERT-based models: BERT_uncased, BioBERT, and Clinical BERT. They utilized the entity texts combined with a context between them as an input for the BERT-based models.

3 Dataset

We evaluate our approaches on the National NLP Clinical Challenges (n2c2) 2018 Adverse Drug Event Dataset[1]. The dataset contains ADE mentions, drug-related attributes, and drug-related relations from 505 patient discharge summaries drawn from the MIMIC-III database [17]. It consists of nine entity types (Drug, Strength, Route, Form, ADE, Dosage, Reason, Frequency) and eight relations between the drug entity and other non-drug entity types. Table 1 shows the number of relations in the training and test data. We use the gold annotated entities of this dataset for RE.

[1] https://www.nlm.nih.gov/research/umls/rxnorm/index.html

Table 1: Relation type statistics of n2c2 2018 data sets.

n2c2 dataset		
Relation	**# Train instances**	**# Test instances**
Strength-Drug	6702	4244
Duration-Drug	643	426
Route-Drug	5538	3546
Form-Drug	6654	4374
ADE-Drug	1107	733
Dosage-Drug	4225	2695
Reason-Drug	5169	3410
Frequency-Drug	6310	4034

4 Methods

In this work, we explore several approaches for ADE extraction: rule-based approaches, two deep learning-based approaches, and a contextualized language model-based approach. The remainder of this section describes the systems in detail.

4.1 Rule-based approach

In our rule-based system, we utilize the co-location information between the drug and the non-drug entity types to determine if the non-drug entity is referring to the drug. We use a breadth-first search algorithm to find the closest occurrence of the drug on either side of the non-drug entity. For each non-drug entity, we traverse both sides until the closest occurrence of the drug is found based on the provided span values of the entities. We explore four traversal mechanisms and report the best traversal mechanism in our results: 1) traverse left-only, 2) traverse right-only, 3) traverse left-first-then-right, and 4) traverse right-first-then-left. This was conducted in two modes: 1) limiting the traversal to only a single relation per relation type (bounded), or 2) allowing for a drug to be linked to multiple entity types with the same relation (unbounded). For example, the sentence *Once her hematocrit stabilized, she was started on a heparin gtt with coumadin overlap.* contains a non-drug entity, *gtt (Route)* and two drugs *Heparin* and *Coumadin*. The non-drug entity has a relation with the closest drug occurrence Heparin but not with Coumadin when applying the left-only traversal mechanism.

4.2 Deep learning-based approach

Here, we describe the CNN architectures used in this work. CNNs consist of four main layers[18]: embedding, convolution, pooling, and feed-forward layers. Initially, the convolution layer which is a filter learns using the backpropagation algorithm and extracts features from the input. Then the max-pooling layer uses the position information and helps to extract the most significant features from the output of the convolution filter. Finally, the feed-forward layer uses a softmax classifier that performs classification. CNNs take pre-trained word vectors obtained from an external resource as input. Here, we explore two word embedding types: word2vec[10] and GloVe[19]. We treat the RE task as a binary classification task building a separate model for each drug-entity type to determine whether a relation exists between two entities.

Sentence CNN. In this architecture, for each drug-entity pair, we extract the sentence containing the relation and feed it into a CNN where each word in the sentence is represented as a vector embedding. We then apply the convolution layer to learn the local features from the embedding vectors and then the max-pooling layer to extract the most important features from the sentence. Finally, the vector is fed into a softmax (fully-connected) layer to perform the classification. The classification error is then back-propagated, and the model is re-trained until the loss is minimized. Figure 1 shows an illustration of the Sentence CNN architecture.

Segment-CNN. In this architecture, the sentence is divided into segments and trained by separate convolutional units. First, we extract the sentence containing the relation, and we divide it into five segments: 1) preceding - tokenized words before the first concept; 2) concept 1 - tokenized words in the first concept; 3) middle - tokenized words between

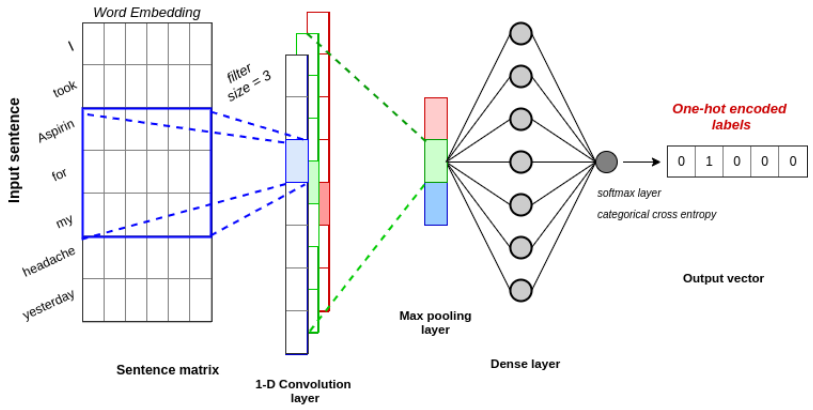


Figure 1: An illustration of our model for Sentence-CNN. It explains the process of both single label and multi-label sentence CNN

the two concepts 4) concept 2 - tokenized words in the second concept; and 5) succeeding - tokenized words after the second concept. Figure 2 explains how an extracted input sentence is divided into five segments.

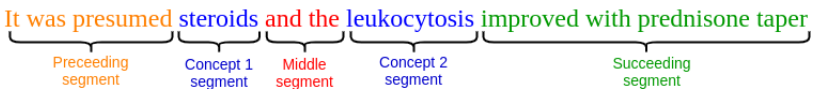


Figure 2: An example of an input sentence that illustrates how the segmentation is done

We construct separate convolution units for each segment and concatenate them before we feed the fixed-length vector into the dense layer that performs the classification. Each convolution unit applies a sliding window that processes the segment and feeds the output to the max-pooling layer to extract essential features independent of their location. The output features of the max-pooling layer of each segment are then flattened and concatenated into a vector before feeding it into the fully connected feed-forward layer. The vector is finally fed into a softmax layer to perform the classification. Figure 3 shows an illustration of Segment-CNN architecture.

4.3 Contextualized language model-based approach

In this approach, we explore using Bidirectional Encoder Representations from Transformers (BERT)[20] contextualized embeddings into a simple feed-forward neural network. We first extract the sentence containing the relation and pass it through a pre-trained BERT model. The output is then fed into a dropout layer and then into a fully-connected dense layer for classification. As with our deep learning-based approaches, we treat the RE as a binary classification task building a separate model for each drug-entity type. We explore the following BERT-based language models:

- *BERT[20] (-cased and -uncased)*. The original BERT models are trained on a large corpus of English data: BookCorpus (800M words) and Wikipedia (2,500M words) in a self-supervised manner (without human annotation). BERT-based models are smaller BERT models intended for environments with limited computational resources. BERT_uncased and BERT_cased have 2-heads, 12-layers, 768-hidden units/layer, and a total of 110 M parameters.
- *BioBERT[21]*. This model is initialized with the general BERT and further trained over a corpus of biomedical research articles from PubMed[2] abstracts and PubMed Central[3] article full texts.
- *Clinical BERT[22]*. This model is initialized with BioBERT and further fine-tuned over the Medical Information Mart for Intensive Care-III[17] (MIMIC-III) clinical note corpus.

[2]https://www.ncbi.nlm.nih.gov/pubmed/
[3]https://www.ncbi.nlm.nih.gov/pmc/

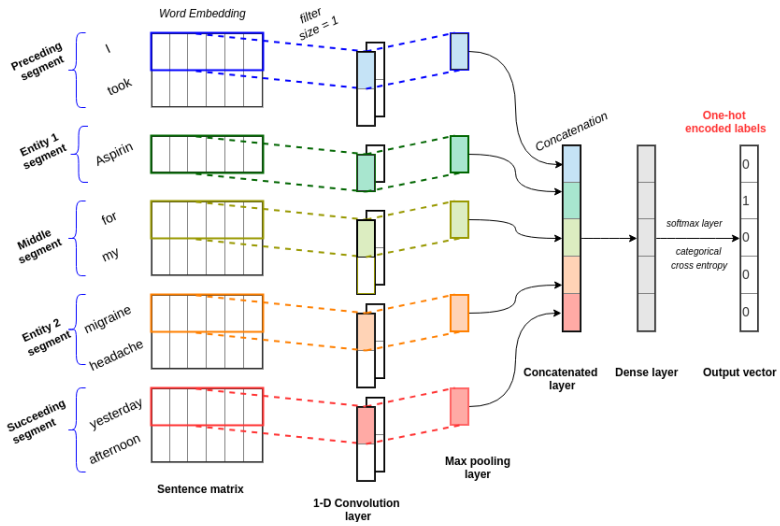


Figure 3: An illustration of our model for Segment-CNN

5 Experimental design

Word representation. For our deep learning-based approaches we use Word2Vec[10] and GloVe[19] representations. The Word2Vec algorithm is trained over the Medical Information Mart for Intensive Care (MIMIC-III), an openly available dataset developed by the MIT Lab for Computational Physiology, comprising 2 million clinical notes from nearly 40,000 critical care patients. The gloVe is trained over Wikipedia (2014) and Gigaword 5.

Text tokenization and vectorization. For the rule-based and the deep learning-based approaches, we use SpaCy tokenizer[4] and Keras tokenizer[23]. For the contextualized language model-based approaches, we use BertTokenizer and AutoTokenizer[22].

Hyper-parameters. We define our model training hyper-parameters by adjusting the batch size, learning rate, and the number of epochs. We use the batch size of 512, rmsprop optimizer with the learning rate of 0.001, and train for 10-20 epochs for our deep learning-based approach. We used the HuggingFaceTransformers[5] to build the BERT model for our contextualized-learning-base approach with Tensorflow 2.0. We use TFRecord to read data into a Dataset object efficiently. We use SparseCategoricalCrossentropy as the loss function and Adam as the optimizer to minimize the loss function.

6 Evaluation criteria

We evaluate our approaches using Precision (P), Recall (R), and F_1 score (F). Precision calculates out of all instances how many instances are predicted correct, and Recall calculates out of all the correct instances that should have been predicted how many instances are correctly predicted. F_1 score is the harmonic mean of Precision and Recall. We also report the micro and macro averages of the system performance. Micro average calculates metrics globally by counting the total true positives, false negatives, and false positives, whereas macro average calculates metrics for each label and the unweighted mean as it does not take class imbalance into account.

7 Results and Discussion

In this section, we describe the results of our three approaches, discuss the results across our three models, and compare previous work.

[4]https://spacy.io/api/tokenizer
[5]https://huggingface.co/transformers/

7.1 Individual model results

Rule-based approach results. Table 2 shows the Precision, Recall, and F_1 scores for our rule-based approach on the test set of the n2c2-2018 dataset for the top three traversal mechanisms described in our method section. Analysis of the various traversal mechanisms over all the non-drug entities showed that the *Left-only* traversal mechanism obtained the best results except for the entity-drug pair Duration-Drug, ADE-Drug, and Reason-Drug. Using the *Left-Right (unbounded)* traversal mechanism obtained the highest F_1 score for these three entities. This is mainly because all other drug attributes are usually mentioned before the drug entity mentions, but the Duration, Reason, and ADE are usually mentioned after the drug mentions. Overall, This approach achieved an overall Precision of 0.88, Recall of 0.83, and F_1 score of 0.86.

Table 2: Results for our rule-based approaches over the n2c2-2018 test set

	Left-only			**Left-Right (unbounded)**			**Left-Right (bounded)**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Strength-Drug	0.96	0.95	**0.95**	0.46	0.90	0.61	0.94	0.94	0.94
Duration-Drug	0.78	0.69	**0.73**	0.58	0.74	0.65	0.46	0.41	0.43
Route-Drug	0.90	0.89	**0.89**	0.45	0.64	0.53	0.37	0.36	0.37
Form-Drug	0.98	0.98	**0.98**	0.62	0.63	0.63	0.67	0.66	0.67
ADE-Drug	0.46	0.39	0.43	0.55	0.75	**0.64**	0.60	0.51	0.55
Dosage-Drug	0.89	0.89	**0.89**	0.61	0.57	0.59	0.89	0.88	0.89
Reason-Drug	0.48	0.35	0.41	0.61	0.57	**0.59**	0.39	0.28	0.33
Frequency-Drug	0.98	0.98	**0.98**	0.39	0.62	0.48	0.10	0.10	0.10
System (Micro)	0.88	0.83	**0.86**	0.50	0.67	0.57	0.56	0.53	0.55
System (Macro)	0.85	0.80	**0.83**	0.61	0.70	0.63	0.58	0.53	0.55

The results indicate that for most entity-drug pairs, co-location information is sufficient to identify most relations. However, the performance of the entity-drug pair ADE-Drug and Reason-Drug are lower compared to the other relation types. Our supposition for this is that the co-location information was insufficient to identify the correct ADE or Reason when multiple drugs were in the same sentence. For example, in the sentence *"Since no new infection was found this was presumed steroids and the leukocytosis improved with prednisone taper."* the non-drug entity *leukocytosis* (ADE) is associated with both *steroids* (Drug) and *prednisone* (Drug).

Deep learning-based results. Table 3 shows the Precision (P), Recall (R) and F_1 scores for our Segment-CNN and Sentence CNN models over the n2c2-2018 test set. The results show that both models performed comparatively similar. In theory, we believed that Segment-CNN should have performed better because the Sentence CNN cannot differentiate the inputs when multiple drug-entity pairs are located in a sentence, but the results contradict the assumption. We believe this is because we treat this as a binary classification problem and build a separate model for each relation type.

Table 3: Results of our deep learning-based approaches over the n2c2-2018 test set

	Segment-CNN			**Sentence CNN**		
	P	**R**	**F**	**P**	**R**	**F**
Strength-Drug	0.91	0.88	**0.90**	0.90	0.91	**0.90**
Duration-Drug	0.39	0.90	0.55	0.41	0.90	**0.57**
Route-Drug	0.77	0.89	**0.83**	0.76	0.91	**0.83**
Form-Drug	0.85	0.95	**0.90**	0.85	0.96	**0.90**
ADE-Drug	0.32	0.85	**0.46**	0.32	0.85	**0.46**
Dosage-Drug	0.83	0.92	**0.87**	0.82	0.93	**0.87**
Reason-Drug	0.27	0.88	**0.42**	0.27	0.88	0.41
Frequency-Drug	0.56	0.88	**0.69**	0.56	0.88	**0.69**
System (Micro)	0.69	0.90	**0.78**	0.68	0.92	**0.78**
System (Macro)	0.68	0.90	**0.77**	0.67	0.91	**0.77**

Segment-CNN performed well with word2vec, whereas Sentence-CNN performed well with GloVe embeddings. We believe this is because the glove embeddings are trained over a more extensive dataset, while the word2vec MIMIC-III embeddings are trained in the same domain.

Contextualized language model-based results. Table 4 shows the Precision (P), Recall (R) and F_1 scores of the four fine-tuned BERT models over the n2c2-2018 test dataset. Results show that the models obtain a similar performance overall and for each entity-drug pairs. Comparatively, BERT_cased model performs better in some categories than the other models.

Table 4: Results of our contextualized langauge model-based approaches over the n2c2-2018 test set

	BERT (uncased)			**BERT (cased)**			**BioBERT**			**Clinical BERT**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Strength-Drug	0.86	0.88	0.87	0.86	0.99	**0.92**	0.86	0.90	0.88	0.87	0.82	0.84
Duration-Drug	0.95	0.93	0.94	0.96	0.93	0.94	0.96	0.93	**0.95**	0.96	0.92	0.94
Route-Drug	0.92	0.99	0.95	0.92	0.97	**0.97**	0.92	0.97	0.94	0.92	0.95	0.93
Form-Drug	0.96	0.97	**0.97**	0.96	0.95	0.96	0.96	0.97	0.96	0.96	0.97	**0.97**
ADE-Drug	0.95	0.99	**0.97**	0.95	0.99	**0.97**	0.95	0.99	**0.97**	0.95	0.99	**0.97**
Dosage-Drug	0.93	0.96	0.94	0.93	0.96	**0.95**	0.93	0.96	0.94	0.93	0.89	0.91
Reason-Drug	0.96	0.98	**0.97**	0.96	0.98	**0.97**	0.96	0.99	**0.97**	0.96	0.99	**0.97**
Frequency-Drug	0.93	0.96	**0.94**	0.93	0.92	0.93	0.93	0.95	**0.94**	0.93	0.95	**0.94**
System (Micro)	0.93	0.96	**0.94**	0.93	0.96	**0.94**	0.93	0.95	**0.94**	0.93	0.96	**0.94**
System (Macro)	0.92	0.95	**0.93**	0.92	0.96	**0.93**	0.92	0.95	**0.93**	0.92	0.95	**0.93**

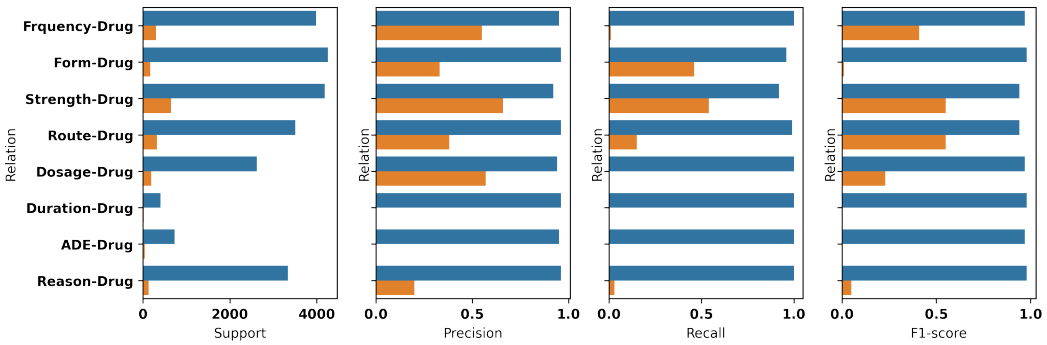


Figure 4: Error analysis of each relation type during the binary classification using BERT (uncased) model. The *Blue* and *Brown* bars represent the positive and negative classes respectively.

7.2 Negation Analysis

In this work, we performed a binary classification for each class: 1) Positive class - there is a relation between the drug and the entity, 2) Negative class - there is no relation between the drug and the entity (no-relation). Figure 4 shows the breakdown of the performance of each class when the binary classification is performed using the BERT_uncased model. We report Support, Precision, Recall, and F_1 score for each class, and *Blue* and *Brown* bars represent the positive and negative classes, respectively. The support shows the number of actual occurrences of the classes. We can see the Precision, Recall, and F_1 score of the positive classes are way higher than the negative classes. We believe this explains the higher performance of the BERT models. The performance of the negative (no-relation) classes is low due to the data imbalance of the classes, as shown in the support. Positive classes are significantly larger than the negative classes, and due to this, the poor performance of the negative classes did not affect the performance of the positive class.

7.3 Comparison across models

Table 5 shows the Precision (P), Recall (R), and F_1 score for the best results of each of our three approaches: 1) rule-based approach using left-only traversal mechanism; 2) deep learning approach using Segment-CNN, and 3) contextualized language model-based approach using BioBERT. Comparing the rule-based approach with our deep learning-based approach shows that the rule-based approach obtained an overall higher Precision, Recall, and F_1 score except for the classes ADE-Drug and Reason-Drug. BERT-based models outperform the other two approaches except for the Strength-Drug, Frequency-Drug, and Form-Drug pairs. The overall Precision and Recall are higher, especially for the entity-drug pairs that performed poorly with the other approaches (ADE-Drug, Reason-Drug, and Duration-Drug). Using pre-trained language representations to fine-tune models is advantageous as they use minimal task-specific parameters and are trained on the downstream tasks by simply fine-tuning all the pre-trained parameters.

Table 5: Comparison across our approaches over the n2c2-2018 test set

	Train	**Test**	**Rule-based**			**Segment-CNN**			**BioBERT**		
	#	**#**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Strength-Drug	6702	4244	0.96	0.95	**0.95**	0.91	0.88	0.90	0.86	0.90	0.88
Duration-Drug	643	426	0.78	0.69	0.73	0.39	0.90	0.55	0.96	0.93	**0.95**
Route-Drug	5538	3546	0.90	0.89	0.89	0.77	0.89	0.83	0.92	0.97	**0.94**
Form-Drug	6654	4373	0.98	0.98	**0.98**	0.85	0.95	0.90	0.96	0.97	0.96
ADE-Drug	1107	733	0.46	0.39	0.43	0.32	0.85	0.46	0.95	0.99	**0.97**
Dosage-Drug	4255	2695	0.89	0.89	0.89	0.83	0.92	0.87	0.93	0.96	**0.94**
Reason-Drug	5169	3410	0.48	0.35	0.41	0.27	0.88	0.42	0.96	0.99	**0.97**
Frequency-Drug	6310	4034	0.98	0.98	0.98	0.56	0.88	0.69	0.93	0.95	**0.94**
System (Micro)			0.88	0.83	0.86	0.69	0.90	0.78	0.93	0.95	**0.94**
System (Macro)			0.85	0.80	0.83	0.68	0.90	0.77	0.92	0.95	**0.93**

7.4 Comparison with previous work

In this section, we compare our results with two previous works utilizing BERT: Wei, et al.[14] and Alimova, et al.[15] To the best of our knowledge, these are the only two works that have applied pre-trained language models of BERT on the n2c2-2018 dataset. Table 6 shows the overall Precision, Recall, and F_1 score of our fine-tuned BERT models with the reported results from the other state-of-the-art BERT-based models on the n2c2-2018 dataset. The F_1 score of all models of Wei et al's and our three models is same, but the Precision of Wei, et al's models is higher whereas the Recall of our models is higher. There is a notable difference between Alimova, et al. models and ours. The F_1 score of all three models of Alimova, et al. are lower than ours, and we believe this is due to the difference in the representation of the inputs for the models.

Table 6: Overall results in comparison with previous work on the n2c2-2018 test data

	Our models				**Wei, et al.[14]**				**Alimova, et al.[15]**		
	Cased	**Uncased**	**Bio**	**Clinical**	**Cased**	**Uncased**	**Bio**	**Clinical**	**Uncased**	**Bio**	**Clinical**
Precision	0.93	0.93	0.93	0.93	**0.98**	**0.98**	**0.98**	**0.98**	**-**	**-**	**-**
Recall	**0.96**	**0.96**	0.95	0.93	0.90	0.90	0.90	0.90	-	-	-
F_1 score	**0.94**	**0.94**	**0.94**	0.93	**0.94**	**0.94**	**0.94**	**0.94**	0.56	0.75	0.75

Table 7 shows a comparison of the F_1 score of our models the results reported by Wei, et al.[14]'s and Alimova, et al.[15] for each class of the dataset. The results show that Alimova, et al.'s models perform lower. However, when comparing the results with Wei, et al., we found the results are complementary; classes that did not perform well with Wei, et al.'s models performed well with our models. Specifically, the classes *Reason-Drug, ADE-Drug, and Duration-Drug* obtained a higher Precision, Recall, and $F-1$ score than Wei, et al.'s. Meanwhile, the Precision, Recall, and $F-1$ score of the class *Strength-Drug* are higher in Wei, et al. We believe this is due to three differences between our systems: 1) Wei, et al. represent an entity-drug pair in an input sentence using the semantic type of an entity to replace the entity itself, whereas we do no such replacement; 2) they perform a multi-class classification, whereas we perform binary classification creating a separate model for each entity; and 3) Wei, et al. Clinical BERT representations were fine-tuned with MIMIC-III over BERT (cased), whereas our representations were fine-tuned over BioBERT.

Table 7: Comparison of F_1 score with previous work over each class of the n2c2-2018 dataset.

	Our models				Wei, et al.[14]				Alimova, et al.[15]		
	Cased	**Uncased**	**Bio**	**Clinical**	**Cased**	**Uncased**	**Bio**	**Clinical**	**Uncased**	**Bio**	**Clinical**
Strength-Drug	0.87	0.87	0.88	0.84	0.98	**0.99**	0.98	**0.99**	0.58	0.68	0.68
Duration-Drug	**0.94**	**0.94**	**0.94**	**0.94**	0.88	0.89	0.88	0.89	0.41	0.66	0.65
Route-Drug	0.95	0.95	0.94	0.93	**0.97**	**0.97**	**0.97**	**0.97**	0.63	0.74	0.74
Form-Drug	0.97	0.97	0.96	0.97	0.97	**0.98**	**0.98**	**0.98**	0.62	0.81	0.81
ADE-Drug	**0.97**	**0.97**	**0.97**	**0.97**	0.80	0.80	0.81	0.81	0.10	0.62	0.62
Dosage-Drug	0.94	0.94	0.94	0.91	**0.97**	**0.97**	**0.97**	**0.97**	0.67	0.82	0.82
Reason-Drug	**0.97**	**0.97**	**0.97**	**0.97**	0.76	0.76	0.76	0.77	0.22	0.73	0.73
Frequency-Drug	0.94	0.94	0.94	0.94	**0.96**	**0.96**	**0.96**	**0.96**	0.53	0.79	0.78

Table 8 shows the comparison of the top five team results reported by the n2c2-2018 challenge participants[1] and our best models. The UTH system developed a joint learning biLSTM+CRF based architecture to identify the entities and relations together; the VA team used a complex traditional machine learning approach that utilized random forests; NaCT proposed a deep learning-based ensemble method; and MDQ used a biLSTM with an attention mechanism. The results show that our BERT based system obtained a higher recall across all of the systems but a lower precision across the first four teams.

Table 8: Our best results in comparison with the top 5 results of the n2c2-2018 competition

	n2c2-2018 Teams					Our systems		
	UTH	**VA**	**NaCT**	**UFL**	**MDQ**	**BERT**	**Segment-CNN**	**Rule**
Precision	**0.96**	0.95	0.94	0.95	0.93	0.93	0.69	0.88
Recall	0.95	0.94	0.94	0.92	0.94	**0.96**	0.90	0.83
F_1 score	**0.96**	0.94	0.94	0.94	0.94	0.94	0.78	0.86

8 Conclusions and Future work

In this work, we have investigated diverse approaches to identify relations between medication information and adverse drug events from clinical notes. We explored a rule-based, deep learning-based, and contextualized language model-based approaches. We evaluated our approaches on the n2c2-2018 dataset and found overall the contextualized language model-based approach using BioBERT outperformed the other approaches. However, our results also showed that the rule-based approach which uses co-location information was sufficient to identify relations between entities whose positions with respect to each other were consistent throughout the text (e.g. Strength-Drug, Form-Drug and Frequency-Drug). In our contextualized language model-based approaches, we represent a drug-entity pair by the entire sentence, but this may not advisable as we can have multiple drug-entity pairs within a sentence. Therefore in the future, we plan to investigate effective ways of unique representations of a drug-entity pair that can capture the positional information of both the drug and entity. In this work, we developed a separate model for each class of the dataset and performed binary classification separately. In the future, we plan to investigate expanding the model to perform multi-class classification for different datasets.

References

1. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association. 2020;27(1):3–12.

2. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality. Jama. 1997;277(4):301–306.

3. Luo Y. Recurrent neural networks for classifying relations in clinical notes. Journal of biomedical informatics. 2017;72:85–95.

4. Schatz SN, Weber RJ. Adverse drug reactions. ACCP (American College of Clinical Pharmacy). CNS. Pharmacy Practice, PSAP. 2015;.

5. Li F, Liu W, Yu H. Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning. JMIR medical informatics. 2018;6(4):e12159.

6. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC medical informatics and decision making. 2015;15(1):37.

7. Miller T, Geva A, Dligach D. Extracting adverse drug event information with minimal engineering. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019. p. 22–27.

8. Dligach D, Bethard S, Becker L, Miller T, Savova GK. Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association. 2014;21(3):448–454.

9. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. Journal of the American Medical Informatics Association. 2020;27(1):65–72.

10. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

11. Xu J, Lee HJ, Ji Z, Wang J, Wei Q, Xu H. UTH_CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017. In: Proceedings of the Text Analysis Conference; 2017. .

12. Sorokin D, Gurevych I. Context-aware representations for knowledge base relation extraction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017. p. 1784–1789.

13. Christopoulou F, Tran TT, Sahu SK, Miwa M, Ananiadou S. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. Journal of the American Medical Informatics Association. 2020;27(1):39–46.

14. Wei Q, Ji Z, Si Y, Du J, Wang J, Tiryaki F, et al. Relation Extraction from Clinical Narratives Using Pre-trained Language Models. In: AMIA Annual Symposium Proceedings. vol. 2019. American Medical Informatics Association; 2019. p. 1236.

15. Alimova I, Tutubalina E. Multiple features for clinical relation extraction: A machine learning approach. Journal of Biomedical Informatics. 2020;103:103382.

16. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2019. p. 1–5.

17. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3:160035.

18. Nguyen TH, Grishman R. Relation extraction: Perspective from convolutional neural networks. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing; 2015. p. 39–48.

19. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532–1543.

20. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018;.

21. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240.

22. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv:190403323. 2019;.

23. Chollet F, et al.. Keras. GitHub; 2015. https://github.com/fchollet/keras.

Developing and Deploying a Scalable Computing Platform to Support MOOC Education in Clinical Data Science

David Mayer[1], Seth Russell, MS[1], Melissa P. Wilson, MS[1], Michael G. Kahn, MD, PhD[1], Laura K. Wiley, PhD[1]
[1]University of Colorado Anschutz Medical Campus, Aurora, CO

Abstract

One of the challenges of teaching applied data science courses is managing individual students' local computing environment. This is especially challenging when teaching massively open online courses (MOOCs) where students come from across the globe and have a variety of access to and types of computing systems. There are additional challenges with using sensitive health information for clinical data science education. Here we describe the development and performance of a computing platform developed to support a series of MOOCs in clinical data science. This platform was designed to restrict and log all access to health datasets while also being scalable, accessible, secure, privacy preserving, and easy to access. Over the 19 months the platform has been live it has supported the computation of more than 2300 students from 101 countries.

Introduction

One of the major challenges faced by data science educators is managing student computing environments. Typically educators must choose between teaching students how to set up an environment on their own computer or hosting a pre-configured server.[1] Setting up a local environment on each student computer, while authentic, is challenging and often time-consuming because of the variety of operating systems and sometimes insufficient user permissions (e.g., for students using employer-provided computers). Server-based solutions shift managing the complexity of the computing environment to instructors, which reduces the authenticity of learning to manage the entire data science pipeline. A number of commercial solutions, like RStudio Cloud, have emerged to support educators providing a hosted solution without having to manage servers directly.[2] While server-based solutions have some costs associated, for students with internet access they can increase equity of education as all students have equal computational power regardless of their own computing hardware.[1,3]

These technology challenges are magnified for those teaching data science focused Massively Open Online Courses (MOOCs). MOOCs are typically offered to thousands of learners across the globe completely asynchronously, increasing the number of unique computing environments and reducing instructor contact for individual-level support. Previous data science MOOCs have devoted an entire 4 week (~13hour) course to setting up students' computational environments.[4,5] Others use technology embedded in the platform's learning management system (e.g., shared JuptyerHub).[3,6] Importantly, these data science MOOCs have not had a particular domain focus and thus can use openly available or non-sensitive data in their courses.

While MOOCs are an attractive solution to the increasing demand for a clinical data science workforce,[7] it is not clear how to support student computing environments when working with sensitive healthcare data. We developed a series of MOOCs ("Specialization") on clinical data science,[8] that uses real clinical data (MIMIC-III demo database).[9] At the time, all individuals seeking access were required to complete data use agreements. Setting up a student's local environment would not allow us to restrict or track data download and potential sharing with external entities. Built in data science solutions on the course hosting site had the same limitation and additionally would not allow instructors to restrict data access to only those students who had completed a data use agreement. In response to these challenges we sought to create a hosted computing platform that would both manage student access to restricted materials and accommodate the unique challenges posed by MOOCs.

Methods

Designing the Computing Platform

The primary goal of developing the computing platform was to create a system that would allow instructors to restrict and monitor access to sensitive clinical data to only those students who had signed required data use agreements. The secondary goals of the computing platform were to support challenges inherent in MOOC education and hosted computing, namely: 1) availability and scalability, 2) secure and privacy preserving, and 3) easy independent access. Given the world-wide access and scope of MOOCs, the platform had to be available across

the globe 24 hours a day, 7 days a week, and be able to support potentially hundreds of thousands of learners.[5] As with all server-based solutions, especially with those hosting clinical data, the platform needed to be secure and have full logging of user activity. Additionally, while it is not legally clear the extent to which MOOCs are subject to the Family Educational Rights and Privacy Act (FERPA)[10] in order to comply with the course hosting company's privacy policies, the computing platform needed to preserve student privacy. Finally, given the limited contact with instructors and large student to instructor ratio, the platform onboarding process and use had to be as simple as possible.

Developing and Deploying the Computing Platform

During development of the computing platform we had a number of resources that shaped the final product. First, we had previously run a version of the specialization as a regular university course that used our university approved Google Cloud platform (GCP)[11] infrastructure to host student computation. This experience led us to develop a partnership with Google Cloud Healthcare to provide financial support for the creation of the specialization and the hosting of student computation. Second, the Google Cloud Healthcare team routinely hosts healthcare datathons using GCP and MIMIC-III datasets. The hosting guide, system configurations, and other pipelines to support these datathons are all publicly available on GitHub.[12] Importantly, outside of the choice to use GCP, these resources informed, but did not dictate, the final computing platform created.

GCP is a suite of cloud computing tools that includes support for computing, data storage and databases, networking, identity and security management tools, and advanced analytics for big data applications with all resources organized, managed, and billed to individual projects.[11] Organizations may have multiple projects and resources can be easily moved between projects as needed. We created a Google Organization (LearnClinicalDataScience) for the computing platform that consisted of two sets of projects - one for developing and prototyping platform improvements (Development) and the other used in production for hosting student work (Production). Within each set of projects, one is devoted to managing student enrollment and platform monitoring (Management) and the other for hosting student work (Computing). All projects use Google Compute Engine[13] (i.e., virtual servers), Cloud Operations Logging and Monitoring,[14] and BigQuery.[15] The Management project handles student enrollment and access with custom R-scripts, Google Forms,[16] Google Groups,[17] Cloud Identity and Access Management (IAM),[18] and SendGrid[19] email service for student communication. The Computing project hosts student computing using RStudio Server Pro (v1.2.5019-6)[20] and R (v3.6.3).[21]

We developed a complete version of the computing platform in the Development projects and then performed a series of beta tests. Initial beta tests consisted of project team members (DM, LW) creating user accounts and performing basic computational tasks to ensure that the account management process performed as designed. Basic security checks and penetration tests were conducted by SR to identify any obvious security risks in the platform. We then conducted a group beta test of 27 local users to test simultaneous user registration and to develop an understanding of computing resources required for sample computational workloads similar to those used in the course. Changes to the computing platform and course materials were made following the group beta test and the platform re-tested by project team members (DM, LW). After completion of all beta testing informed improvements, copies of the Development machines were created in the Production projects. The final computing platform was put into production in January 2019. After moving to production, the Development projects were suspended (e.g., shut down, but available for access as needed) and used intermittently to identify the impact of software updates and prototype new platform modifications.

Evaluating Performance of the Computing Platform

We analyzed data from the first 19 months of computing platform usage (January 15, 2019-August 15, 2020) to understand overall platform performance and associated costs. We identified the number of distinct students who registered for the computing platform (and accepted the data use agreement), and created a frequency map (by country) of all registered students using the city and/or country reported in their signed data use agreement. Computing logs (e.g., unique R sessions, R code inputs, and queries run in BigQuery) were analyzed to determine overall student usage (e.g., number with at least one entry of each type). We also investigated the average volume of platform usage across each day of the week, both with respect to a constant timezone (e.g., overall computing load at any single time point), and student timezone (e.g., what time of day students access course work). All computing logs are captured in UTC. We inferred student local time by mapping student's reported country/city combination to the regionally observed local time zone. When city level data was not provided, a best attempt to assign a timezone

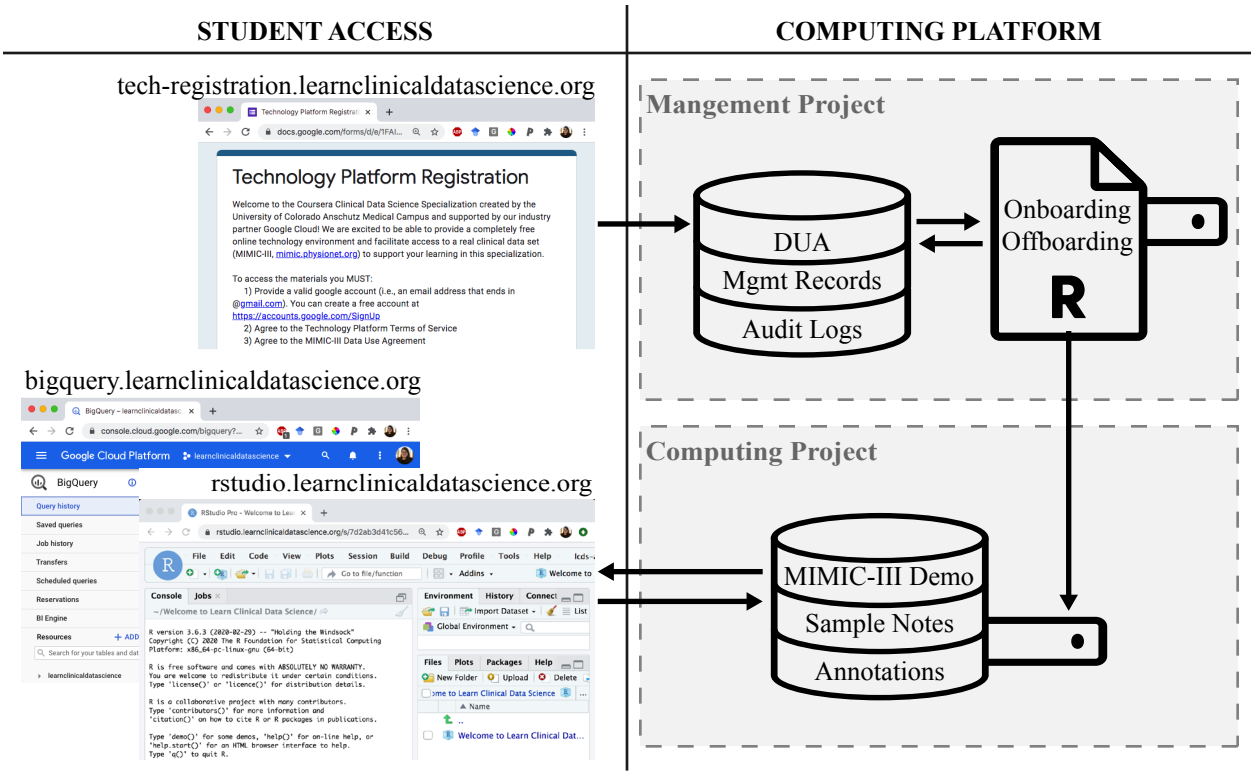


Figure 1. Overview of Computing Platform Design and Student Access Process

was made by assuming the most prevalent timezone for the area. System performance and reliability were assessed by analyzing all system performance and reliability logs to identify types of errors encountered and total system downtime. Finally, all available discussion forums on the main course hosting site (Coursera), were manually reviewed (DM) and categorized by type of question asked to assess the overall frequency of questions generated by the computing platform. All platform metrics were analyzed with R version 4.0.2 and a variety of packages for data processing, graphing and reporting.[22–25]

Results

The Clinical Data Science Specialization and the associated course computing platform launched on January 15, 2019. Course programming assignments consist of html-based tutorials with associated RMarkdown documents requiring ~16MB of storage. As of August 17, 2020 a total of 7,109 students had registered for the first course in the specialization - where students are on boarded to the computing platform. A diagram of the final computing platform implemented is shown in **Figure 1**.

Computing Platform Technical Details

Each Google project (Management, Computing) consists of one or more virtual servers and a set of associated BigQuery datasets. The Management server has 1vCPU, 3.75GB RAM, and a 50GB SSD boot disk running Ubuntu 18.04. This machine is used to process student enrollments and manage all platform logging activities. The associated BigQuery datasets consist of student management logs (1.2GB), data use agreements (888MB), and R-Session, R-Console, and BigQuery Web UI logs (15.6TB). Earlier versions of the computing platform also used a License Management Server (0.5vCPU, 1.70GB RAM) to host the license key for RStudio Server Pro within the Management Project. All servers in the Management project have firewalls limiting access to the University of Colorado campus. The Computing server has 2vCPU, 7.75GB RAM, a 50GB SSD boot disk and an additional 100GB SSD disk for student file storage. Students complete their coursework on this machine using R and a professional license of RStudio Server. The RStudio Server is configured to have all course related packages pre-installed, restrict terminal access, and set limits on file upload sizes (to attempt to limit loading of non-course data). A firewall is configured such that only https encrypted web traffic on the RStudio interface is accepted. Student's browsers must support a minimum TLS version of 1.0 and HTTP Strict Transport Security is enabled. Each student

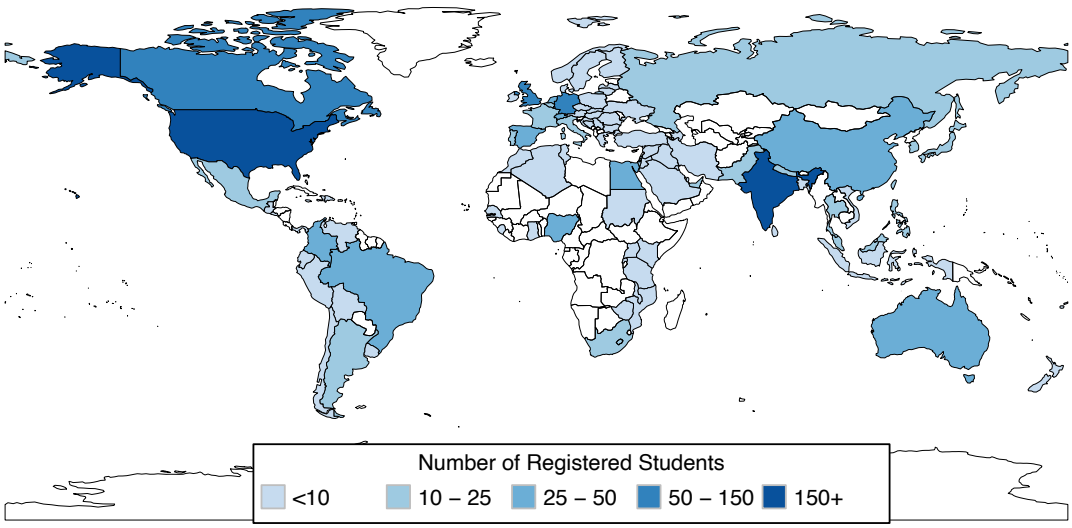


Figure 2. Volume of Computing Platform Registrants by Country

has a disk quota (150MB soft, 200MB hard limit). Process time and memory allocation are capped on a per-user basis with calculations limited to 10-min & 2 GB memory. R-sessions are configured with 30 minute idle timeout value. Students have access to 7.9GB of data through BigQuery, including the MIMIC-III demo dataset (in both the original data model and as an OMOP transformation), gold-standard labels for computational phenotyping and NLP courses, and a series of sample clinical notes generated from medical transcription training samples. Every server in the platform has full logging of all system commands and resource usage (i.e., free memory, free CPU cycles, and free disk space) and OS patches are applied automatically at regular intervals.

Accessing the Computing Platform

Students register for access through a vanity url (tech-registration.learnclinicaldatascience.org) that connects them to a Google Form where they sign the MIMIC-III data use agreement and platform terms of service (e.g., only use the platform for course work, do not attempt to identify other students). These form responses are stored in a Google Sheet that is also accessible from BigQuery. Within the Management Project the Management server runs a custom R-script at 5-minute intervals checking for new registrations. This script determines if the stuwdent is new or returning. New students are assigned a unique student id number, provisioned a linux user account on the Computing server (which is linked to their Google account), added to a private Google Group that has been assigned the appropriate IAM roles for accessing course BigQuery datasets, and a platform expiration date calculated (~6mo). For returning students (i.e., those with an expired account) their original student id number is identified, their prior linux account reassociated with their Google account, and they are readded to the Google Group. When these onboarding processes are complete, an email is sent to the student confirming their registration and providing links to the course resources. Students can access course data through bigquery.learnclinicaldatascience.org (redirects to the Computing project on BigQuery) and R/RStudio through rstudio.learnclinicaldatascience.org. Students log in to RStudio using their Google account, no passwords or personally identifying information is available on the Computing server. Six months after user registration, the student's linux account-Google link is removed. While students are not notified of this change, if they try to login they will receive a custom error that asks them to re-register (at this time all their data is preserved).

Computing Platform Usage

From the launch of the computing platform on January 15, 2019 through August 15, 2020, 2,308 students have requested access to the platform and datasets. Of these registered students, 1,566 (67.9%) have logged in to the computing platform, 1,215 (52.6%) queried course data, and 904 (39.2%) have run R code on the platform. Registered students' location reported on their data use agreement shows that they live in more than 101 different countries. The overall volume of student registrants by country is provided in **Figure 2**.

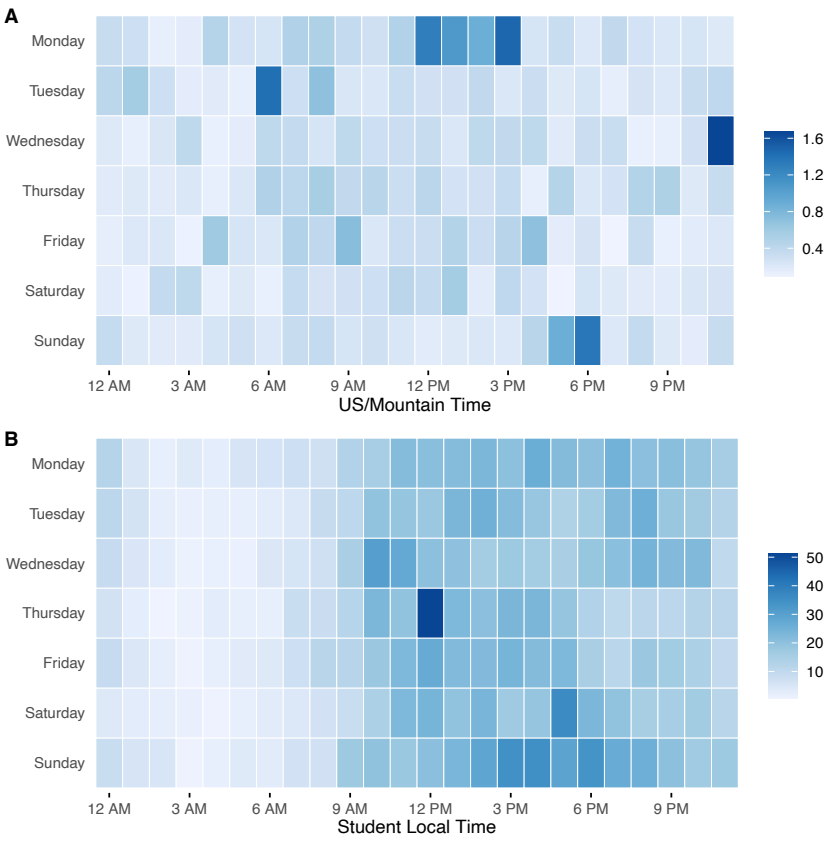


Figure 3. Computing Platform Average Weekly Access
A) This panel shows the weekly average number of R sessions started during each hour/day combination with respectic to the US Mountain Time Zone. B) This panel shows the weekly average number of lines of code entered in the computing platform within each hour/day combination with respect to the student's local time zone.

Based on data use agreement location, students are accessing the platform from 83 different time zones. With respect to the United States Mountain time zone (MT), the highest total number of connections have occurred at 11pm on Wednesday with 138 connections. The least common hour/day combinations (5pm Saturday, 7pm Friday) have seen only 8 connections total across the 82 weeks the platform has been live. **Figure 3A** shows a summary of the weekly average number of logins by time/day of the week with respect to MT. **Figure 3B** shows a summary of the weekly average number of code commands executed by time/day of the week with respect to the students' time zone. The most frequent time/day combination was Thursday at noon, which was driven by work performed by 24 unique students from India (GMT+5:30) across 18 different Thursdays in the evaluation period. Manual review confirmed that this activity was related to course content. Overall, among students who ran R code on the platform, students entered a median of 65.5 [IQR: 11, 187.8] code commands with a single student entering 5,356 code commands. The median amount of storage used by students was < 1MB with a range of 0 - 227.5 MB.

Computing Platform Performance

Over the 82 weeks the computing platform has been running there were 25 unexpected alerts or other errors, of which 6 resulted in downtime where students were unable to access the system. The majority of alerts originated on the Computing server (n=18), 13 alerts were related to CPU and memory utilization (i.e., available CPU < 50%, free RAM < 25%). Five uptime alerts on the Computing server were related to: DNS outages (n=2, downtime only for affected regions), RStudio Server errors (n=2), and a License Management Server outage (n=1). The License Management Server had 5 unexpected outages totaling just under 3 hours of downtime, however the majority (n=4) of these outages lasted less than 30min - the interval the Computing server uses to confirm an active licence. One incident occurred overnight in MT and accounted for approximately an hour and a half of downtime on the Computing server and resulted in at least two students unable to access the machine. The Management server had no unscheduled downtime. **Table 1** provides a summary of the total number of unexpected alerts and computing platform outage durations. Overall the computing platform has had a system uptime percentage of 99.8%.

There have been 13 instances of user onboarding errors. First, although we require all students to register with an "@gmail.com" address, some accounts are treated in IAM as an "@googlemail.com" address. One student registered with such an account and was unable to access the platform until we adjusted our onboarding process to account for this scenario. Three students registered with the same Google account multiple times within 5 minutes (the onboarding script interval) which resulted in them unexpectedly being assigned to the same user account. When multiple registrations occur in the same batch, a student id was only assigned to the first registration with subsequent

Table 1. Summary of Computing Platform Alerts and Outages

	Count Unexpected Alerts	Alert/Outage Duration	Computing Platform Outage Duration
License Management Server	5	2hr 54min	1hr 24min
Computing Server			
CPU Utilization Alert	11	5hr 33min	-
Memory Utilization Alert	2	183hr 16min	-
Uptime Alert	5	23hr 10min	22hr 7min
Total Computing Platform Unavailability:			23hr 31min

registrations labelled as NA, however an account and account/mapping were performed for both accounts. As the RStudio-Google authentication map uses the most recently added mapping, all students accessed a single (shared) "NA" user account. Finally, 9 students were not initially granted data access when registering. This was due to exceeding a Google limit on IAM users that can share a single role (n=1,500).

Across the four launched courses of the Clinical Data Science Specialization there were 196 forum posts available for analysis, with 259 questions/issues raised (students can comment on forum posts to answer a question or echo concerns). Of these issues, 61 (23.5%) related to the computing platform with the majority (n=38, 62.3%) due to the students not registering for access. The remaining 23 included technical issues with students' registration (n=10) or students who had issues locating the resources they needed (n=13).

Discussion

Managing student computing environments is a major challenge in data science education which is magnified when dealing with MOOCs and sensitive clinical data. When developing a series of clinical data science MOOCs we identified the need to develop a computing platform with the primary objective of restricting access to and logging access of sensitive health data. Here we present the results of that work - a computing platform that can only be accessed by students who have completed all requirements for data access (signed data use agreement) where all data access (e.g., SQL queries) and analysis (e.g., R-console commands) are logged and available for review. In addition to these fundamental objectives, we also had three technical requirements related to perceived challenges with hosted computing in MOOCs: 1) availability and scalability, 2) secure and privacy preserving, and 3) easy independent access.

Although we prepared our courses with the potential for an international audience, it wasn't well known whether clinical data science coursework with a particular focus on customs/regulations in the United States would be of global interest. However given the global popularity of other data science MOOCs[5], we designed the platform to be accessible worldwide. Indeed, we had students from around the globe register for access and since deployment of the platform at least one student has started an R session at every single hour of every day of the week. The global use of the platform is confirmed by analyzing platform usage relative to the students' time zone, where most students access the platform during more traditional learning hours (9am-midnight). This platform has also proven to be very stable with only 6 incidents resulting in system-wide outages for only 23.5 hours across the 13,896 hours the platform has been available.

One of the benefits of using a cloud infrastructure is that the platform has proven to be scalable on demand. Initial deployments used larger servers and directory storage (500GB), as actual platform usage started we were able to initially reduce our storage to 100GB and eventually increase it to 150GB after a year of additional student access. Similarly, we can adjust the server specifications (CPUs and RAM) as needed with limited downtime (simply requires restarting the server). This flexibility has had the additional benefit of allowing for cost control measures as we can shrink and grow the machine as needed. An additional benefit has been the query caching function within BigQuery where identical repeated queries are not charged. These types of repeated queries are common for students working through set exercises. Even without aggressive system size optimization, the computing platform development cost $1,431 and continuous access for more than 2300 students across 19 months on the production platform only cost $2,941.

We took multiple steps to ensure the system security and student privacy. First, our platform design uses standard web-based security steps including enabling SSL and limiting server access for students as much as possible beyond those resources required for completing course content. In addition to logging all access and activity, we routinely

monitor those logs for unapproved access attempts including running system commands. Even if a student somehow bypassed our permissions restricting access for viewing/modifying system files, all student files are labeled by student number only. Additionally, although we use Google Groups for managing data access roles, by using an Organization Google account the membership of this group (and group enrollment status) is kept entirely hidden from students. To our knowledge, outside of the single instance of three students getting mapped to the same user account, no private student information has been inappropriately accessed.

Given the exceptionally high student to instructor ratios in MOOCs, it was critical that our platform be accessible without extensive instruction or hands-on attention. To this end students interact with only three sites, all with custom vanity URLs - one for registration and two for platform access. By all available measures our approach has worked well. Technical issues accounted for a minority of forum posts across the courses. The majority of the issues raised were related to students having not attempted to register - suggesting that the primary improvement is needed within our learning management system to highlight the need to access the external site. Although a minority of students did have actual technical issues accessing the site due to registration errors, these issues were usually fixed within the same or subsequent day after reporting on the forum. Although available data supports the easy accessibility of the platform, the number of students registered is dramatically lower than course registrants. The only training students received on the platform was a course reading describing the registration process, and it is possible that this drop is due to unreported student issues with the platform. Alternatively, MOOCs experience a well known phenomena where the number of students registering far exceeds those who perform any course work with even fewer completing the course.[26]

While we have pleasantly been surprised by the overall efficacy and efficiency of the system, there are numerous limitations to our approach. First, and most importantly, this approach was not without cost. Although as highlighted above the expenses are manageable by right-sizing the system size, there is a non-trivial cost associated with hosting student computation. Our numbers are also much smaller than would be expected for data science courses that use much larger datasets/computationally intensive algorithms because we only required small servers and limited storage space. The MIMIC-III demo data contains the records of only 100 patients and though the sample note corpus is larger, most of the computation performed on the platform is happening within the database. Second, we were unable to specifically assess students' feedback on the platform outside of their voluntary posts in the course forums. This assessment would be helpful to inform future iterations or applications of the platform. Third, although there is renewed interest in online computing solutions due to the COVID-19 pandemic requiring remote education, this solution is likely over-engineered for the majority of university courses. Finally, due to platform security concerns we have not made the processing scripts available publically.

Finally, we have found a number of unexpected benefits from the course platform. First, logistically having full logging of all issues commands has been invaluable when answering student questions, both technical and around course content. When students report that they can't access course resources it's easy to pinpoint whether they simply haven't registered or if they are having some other issue. For course problems we are able to see what commands they are running and provide targeted recommendations for "common issues" students encounter. Second, as educators and content creators, these logs also allow for valuable insight into how students interact with course materials. We hope to perform more robust studies of these data in the future to inform both clinical data science education best practices and to improve our own course materials.

Conclusion

We have created a computing platform to support clinical data science MOOC education that has been scalable, globally available, secure, privacy preserving, and generally supported independent access by a large number of students.

Acknowledgements & Funding

We thank our Google partners, especially Marianne Slight, Kate Strasburger, and Stuart O'Brian for their support and their team's technical advice. We are especially grateful to the MIT Laboratory for Computational Physiology whose work and support allowed us to provide students with real clinical data. Our beta testers and students who have provided feedback on the platform have all significantly improved our final production platforwm. Finally we would like to thank the extraordinary team who have contributed to the creation of the clinical data science MOOC: Chan Voong, Christine Mousavi, Jay Billups, Janet Corral, Deborah Keyek-Franssen, Jill Taylor, Jill Lester, Jaimie Henthorn, Aileen Sanders, Alesia Blanchard, Ashley Boshoven, and Benita Bazemore-Cook. Computing platform

use and development costs were supported by our partnership with Google Cloud Healthcare, and complimentary RStudio Server Pro licenses were provided by the RStudio Education team.

References

1. Kross S, Guo PJ. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges. 2019 May 2 [cited 2020 Aug 18];1–14. Available from: https://dl.acm.org/doi/pdf/10.1145/3290605.3300493
2. RStudio Cloud - Do, share, teach, and learn data science [Internet]. [cited 2020 Aug 19]. Available from: https://rstudio.cloud/
3. Suen A, Norén L, Liang A, Tu A. Equity, Scalability, and Sustainability of Data Science Infrastructure. In: Proceedings of the 17th Python in Science Conference doi [Internet]. 2018. Available from: http://conference.scipy.org/proceedings/scipy2018/pdfs/anthony_suen_laura_noren_alan_liang_andrea_tu.pdf
4. Leek J. The Data Scientist's Toolbox [Internet]. Coursera. Available from: https://www.coursera.org/learn/data-scientists-tools?specialization=jhu-data-science
5. Kross S, Peng RD, Caffo BS, Gooding I, Leek JT. The Democratization of Data Science Education. Am Stat [Internet]. 2020 Jan 2;74(1):1–7. Available from: https://doi.org/10.1080/00031305.2019.1668849
6. Brooks C. Introduction to Data Science in Python [Internet]. Coursera. Available from: https://www.coursera.org/learn/python-data-analysis?specialization=data-science-python
7. Bresnick J. Lack of Talent, Direction Afflict Healthcare Data Analytics Plans [Internet]. Health IT Analytics. [cited 2020 Aug 19]. Available from: https://healthitanalytics.com/news/lack-of-talent-direction-afflict-healthcare-data-analytics-plans
8. Learn Clinical Data Science [Internet]. [cited 2020 Aug 19]. Available from: https://www.learnclinicaldatascience.org/
9. Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database Demo [Internet]. PhysioNet; 2019. Available from: http://dx.doi.org/10.13026/C2HM2Q
10. Young EM. Educational Privacy in the Online Classroom: FERPA, MOOCS, and the Big Data Conundrum. Harvard Journal of Law and Technology [Internet]. 2015 [cited 2020 Aug 21]; Available from: https://www.semanticscholar.org/paper/5bcd4885d7cd291fcae72c05c802338869d55859
11. Google Cloud Computing, Hosting Services & APIs [Internet]. [cited 2020 Aug 21]. Available from: https://cloud.google.com/gcp/
12. Datathon Support by Google Cloud Healthcare [Internet]. GitHub. 2019 [cited 2020 Aug 21]. Available from: https://github.com/GoogleCloudPlatform/healthcare
13. Compute Engine: Virtual Machines (VMs) [Internet]. [cited 2020 Aug 21]. Available from: https://cloud.google.com/compute
14. Operations: Cloud Monitoring & Logging [Internet]. [cited 2020 Aug 21]. Available from: https://cloud.google.com/products/operations
15. BigQuery: Cloud Data Warehouse [Internet]. [cited 2020 Aug 21]. Available from: https://cloud.google.com/bigquery
16. Google Forms: Free Online Surveys for Personal Use [Internet]. [cited 2020 Aug 21]. Available from: https://www.google.com/forms/about/
17. Google Groups [Internet]. [cited 2020 Aug 21]. Available from: https://groups.google.com/
18. Cloud Identity and Access Management [Internet]. [cited 2020 Aug 21]. Available from: https://cloud.google.com/iam
19. Email Delivery Service [Internet]. [cited 2020 Aug 21]. Available from: https://sendgrid.com/
20. RStudio Server Pro [Internet]. [cited 2020 Aug 21]. Available from: https://rstudio.com/products/rstudio-server-pro/
21. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available from: http://www.R-project.org/
22. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS [Internet]. 2019 Nov 21;4(43):1686. Available from: https://joss.theoj.org/papers/10.21105/joss.01686
23. Neuwirth E. RColorBrewer: ColorBrewer Palettes [Internet]. 2014. Available from: https://CRAN.R-project.org/package=RColorBrewer
24. Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots [Internet]. 2020. Available from: https://CRAN.R-project.org/package=ggpubr
25. South A. rworldmap: A New R package for Mapping Global Data [Internet]. Vol. 3, The R Journal. 2011. p. 35–43. Available from: http://journal.r-project.org/archive/2011-1/RJournal_2011-1_South.pdf
26. Reich J, Ruipérez-Valiente JA. The MOOC pivot. Science [Internet]. 2019 Jan 11;363(6423):130–1. Available from: http://dx.doi.org/10.1126/science.aav7958

CONSIDER Statement: Consolidated Recommendations for Sharing Individual Participant Data from Human Clinical Studies

Craig S. Mayer, MS[1], Nick Williams, Ph.D[1], Vojtech Huser MD, Ph.D[1]
[1]Lister Hill National Center for Biomedical Communication, National Library of Medicine, NIH Bethesda, MD

Abstract

Many research sponsors require sharing of data from human clinical trials. We created the CONSIDER statement, a set of recommendations to improve data sharing practices and increase the availability and re-usability of individual participant data from clinical trials. We developed the recommendations by reviewing shared individual participant data and study artifacts from a set of completed studies, as well as study data deposited on ClinicalTrials.gov and on several data sharing platforms. The CONSIDER statement is comprised of seven sections including: format, data sharing, study design, case report forms, data dictionary, data de-identification and choice of data sharing platform. We developed several different forms of CONSIDER which includes a brief form (the checklist), a full form (detailed descriptions and examples), and a scoring methodology. The checklist can be used to evaluate adherence to various progressive data sharing recommendations. We are currently in Phase 2 of collecting feedback on the CONSIDER statement.

Introduction

In the past, influential policy documents that guide publishing of medical journal articles achieved advances on how results of clinical research are reported. For example, the CONSORT statement and checklist targeted improved reporting of participant flow diagrams and interventional trial designs, analysis and interpretations.[1] Similarly, the STROBE statement aimed to improve the reporting of observational studies.[2] The creation and use of these policy documents have had a profound positive effect in their respective areas.[3,4] We assume that same mechanism may help improve sharing of de-identified individual participant data (IPD) from completed human clinical studies (interventional trials and observational studies).

By policy, many research sponsors currently require sharing of data from human clinical studies.[5] In some of these cases there is not a consensus on a method or format for sharing IPD data. This results in Principle Investigators (PIs) and study sponsors having to decide how best to ensure they comply with this requirement.

For secondary data users there are many challenges when dealing with acquiring and using shared data. These challenges include data integration and interoperability issues, ethical and policy considerations, financial and time constraints and data quality and annotation problems.[6,7] The burden of many of these challenges are shared with PIs who are looking to make their data available, but PIs can mitigate these challenges by taking pre-emptive action when anticipating sharing their study data.[8] Despite the challenges, there are many benefits to sharing study data and data reuse, such as significant reductions in the time and funding that would otherwise be needed to produce new data, as well as the ability to generate and test new hypotheses and compare and contrast multiple sources of clinical trial data.[9,10]

There are many considerations PIs need to account for in order to minimize these challenges of sharing their data and maximize the capabilities of its re-use. These considerations include: what data and data artifacts should be shared, what formats study resources should be shared in, and how to best make accessible and share these resources, either via a data sharing platform or by request from interested secondary researchers.

There are different formats to be considered in study design when it comes to data collection and reporting. While in the past, many studies have used custom formats developed specifically for the study for data collection and reporting, there are various initiatives that have been developed over recent years that maximize the interoperability and reuse of collected participant data by being implemented during study design. These initiatives include the use of data standards or common data elements (CDEs) that aid in the harmonization of different clinical studies.[11] These efforts are to encourage and improve the re-use capabilities of clinical study data. The most popular of these initiatives are standards set by the Clinical Data Interchange Standards Consortium (CDISC), which despite existing for over two

decades, has not become widely adopted by academic medical centers and are mainly used by pharmaceutical research sponsors, thanks to the Food and Drug Administration (FDA) mandate.

As far as making data available for re-use there are many methods that can be chosen. Over the past few years several data sharing platforms with various capabilities have been developed to share and acquire IPD from completed human clinical trials. Some platforms such as NIDA Data Share, are domain specific, while others, such as Vivli or Clinical Study Data Request (CSDR), are general and include studies from a variety of clinical domains.

We present a set of recommendations to improve the practice of data sharing and promote data re-use, known as the Consolidated Recommendations for Sharing Individual Participant Data from Human Clinical Studies (CONSIDER statement). The acronym is loosely based on letters contained in the title: CONSolidated REcommendations for sharing Individual participant Data. Letters E and R are re-ordered to create a more memorable acronym. This set of recommendations provides a checklist and a set of recommendations to guide PIs, study team members, study sponsors and data sharing platform representatives for optimal clinical research data sharing. We use the term study to refer to both interventional trials and observational studies.

Methods

Set of reviewed studies

To develop the CONSIDER statement, we looked at the current state of sharing data from human clinical studies. Our analysis of shared study materials included reviewing the structure and features of shared IPD data and analyzing the presence and format of different study artifacts (e.g. data dictionary, case report forms [CRF], etc.). We also analyzed the accessibility and availability of study resources by reviewing and following the process for requesting study data on data sharing platforms and analyzing study records on a clinical trial registry where we identified which fields are commonly included and excluded by record administrators.

Our analysis included a set of HIV clinical studies which were obtained as part of a larger project focusing on CDEs in HIV studies.[12] While the trials underpinning the recommendations are primarily made up of HIV-related studies, the recommendations are not limited to HIV, and were generated as general recommendations that can be applied to studies from any clinical field. Further adding to the generalization of our recommendations, the data platforms and sources, as well as certain acquired data artifacts used in developing the recommendations are not HIV specific and include non-HIV studies and data sources that exemplify good data sharing practices or present popular challenges associated with data sharing.

The CONSIDER statement was developed from the review of IPD from 30 studies, study data artifacts from 48 studies, and an analysis of 10 data sharing platforms and 6 clinical trial networks. We also did a comprehensive review of clinical study registration data, the presence of data artifacts and the plan to share IPD for HIV trials on ClinicalTrials.gov.[13]

Design assumptions

The development and use of CONSIDER requires a few comments and assumptions. First, we do not recommend any one data sharing platform or data standard or structure. These recommendations are intended to recommend specific features and capabilities rather than a specific entity.

Second, the CONSIDER statement was also developed with the knowledge of the constraints and limited resources PIs face when it comes to staff, time, funding, and privacy. With that in mind, we know this limits the capabilities for a PI to prepare data for sharing, use a specific sharing platform or use certain data structures. The CONSIDER statement is intended to be used as a checklist to use the optimal practices, where plausible, to maximize the visibility, shareability and re-usability of clinical study data.

Results

Sections

Based on our analysis we structured CONSIDER into seven sections that reflect key areas relevant to data sharing. Each section contains between 1 and 13 checklist items. The sections are:

1. *Data Format:* Recommendations on the data structure and the inclusion of certain aspects and elements of the IPD. Using certain methods when formatting IPD can greatly improve the functionality of the data to data re-users and ensure an effective analysis of the shared data.

2. *Data Sharing:* Includes how to make the study available and visible to potential secondary researchers. This section also includes how to share information about the study that can give data re-users a complete understanding of the study and how best to use the available data.

3. *Study Design:* Includes data collection and data sharing recommendation that are important to consider by PIs during study design. The recommendations aim to improve the data usability and data comparability (to other similar studies).

4. *Case Report Forms:* Recommendations about the inclusion of CRFs. CRFs are valuable study artifacts for data recipients to understand the data collection process and the underlying documents that generated the data being analyzed.

5. *Data Dictionary:* Provides recommendations about the availability, format, and features to include when sharing the data dictionary of a study. This section includes making data dictionaries as widely and publicly available as possible and in a format that is easy to use (machine readable). This section also includes recommendations on key information about each data element or form to include in the dictionary. A more comprehensive analysis of data dictionaries leading to the recommendations included in this section can be seen in our previous work.[14]

6. *Data De-identification:* Focuses on describing how data was redacted during de-identification and what should be provided to data recipients when referring to de-identification techniques so they have an understanding of how the data has been changed from the raw collected data. Data de-identification also include the clear communication about the rights and restrictions of data recipients when referring to the process and risks of potential re-identification.

7. *Choice of a Data Sharing Platform:* Provides features and capabilities to look for when choosing a data sharing platform to deposit data (at study completion). The recommendations specify desired platform features such as the ability to quickly find relevant studies (search capabilities), the presence of available study resources and metadata, and the ability to quickly request and acquire available IPD. Different features and capabilities of data sharing platforms improve the effectiveness and efficiency for sharing IPD from clinical studies.

Recommendations

Table 1 shows CONSIDER as a list of recommendations (or checklist format) structured by the previously mentioned sections. For the full version of the CONSIDER statement go to w3id.org/CONSIDER.

Table 1. Individual CONSIDER recommendations by section.

Section	Title of recommendation
Format	Share person table in CDISC or OMOP format
	Group data and data elements into relevant data domains (e.g., medication history, laboratory results history, medical procedure history)
	Follow a convention when using relative time.
	Utilize previously defined Common Data Elements and reference them by their identifiers
	Use formats that can be natively loaded (without highly specialized add-ons) into multiple statistical platforms
Data Sharing	Register your study at ClinicalTrials.gov registry
	Do not limit study metadata to the legally required elements. Also populate optional elements (such as data sharing metadata)
	Fully populate data_sharing_plan text filed on ClinicalTrials.gov (if sharing data)
	If Individual Participant Data is shared on a data sharing platform, update the ClinicalTrials.gov record with the URL link to the data.
	Provide basic summary results using results registry component of Clinicaltrials.gov
	Utilize ClinicalTrials.gov fields for uploading study protocol, empty case report forms, statistical analysis plan and study URL link
	Provide de-identified Individual Participant Data
Study Design	Adopt previously defined applicable Common Data Elements
Case Report Forms	Share all Case Report Forms used in a study
	List all CRFs
Data Dictionary	Provide data dictionary
	Provide data dictionary in machine readable format
	Separate data dictionary from de-identified individual participant data. Since it contains no participant level data, do not require local ethical approval as a condition of releasing the data dictionary (avoid a requestwall for data dictionary).
	Share a data dictionary as soon as possible. Do not wait until the data collection is complete.
	Provide data dictionary in a single, machine-readable file.
	For each data element, provide a data type (such as numeric, date, string, categorical)
	For categorical data elements, provide a list of permissible values and distinguish when numerical code or string code is a code for a permissible value (versus actual number or string)
	Distinguish categorical string data elements from free-text string data elements
	Link utilized Common Data Elements adopted by your study to appropriate terminologies
	Link data elements or permissible values to applicable routine healthcare terminologies (either because you designed them to be linked or post-hoc, they can be semantically linked as equivalent)
	Provide complete data dictionary (all elements in data are listed in a dictionary) and all types of applicable dictionaries (date elements, forms [or groupings], and permissible values)
	Include sufficient description for data elements
	Use identifiers (unique where applicable) for data element, forms and permissible values.
Data de-identification	Provide data de-identification notes
Choice of a Data Sharing platform	Use platforms that allows download of all studies available on the platform
	Choose a platform that supports batch request (ability to request multiple studies with one request)

CONSIDER formats

We developed three views of the CONSIDER statement which have varying levels of detail and serve different purposes. The first format is the *brief view*, which is just a list of the different recommendations and their associated sections as seen in Table 1. The second format is the *full view* that on top of what is included in the brief format, also includes a detailed description of the recommendation, a positive example that features a study or platform that demonstrates full or partial compliance with the recommendation, and optionally a challenging example where the recommendation was not followed and what challenge that leads to during data re-use. For some recommendations, we used positive and challenging examples outside the input set of studies described in Methods. Finally, the third view is the *score sheet view* that assumes familiarity with the individual recommendations and is meant to facilitate individual study scoring. It lists each recommendation (by section) and the scoring instructions.

We acknowledge that some CONSIDER items depend on each other and can be considered partially overlapping. For example, the requirement to list data type for each data element in the data dictionary partially overlaps with properly handling categorical data elements. The description field (in the full view) for each CONSIDER item contains an explanation and rationale for this overlap. We chose to allow some overlap because we saw studies that formally comply with some recommendation, but closer scrutiny reveals additional deficiencies. The seemingly overlapping recommendations are meant to fully clarify and describe the best data sharing practices.

CONSIDER score and scoring approach

We developed a scoring system that scores each recommendation separately and then counts the score by each section. The higher the score the better the practices are implemented by the given study.

Each checklist item is scored on a zero to one scale. For binary items, the possible values are either one for practicing the recommendation or zero if it does not. For items where partial assessment of compliance is possible, a range of values between zero and one can be assigned (with 2-digit precision) depending on how completely the recommendation is followed. For example, for the recommendation 'for each data element provide a data type' if the study provides a data type for 47.15% of the data elements then it will get a score of 0.47 for that checklist item.

Certain recommendations included in CONSIDER rely on publicly available information from ClinicalTrials.gov (CTG) registry. To facilitate the easy application of the CONSIDER checklist, we created an R script that uses the relational database version of CTG, known as the Aggregated Analysis of ClinicalTrials.gov (AACT database; published and maintained by Duke University), to automatically score a subset of checklist items that can be assessed by CTG study registration metadata.[15] This script (located at w3id.org/CONSIDER) takes as input a set of study CTG identifiers (called NCTs) and returns their CONSIDER scores for the subset of checklist items that can be automated.

Example application of the CONSIDER checklist to individual studies

We applied CONSIDER to two trials from our set of analyzed trials to show how the checklist can be applied to individual studies. We scored NCT01751646 'Vitamin D Absorption in HIV Infected Young Adults Being Treated With Tenofovir Containing cART', which has IPD data deposited on the National Institute of Child Health and Human Development's Data and Specimen Hub platform (NICHD DASH), and NCT01233531 'Effects of Cash Transfer for the Prevention of HIV in Young South African Women', which has IPD available upon request from the HIV Prevention Trial Network (HPTN). Both are also registered at CTG. Table 2 shows the scores and percentages for these two studies when applying the CONSIDER checklist.

Table 2. Results of applying CONSIDER scoring for interventional trials NCT01751646 and NCT01233531.

Section	Best Possible Score	NCT01751646	NCT01233531
Format	5	3 (60.0%)	3 (60.0%)
Data Sharing	7	4 (57.1%)	3 (42.9%)
Study Design	1	0 (0.0%)	0 (0.0%)
Case Report Forms	2	2 (100.0%)	1 (50.0%)
Data Dictionary	13	9.45 (72.7%)	6.72 (51.7%)
Data de-identification	1	1 (100.0%)	0 (0.0%)
Choice of a Data Sharing platform	2	0 (0.0%)	2 (100.0%)

Discussion

Seeking feedback

We expect evolution of the CONSIDER checklist and we welcome feedback to any checklist item or section at craig.mayer2@nih.gov. CONSIDER was first developed in May 2019. We performed a Phase 1 feedback stage from Sept. 2019-Dec. 2019, where we elicited feedback from an internally selected group of experts. We are currently (since Jan. 2020) in Phase 2 of collecting feedback from the larger CRI community. As part of the Phase 2 feedback process, we created a mechanism and set of questions specific to the different perspectives involved in the data sharing process. This includes targeted feedback questions intended for PI's involved in study design, study data custodians involved in data housing and distribution, data sharing platform administrators, and data recipients. The developed targeted feedback tools are intended to better assist in understanding the capabilities and challenges for each individual involved in the data sharing process and allow us to better formulate well-rounded recommendations that are both feasible and beneficial for all involved.

Limitations

The resulting CONSIDER statement has several limitations. First, we focused on a US context and considered ClinicalTrials.gov registry. Second, we only used a limited set of studies to arrive at the recommendations. Using a larger set may result in a more comprehensive coverage of best practices. Third, the scoring system assumes equal importance (and weight) of each item. It would be feasible to develop a weighted score if agreement on prioritization can be reached. We also currently do not attempt to combine the score of individual CONSIDER sections into a single score. Fourth, we received only limited feedback from some stakeholder groups (platform administrators and PIs) and plan a focused feedback seeking campaign to address this.

Future work

To further develop CONSIDER we will continue to assess the state of data sharing and accept feedback to add recommendations and sections as they become necessary. We will also look to improve the capabilities of scoring individual studies by automating the process (as we have done with CTG related recommendations), which may include linking directly to other clinical trial registries, data sharing platforms and individual study pages. We also understand the different recommendations may present certain challenges and we intend to assess how demanding and resource intensive the implementation of each recommendation is to better recommend the most practical and implementable practices.

Conclusion

We analyzed data sharing platforms, data artifacts, study registration metadata, and shared IPD for completed clinical trials and created a set of recommendations, called the CONSIDER statement. The CONSIDER statement consists of seven key sections devoted to data format, data sharing, study design, CRFs, data dictionary, data de-identification, and choice of data sharing platform. These recommendations can be used to score existing studies to evaluate adherence to good data sharing practices. The recommendations can also be used to guide PIs and study sponsors to improve data sharing of future studies. We expect evolution of the CONSIDER statement based on input from clinical research informatics experts and the wider research community.

Acknowledgement

This work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

References

1. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. PLoS Med. 2010 Mar 24;7(3):e1000251.

2. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. Ann Intern Med. 2007 Oct 16;147(8):573.

3. Bastuji-Garin S, Sbidian E, Gaudy-Marqueste C, Ferrat E, Roujeau J-C, Richard M-A, et al. Impact of STROBE Statement Publication on Quality of Observational Study Reporting: Interrupted Time Series versus Before-After Analysis. Schooling CM, editor. PLoS ONE. 2013 Aug 26;8(8):e64733.

4. Poorolajal J, Cheraghi Z, Irani AD, Rezaeian S. Quality of Cohort Studies Reporting Post the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement. Epidemiol Health. 2011 Jun 7;33:e2011005.

5. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data sharing statements for clinical trials. BMJ. 2017 Jun 5;357:j2372.

6. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress. Yearb Med Inform. 2017;26(01):38–52.

7. Wilkinson T, Sinha S, Peek N, Geifman N. Clinical trial data reuse – overcoming complexities in trial design and data sharing. Trials. 2019 Dec;20(1):513.

8. Mbuagbaw L, Foster G, Cheng J, Thabane L. Challenges to complete and useful data sharing. Trials. 2017 Dec;18(1):71.

9. Kawahara T, Fukuda M, Oba K, Sakamoto J, Buyse M. Meta-analysis of randomized clinical trials in the era of individual patient data sharing. Int J Clin Oncol. 2018 Jun;23(3):403–9.

10. Rosenblatt M, Jain SH, Cahill M. Sharing of Clinical Trial Data: Benefits, Risks, and Uniform Principles. Ann Intern Med. 2015 Feb 17;162(4):306.

11. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of Common Data Elements. Clin Trials J Soc Clin Trials. 2016 Dec;13(6):671–6.

12. Huser V, Mayer CS, Williams N. Real World Data and Research Common Data Elements: a Case Study in HIV. AMIA Clinical Informatics Conference, May 2020.

13. Huser V. Sharing of de-identified patient level data from human clinical trials: analysis of US-based studies in the ClinicalTrials.gov registry. In 2017.

14. Mayer CS, Williams N, Huser V. Analysis of data dictionary formats of HIV clinical trials. Huy NT, editor. PLOS ONE. 2020 Oct 5;15(10):e0240047.

15. AACT Team. Aggregate Analysis of ClinicalTrials.gov (AACT) database [Internet]. 2020 [cited 2020 Mar 20]. Available from: https://aact.ctti-clinicaltrials.org/learn_more

Empirical Findings on the Role of Structured Data, Unstructured Data, and their Combination for Automatic Clinical Phenotyping

Asher Moldwin, Dina Demner-Fushman, MD, PhD, Travis R. Goodwin, PhD
U.S. National Library of Medicine, Bethesda, MD, USA

Abstract

The objective of this study is to explore the role of structured and unstructured data for clinical phenotyping by determining which types of clinical phenotypes are best identified using unstructured data (e.g., clinical notes), structured data (e.g., laboratory values, vital signs), or their combination across 172 clinical phenotypes. Specifically, we used laboratory and chart measurements as well as clinical notes from the MIMIC-III critical care database and trained an LSTM using features extracted from each type of data to determine which categories of phenotypes were best identified by structured data, unstructured data, or both. We observed that textual features on their own outperformed structured features for 145 (84%) of phenotypes, and that Doc2Vec was the most effective representation of unstructured data for all phenotypes. When evaluating the impact of adding textual features to systems previously relying only on structured features, we found a statistically significant ($p < 0.05$) increase in phenotyping performance for 51 phenotypes (primarily involving the circulatory system, injury, and poisoning), one phenotype for which textual features degraded performance (diabetes without complications), and no statistically significant change in performance with the remaining 120 phenotypes. We provide analysis on which phenotypes are best identified by each type of data and guidance on which data sources to consider for future research on phenotype identification.

Introduction

In recent years the use of Electronic Health Records (EHRs) has become standard practice across hospitals in the United States,[1] with medical professionals regularly recording patient information digitally and using the recorded information to aid in diagnosis and clinical decision making.[2] In addition to the direct clinical benefits of easily being able to look up individual patient records, large collections of EHRs are also useful to researchers who wish to understand medical conditions, disease processes, and hospitalization patterns based on retrospective records. Data documenting patient care is recorded in EHRs through (a) structured measurements such as lab results, vital signs, demographic information, etc., as well as (b) unstructured clinical narratives such as those found in admission reports, nursing notes, or surgical reports. Clinical notes are routinely recorded to provide relevant contextual information about the patient's medical background, condition, and care. While there can be overlap between the information included in the structured and unstructured portions of EHRs, the unstructured nature of clinical notes is uniquely suited for extracting unique or contextual information that may not conform to a preset field or measurement.[3] Both text and structured data have been used to model medical processes for Clinical Decision Support[4,5] and Disease Prediction[6] applications.

Denny (2012)[7] points out that phenotype identification is one task for which clinical notes are particularly useful, noting that this is often because salient observations in text documents such as pathology and radiology reports are often not also included in tabular data. However, while clinical notes are clearly a useful data source, it is still important to know specifically where (i.e. for which phenotypes) notes are most likely to be beneficial and to identify whether there are situations where they are not worthwhile to include at all. Previous studies have used custom-engineered search terms in clinical notes to identify specific clinical phenotypes,[8] while others used a Bag of Words or Bag of Concepts representation to identify clinical phenotypes based on the words that appear in clinical notes.[9,10] However, it is difficult to determine based on these studies whether clinical notes, structured data, or a combination of both should be used to identify a given phenotype that has not previously been automatically identified. Helpfully, Scheurwegs et al. (2015)[11] show that the best performance can be consistently achieved by combining structured data with a Bag of Words representation of clinical notes when predicting ICD-9 billing codes from fourteen different medical specialties based on EHR data. They conclude that adding structured data is consistently beneficial when compared with using only a Bag of Words representation of clinical notes. However, it is not clear if the performance increase observed by Scheurwegs et al. (2015)[11] when adding structured data is due to information that is present in structured data but missing in clinical notes, or if the information is in fact present in clinical notes but is simply not captured using a Bag of Words representation.

For the purposes of this study, we use the term *clinical phenotype* to refer to a clinically significant group of medical

abnormalities that are characteristic of a single disease (sometimes referred to as a disease phenotype).[12] We determine the individual clinical phenotypes of a patient by means of the ICD-9 diagnostic codes assigned to the patient upon discharge from the hospital, as demonstrated in previous phenotyping work.[13,14] Specifically, we rely on the clinically meaningful groupings of ICD-9 codes as defined by the Agency for Healthcare Research and Quality (AHRQ)'s Clinical Classification Software (CCS). We consider the presence of any ICD-9 code in each of these CCS groupings as evidence for the corresponding clinical phenotype.

In this paper, we train a Long Short-Term Memory network (LSTM)[15] to identify 172 phenotypes using structured data features, textual features, and their combination. We explore three methods for representing textual features: Bag of Words, Bag of Concepts, and Doc2Vec. Finally, we apply statistical analysis on phenotyping performance (measured in AUC) to investigate which categories of phenotypes are most likely to benefit from each data source, and which benefit from their combination. Our results indicate that the combination of text and structured features provides a statistically significant ($p < 0.05$) increase in performance compared to using structured features alone for 51 phenotypes, a decrease in performance for a single phenotype, and no statistically significant benefit for the remaining 117 phenotypes. We analyse the categories of phenotypes that are best identified with each data set, provided analyses which we hope can serve as a practical guide for determining which data sources are most likely to be of value for a given phenotype.

Background and Related Work

The task of automatic clinical phenotype identification consists of computationally processing EHRs to determine whether a given patient's clinical phenotype indicates a particular disease. This can be useful when a disease may not have been explicitly documented in the EHR despite the presence of markers that are characteristic of the disease in the patient's record. In addition to being useful for Clinical Decision Support, research into automatic phenotype identification can lead to a better understanding of disease mechanisms and outcomes.[16] Automatic phenotype identification is also a preliminary step in clinical research that requires selecting patient cohorts based on their diseases or medical conditions.[17,18]

The relationship between structured and unstructured data for phenotyping has been explored in the past. For example, Liao et al. (2010)[10] show that rheumatoid arthritis can be automatically identified most effectively when using a combination of both clinical notes and structured data, and Nunes et al. (2016)[19] reached a similar conclusion when identifying hypoglycemic patients based on EHRs. However, neither of these studies compares systems relying on clinical notes with systems relying on structured data for a wide range of different phenotypes, meaning that their findings may not be completely generalizable to the broader task of phenotype identification. Gehrmann et al. (2018)[9] compared different Natural Language Processing approaches for using clinical notes to identify 10 different clinical phenotypes, but did not focus on comparing notes with structured data (though they do express interest in doing this in their *"Future Extensions"* section). Consequently, this study can be viewed as a generalization of these approaches wherein we test several different formulations of text, structured data, and a combination of both to predict a broad range of phenotypes, thus providing insight into whether text is useful for identifying all phenotypes, or just a specific subset of them.

Data

In this work, we extracted general structured data features from patients' laboratory and chart values in the MIMIC-III Critical Care Database.[20] The MIMIC-III Critical Care Database[20] contains de-identified structured data and timestamped clinical notes for *46*, *520* patients and *61*, *532* ICU stays, based on data collected between 2001 and 2012. This includes clinical notes, lab results, demographic information, logs of hospital admission and discharge, prescriptions, and admission-level ICD-9 discharge diagnoses and procedures. For the sake of reproducibility and comparison, we used the 38 structured features extracted from the MIMIC-III Benchmark[21] to identify phenotypes. The MIMIC-III Benchmark is a set of four clinical prediction tasks using a consistent set of structured data features for all tasks: capillary refill rate, diastolic blood pressure, fraction inspired oxygen, Glasgow coma scale eye opening, Glasgow coma scale motor response, Glasgow coma scale total, Glasgow coma scale verbal response, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH. We adapt the same set of structured features in this work. In addition, we used all types of clinical notes present in MIMIC-III (e.g. admission and discharge reports, nursing notes, radiology reports) in our experiments using notes.

As in Harutyunyan et al. (2019)[21], we considered each ICU stay as an independent episode and used discharge

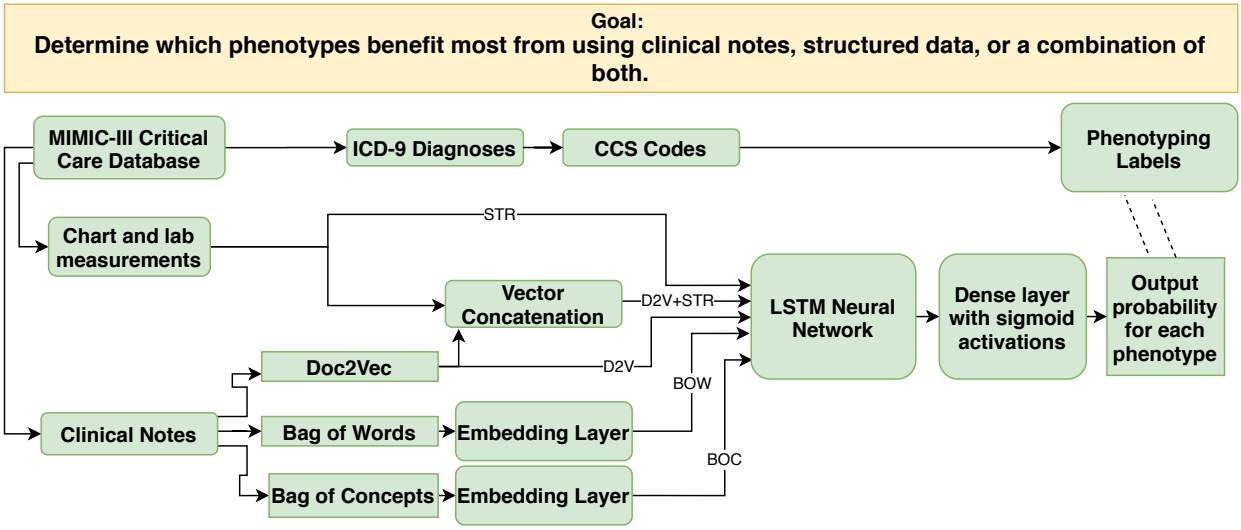


Figure 1: Schematic of the our pipeline including data sources and neural network.

Table 1: Top-level CCS phenotype categories with their size (i.e., the number of phenotypes in the group after filtering phenotypes with < 30 episodes) as well as the prevalence of that group in the testing data.

CCS Category Name	CCS Codes	Size	Prevalence
Infectious and parasitic diseases	1-10	8	26.6 %
Neoplasms	11-47	17	22.2 %
Endocrine; nutritional; and metabolic diseases and immunity disorders	48-58	9	66.9 %
Diseases of the blood and blood-forming organs	59-64	5	35.4 %
Mental illness	650-663, 670	10	35.6 %
Diseases of the nervous system and sense organs	76-95	15	28.1 %
Diseases of the circulatory system	96-121	24	81.6 %
Diseases of the respiratory system	122-134	9	47.3 %
Diseases of the digestive system	135-155	15	41.2 %
Diseases of the genitourinary system	156-175	9	39.7 %
Diseases of the skin and subcutaneous tissue	197-200	4	9.6 %
Diseases of the musculoskeletal system and connective tissue	201-212	11	20.7 %
Congenital anomalies	213-217	2	2.7 %
Injury and poisoning	225-244	16	43.7 %
Symptoms; signs; and ill-defined conditions and factors influencing health status	245-258	9	25.3 %
Residual codes; unclassified; all E codes	259-260	9	41.4 %

diagnoses to determine phenotype labels. Specifically, phenotypes were defined by CCS (Clinical Classifications Software) codes. Developed by the Agency for Healthcare Research and Quality (AHRQ), CCS codes are groups of ICD-9 codes that correspond to specific diseases; these CCS codes are further grouped into a hierarchy based on organ systems and disease categories. Table 1 shows the different top-level CCS categories and the number of phenotypes in each category that we evaluated in this study. Note: unlike Harutyunyan et al. (2019)[21] which considered only 25 phenotypes, we considered all phenotypes associated with at least 30 episodes in MIMIC-III resulting in the 172 phenotypes shown in Table 1. Moreover, we filtered episodes such that all episodes had at least one data point from both the textual and structured data sources. In this study we used the same 14:3:3 splits for training, validation, and testing used by Harutyunyan et al. (2019)[21]. Due to the potential for different disease manifestation, we omitted patients under the age of 18 from this study.

Methods

As shown in Figure 1, our phenotype-identification pipeline consists of preparing training, validation, and testing data based on MIMIC-III and then repeatedly training and evaluating the same deep neural network using different combinations of structured data and three different representations of clinical notes. Our data preparation pipeline consists primarily of (1) extracting clinical episodes based on ICU stays in MIMIC-III, (2) selecting a data source or combination thereof, (3) producing fixed-length continuous vectors capturing the information in the selected data source, and (4) using these fixed-length vectors as the input to train a shared deep neural network for joint phenotype identification.

A. Data Representations

To compare the impact of different data sources, it was necessary to represent the structured data and clinical notes recorded at each timestep with fixed-length, continuous vector encodings. We considered the following methods for representing data in each episode:

1. **Structured Data:** As in Harutyunyan et al. (2019)[21], measurements were aggregated every hour. For measurements consisting of a continuous number (such as height, weight, and temperature), we dedicate one vector dimension to the raw numerical measurement, and for categorical measurements (such as the Glasgow coma scale fields) we use a one-hot encoding with one vector dimension reserved for each possible value of the given measurement. All values were then normalized by subtracting the mean value of each field and dividing by the standard deviation. The resultant structured data vector included a total of 76 elements corresponding to these continuous and categorical variables.
2. **Bag of Words:** We generated vocabularies based on all of the clinical notes in MIMIC-III. For words, the vocabulary size was 1,891,434 words. For each clinical note, we created a Bag of Words vector representation by using a vocabulary-length vector with ones for all words that occur in the note and zeros for all words that do not occur.
3. **Bag of Concepts:** We used MetaMap Lite[22] to identify concepts in clinical notes based on UMLS[23]. Defining the vocabulary to consist only of concepts occurring in clinical notes, we then created a vocabulary-length vector to represent the medical concepts in each note, similarly to the Bag-of-Words representation above. The vocabulary size for concepts was 51,893 concepts.
4. **Doc2Vec:** Because the order in which words appear is not considered by either the Bag of Words nor the Bag of Concepts approaches, we considered a more sophisticated representation – Doc2Vec[24] – for generating document-level representations of each clinical note. We used a vector dimension of 300, an initial learning rate of 0.025, and 100 iterations.*
5. **Structured and Doc2Vec:** We considered a final representation in which the Doc2Vec and Structured Data encodings were concatenated together to form a single multi-datasource representation. Because structured data was available every hour while only 3.2 notes were produced per day (on average), the features corresponding to Doc2Vec were left as zero for hours in which no notes were generated.

Note: an important advantage of the three text encoding schemes is that they can all be applied to entire documents regardless of the document length, unlike other text-based neural network systems such as the Universal Sentence Encoder[25] or BERT[26].

B. Neural Network Architecture

We identify phenotypes by training a Long-Short-Term-Memory (LSTM) network similar to that used in Harutyunyan et al. (2019)[21]. Specifically, we train the LSTM to sequentially process the feature representations from each time-step for a patient such that the output of the LSTM after processing the last time-step can be used to predict the clinical phenotypes of the patient. Because our goal was to test the effect of different data sources (structured vs various forms of text), our approach was to keep the neural network architecture as constant as possible across all of our experiments while varying only the input data. However, it was necessary to modify the network's architecture slightly by adding a fully-connected layer to embed the inputs when using a Bag of Concepts or Bag of Words representation for clinical note text, before input to the first LSTM layer. Inputs are then passed into a bidirectional LSTM network using hyperbolic tangent activation functions. Finally, all outputs are passed through a dense layer using 172 parallel sigmoid activation functions to produce the probability of identifying each of the 172 phenotypes considered in this work.

*Determined empirically

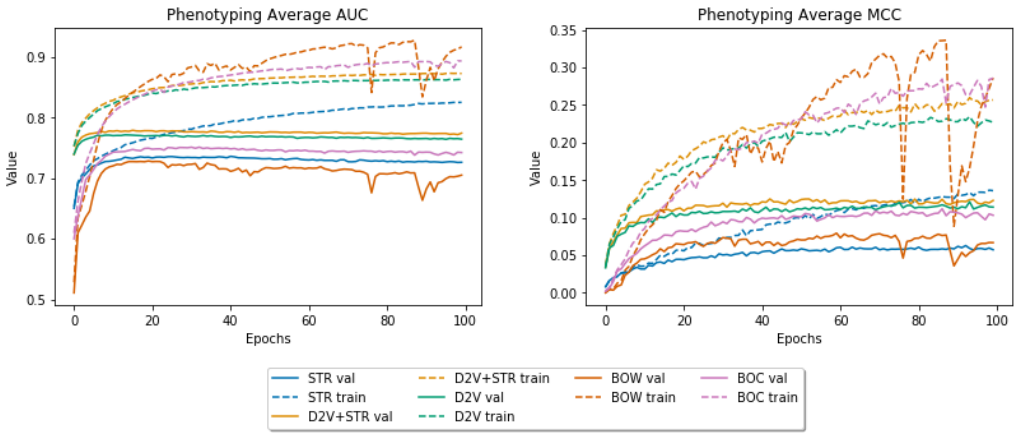


Figure 2: Validation MCC score and AUC-ROC score at each epoch of training systems on the 172-class phenotyping task with the standard LSTM architecture.

Experiments

We compared the LSTM's phenotyping performance when provided with the five different data source representations described earlier: (1) structured data from the patient's chart and lab results only (STR), (2) clinical notes only, encoded as a Bag of Words (BOW); (3) clinical notes only, encoded as a Bag of UMLS Concepts (BOC); (4) clinical notes only, encoded using Doc2Vec (D2V); and (5) a combination of structured data and Doc2Vec-encoded clinical notes (STR+D2V). In all experiments, models were trained for up to 100 epochs, using early stopping based on validation MCC (described below). The AUC and MCC on the validation set are shown at all epochs of training in Figure 2, where it is clear that, without early stopping, all data representations enable the LSTM to overfit well before reaching 100 epochs.

A. Evaluation Metrics

While accuracy is commonly used to evaluate the performance of classification systems, it is not particularly useful for imbalanced classes (in our case, less-prevalent phenotypes). To account for this, we reports metrics that are robust to class imbalance by taking into account the number of false positives associated with each class. We primarily used the Matthews Correlation Coefficient (MCC) as our metric for selecting the best epoch and combaring models during development. Equation (1) shows the Matthews Correlation Coefficient in terms of the number of False Negatives (FN), True Negatives (TN), False Positives (FP), and True Positives (TP):

$$\text{MCC} = \left(\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \right) \tag{1}$$

The advantage of using MCC for phenotype identification is that it gives a reliable measure of the model's performance on both common and uncommon phenotypes, whereas Recall, Precision, and AUC are hard to interpret when the class frequencies are imbalanced. Because we are evaluating joint-phenotype identification, we report the macro-average MCC across all 172 phenotypes (using a classification threshold of 0.5).

For evaluating the performance of our models we report the Area under the Receiver Operating Characteristic curve (AUC). This curve is obtained by plotting the model's True Positive Rate (TPR) against the False Positive Rate (FPR) for different values of the binary classification threshold. To determine whether different data sources resulted in statistically significantly different AUCs, we relied on the AUC confidence interval obtained using DeLong's Test.[27]. DeLong's Test is an asymptotically exact method to evaluate the uncertainty of an AUC plot, allowing us to determine 95% confidence intervals for the AUC produced by each data source for each phenotype.

Results

We report the aggregate phenotyping performance for all five data representations in Table 2. Specifically, we report the AUC, MCC, average Precision and average F_1 scores for 172 phenotypes as evaluated using a held-out testing set

of 5424 episodes.

For the sake of comparison, we also report the impact of each data representation for the other three tasks evaluated in Harutyunyan et al. (2019)[21] – In-hospital Mortality, Length of Stay, and Decompensation Prediction – as well as the 25 CCS phenotypes considered in that work. Table 3 reports these results, using the same metrics reported in Harutyunyan et al. (2019)[21]. For the Decompensation and Length-of-Stay tasks, we used the "deep supervision" formulation of the task in which the targets are predicted after every timestep. Note, for Length of Stay, MCC could not be computed; thus, we always tested the Length of Stay model that had the best kappa score when evaluated using the validation set. We found that the only task for which the clinical notes were definitively helpful was phenotyping.

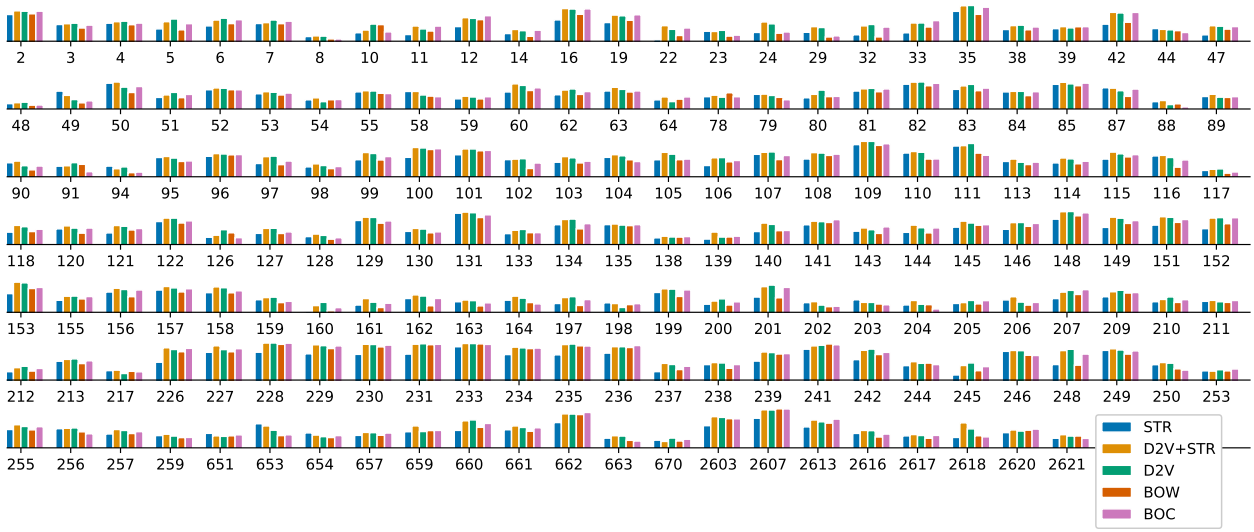


Figure 3: AUC with each data source for all 172 phenotypes (labeled by their CCS codes); the vertical axis represents AUC between 0.55 and 1.0.

Discussion

As shown in Figure 3, clinical notes had a strong positive impact on phenotyping performance. These results indicate that, on average, the combination of structured data and clinical notes was best, with regard to AUC-ROC, MCC, and average precision. In addition, clinical notes alone are superior to structured data alone.[†] Nonetheless, this trend is not true for every phenotype. Analyzing the performance on individual phenotypes and clinically-relevant categories of phenotypes can offer a more fine-grain explanation of why some phenotypes are highly amenable to being identified through clinical notes while others are not. We make use of the multi-level CCS categories (shown in Table 1) to group the clinical phenotypes and determine if phenotypes within certain categories are more easily identified using text (i.e., clinical notes) or text and structured data. Figure 3 shows the AUC-ROC obtained for each of the 172 phenotypes using

[†]We conducted additional experiments re-sampling structured data only during timesteps with notes, and found that even accounting for the frequency of structured data, clinical notes were still superior.

Table 2: Performance when identifying 172 clinical phenotypes from MIMIC-III hospital stays in the test set. Clinical phenotyping performance; Precision, Recall, and F_1 are weighted macro-averages based on phenotype prevalence.

Data	AUC (Micro)	AUC (Macro)	AUC (Weighted)	MCC	Precision	Recall	F_1
STR	85.55%	72.67%	73.04%	6.67%	30.56%	30.56%	16.69%
D2V+STR	87.84%	77.52%	77.37%	13.23%	37.20%	37.20%	25.05%
D2V	87.54%	76.83%	76.55%	12.58%	36.31%	36.31%	23.83%
BOW	84.50%	72.54%	72.48%	7.95%	30.87%	30.87%	20.94%
BOC	86.22%	74.85%	74.59%	10.46%	34.05%	34.05%	23.43%

Table 3: Performance on In Hospital Mortality, Decompensation, Length of Stay, and 25-Class Phenotyping using the metrics reported by Harutyunyan et al.[21]. The relative ranking of each approach is provided in parenthesis.

	In-Hospital Mortality			Decompensation		
Data	AUC-ROC	AUC-PRC	MCC	AUC-ROC	AUC-PRC	MCC
STR	0.854 (#1)	0.470 (#1)	0.383 (#1)	0.900 (#1)	0.309 (#1)	0.326 (#1)
BOW	0.716 (#5)	0.225 (#5)	0.168 (#5)	0.822 (#4)	0.167 (#4)	0.207 (#4)
BOC	0.770 (#4)	0.322 (#4)	0.263 (#4)	0.846 (#3)	0.211 (#3)	0.272 (#3)
D2V+STR	0.848 (#2)	0.461 (#2)	0.369 (#2)	0.885 (#2)	0.236 (#2)	0.273 (#2)
D2V	0.790 (#3)	0.345 (#3)	0.295 (#3)	0.821 (#5)	0.151 (#5)	0.192 (#5)

	Length of Stay			25-Class Phenotyping		
Data	MAD	MAPE	Kappa	AUC	MCC	Precision
STR	108.718 (#2)	185.011 (#2)	0.437 (#1)	0.739 (#4)	0.238 (#5)	0.448 (#4)
BOW	115.345 (#4)	161.749 (#1)	0.301 (#5)	0.732 (#5)	0.251 (#4)	0.445 (#5)
BOC	108.815 (#3)	260.538 (#4)	0.426 (#3)	0.752 (#3)	0.307 (#3)	0.481 (#3)
D2V+STR	108.348 (#1)	191.180 (#3)	0.427 (#2)	0.792 (#1)	0.369 (#1)	0.540 (#1)
D2V	118.734 (#5)	264.235 (#5)	0.404 (#4)	0.783 (#2)	0.352 (#2)	0.533 (#2)

each data representation.

A. How to Represent Clinical Notes

There where no phenotypes for which the Bag of Words had an AUC that was statistically-significantly higher than that of the Bag of Concepts. There were however 8 phenotypes for which the Bag of Concepts performed significantly better than the Bag of Words. Neither Bag of Words nor Bag of Concepts performs significantly better than Doc2Vec for any phenotype. For this reason, we did not consider combining structured data with BoW or BoC in our experiments.

B. When to Combine Clinical Notes with Structured Data

For 160 out of the 172 clinical phenotypes that we considered, we observed an increase in AUC when including clinical notes rather than solely using structured data, and for 51 of these the benefit is statistically significant. When breaking the phenotypes into the CCS categories shown in Table 1, the two categories in which the majority of the phenotypes exhibited a statistically significant increase in AUC when including text along with structured data were "diseases of the circulatory system" and "Injury and poisoning", indicating that it is generally worth including text when identifying phenotypes in these categories. These two categories of phenotypes accounted for 46% of the phenotypes for which combining clinical notes with structured data yielded a statistically significant improvement (diseases of the circulatory system accounting for 26% and injury and poisoning the remaining 20%).

The five phenotypes with the largest statistically significant increase in improvement when adding Doc2Vec to structured features were: "Fracture of neck of femur (hip)" (226), "Melanomas of skin" (22), "Gangrene" (248), "Secondary malignancies" (42), and "Gastrointestinal hemorrhage" (153). These conditions likely benefit from textual features because they all are primarily documented in radiology or surgery reports and are not directly indicated by vital signs or numerical lab measurements. Thus, we recommend combining clinical notes and structured data for phenotypes characterized by both discrete observations and continuous or structured measurements.

Figure 4 shows ROC curves for "Gastrointestinal hemorrhage", where clinical notes were particularly useful, alongside "Diabetes mellitus without complication", the phenotype for which notes were the most detrimental. The positions of the ROC curves for text and structured data sources are reversed between these two plots, with the ROC curve for structured data appearing below all other sources in the case of "Gastrointestinal hemorrhage", but above all the others in the case of "Diabetes mellitus without complication".

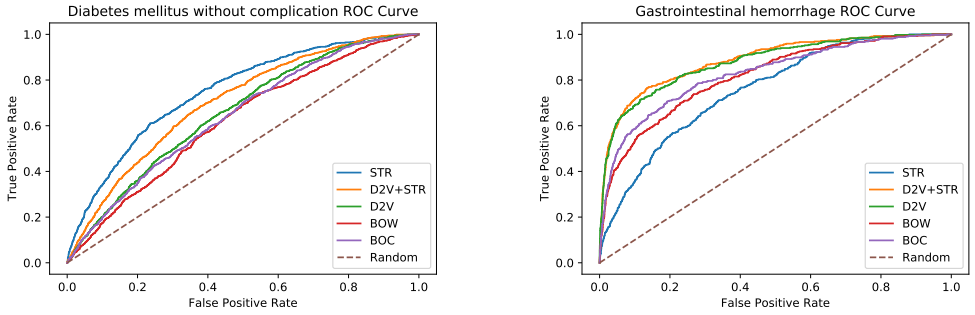


Figure 4: Receiver Operating Characteristic (ROC) curve when identifying phenotypes in the test set with the standard LSTM architecture trained on 172 phenotypes.

C. When to Use Only Structured Data

There were 121 phenotypes for which there was no statistically significant gain from combining clinical notes with structured data. Out of these, there were 12 phenotypes for which adding clinical notes to the structured data harmed the AUC performance at least slightly, but only one of these differences was statistically significant. The only phenotype for which the combination of the combination of text and structured data performed statistically-significantly worse than structured data alone was "Diabetes mellitus without complication" (49), from the "Endocrine; nutritional; and metabolic diseases and immunity disorders" category. A likely reason for this result is the prevalence of Glucose, which is the main measure used for diagnosing Diabetes,[28] in the chart and lab measurements and the inability of Doc2Vec to identify glucose values or severity indicators from the text.

In addition, there were 5 CCS categories for which no phenotypes had a statistically significant improvement when including clinical notes with structured data: "Congenital anomalies", "Diseases of the musculoskeletal system and connective tissue", "Diseases of the nervous system and sense organs", "Diseases of the skin and subcutaneous tissue", and "Endocrine; nutritional; and metabolic diseases and immunity disorders". Table 1 shows the prevalence of each of these groups in our data set. While it is possible that under-performance of text may be influenced by the fact that these phenotype categories are underrepresented in MIMIC-III compared with categories that are frequently treated in an ICU setting such as "Injury and Poisoning", we believe the low level of difference in performance for these categories between text-based and structured-data-based identification seems to indicate that text is less useful for these categories of phenotype. Moreover, because Doc2Vec learns a task-agnostic representation of clinical notes, it will naturally be better able to encode information about more commonly documented observations, suggesting that perhaps context-aware representations such as those learned by BERT may be better suited for recognizing less prevalent phenotypes.

D. When to Use Only Clinical Notes

While adding Doc2Vec-encoded text to structured data often improved results when compared with using structured data alone (as explained above), using the Doc2Vec encoded-text alone was never statistically significantly worse than using the combination of Doc2Vec and structured data, with the exception of two phenotypes: "Diabetes mellitus with complications" and "Diabetes mellitus without complication". This leads us to believe that while Scheurwegs et al. (2015)[11] found that clinical notes and structured data were generally complimentary when using a Bag of Words to represent clinical notes, switching to a better-performing representation of clinical notes such as Doc2Vec can render the addition of structured data mostly redundant.

E. Limitations and Future Work

It is possible that many of the advantages of clinical notes come from the fact that medication names, dosages, and administration instructions are often strongly associated with specific medical conditions. For this reason it would be worth considering including other types of structured data such as prescriptions rather than just laboratory and chart measurements. In addition, we would like to identify specific clinical and textual features that make some phenotypes particularly easy to identify based on note text. While our LSTM-based approach was effective for taking advantage of the chronological nature of EHR data, due to nonlinearities in the network, it does not allow us to easily determine

which features were most important when identifying each phenotype.

In future work, we would like to use a neural network that is more likely to take advantage of the linguistic complexities of clinical notes. For example, it is possible that other methods such as transformer models[29] would be more successful at extracting useful information from text, and could also potentially be useful for improving the timeseries processing of structured data. We would also like to explore in future work the extent to which the information contained in clinical notes and structured data tend to overlap to better inform our analysis and intuitions of whether using clinical notes will be helpful for a given task. We would also like to continue our work with the other three tasks from the MIMIC-III Benchmark, to determine whether there is a way to better harness text such that using clinical notes would improve rather than reduce performance on those tasks.

Code Availability

The code for this work can be found at `github.com/amoldwin/notes_benchmark`.

Conclusion

Exploring the effect of including clinical notes for EHR-based phenotyping, we found that phenotyping performance generally benefited greatly from the inclusion of clinical notes. We observed that while on average there was a significant difference in performance between systems that use structured data exclusively and those that use a combination of text and structured data, some groups of phenotypes, such as "diseases of the circulatory system" and "injury and poisoning" are most likely to benefit in a statistically significant way from the combination of structured data and text, while others can be detected reliably from text alone. When comparing different text representations, the Bag of Concepts tended to be more effective than the Bag of Words approach, but was consistently less effective than Doc2Vec, indicating that the word order and document structure are in fact important for clinical phenotyping. By utilizing a broad array of common phenotypes, we were able to compare the efficacy of these systems across a wide array of phenotypes, allowing us to determine how generalizable our findings were. By taking into account our findings, future researchers may decide to rely on text rather than structured data for phenotyping applications, if given a choice between the two. We hope that this study will help inform researchers and clinicians in situations where it is necessary to design task-specific phenotyping systems for cohort selection and clinical decision support applications.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health and utilized the computational resources of the NIH HPC Biowulf cluster (`http://hpc.nih.gov`).

References

1. Atasoy H, Greenwood BN, Mccullough JS. The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization. Annual Review of Public Health. 2019;40(1):487–500.
2. Romano MJ, Stafford RS. Electronic Health Records and Clinical Decision Support Systems. Archives of Internal Medicine. 2011;171(10).
3. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. Journal of the American Medical Informatics Association. 2011;18(2):181–186.
4. Apostolova E, Wang T, Tschampel T, Koutroulis I, Velez T. Combining Structured and Free-text Electronic Medical Record Data for Real-time Clinical Decision Support. Proceedings of the 18th BioNLP Workshop and Shared Task. 2019;.
5. Rossetti SC, Knaplund C, Albers D, Tariq A, Tang K, Vawdrey D, et al. Leveraging Clinical Expertise as a Feature - not an Outcome - of Predictive Models: Evaluation of an Early Warning System Use Case. AMIA Annu Symp Proc. 2019;2019:323–332.
6. Sun M, Baron J, Dighe A, Szolovits P, Wunderink RG, Isakova T, et al. Early Prediction of Acute Kidney Injury in Critical Care Setting Using Clinical Notes and Structured Multivariate Physiological Measurements. Studies in Health Technology and Informatics. 2019;264:368–372.
7. Denny J. Chapter 13: Mining Electronic Health Records in the Genomics Era. PLoS computational biology. 2012 12;8:e1002823.
8. Ludvigsson JF, Pathak J, Murphy S, Durski M, Kirsch PS, Chute CG, et al. Use of computerized algorithm to identify individuals in need of testing for celiac disease. Journal of the American Medical Informatics Assn. 2013;.

9. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. Plos One. 2018;13(2).
10. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-Treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care & Research. 2010;62(8):1120–1127.
11. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Bulcke TVD. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. Journal of the American Medical Informatics Association. 2015;23(e1).
12. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. Summit Transl Bioinform. 2009 Mar;2009:116–120.
13. Sinnott JA, Cai F, Yu S, Hejblum BP, Hong C, Kohane IS, et al. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. Journal of the American Medical Informatics Association. 2018 05;25(10):1359–1365. Available from: https://doi.org/10.1093/jamia/ocy056.
14. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. Journal of the American Medical Informatics Association. 2015 09;23(e1):e20–e27. Available from: https://doi.org/10.1093/jamia/ocv130.
15. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997 Nov;9(8):1735–1780. Available from: https://doi.org/10.1162/neco.1997.9.8.1735.
16. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. npj Digital Medicine. 2019;2(1).
17. Alzoubi H, Alzubi R, Ramzan N, West D, Al-Hadhrami T, Alazab M. A Review of Automatic Phenotyping Approaches using Electronic Health Records. Electronics. 2019;8(11):1235.
18. Rethinking Clinical Trials;. Available from: https://sites.duke.edu/rethinkingclinicaltrials/informed-consent-in-pragmatic-clinical-trials/.
19. Nunes AP, Yang J, Radican L, Engel SS, Kurtyka K, Tunceli K, et al.. Assessing occurrence of hypoglycemia and its severity from electronic health records of patients with type 2 diabetes mellitus. U.S. National Library of Medicine; 2016. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27744128.
20. Johnson AE, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3(1).
21. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Scientific Data. 2019;6(1):96. Available from: https://doi.org/10.1038/s41597-019-0103-9.
22. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. Journal of the American Medical Informatics Association. 2017;.
23. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004;32(90001).
24. Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32; 2014. .
25. Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, John RS, et al. Universal Sentence Encoder for English. ACL Anthology;Available from: https://www.aclweb.org/anthology/D18-2029/.
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT; 2019. .
27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837–845.
28. Sacks DB, Bruns DE, Goldstein DE, Maclaren NK, McDonald JM, Parrott M. Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. Clinical Chemistry. 2002 03;48(3):436–472. Available from: https://doi.org/10.1093/clinchem/48.3.436.
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 5998–6008.

EFFECT OF REOPENING ORDERS ON COVID-19 HOSPITALIZATIONS IN THE US

Ron N. Nachum[1,*], William A. Ding[2,*], Logan B. Pageler[3,*], Nikhil R. Majeti[4,*], Jesutofunmi A. Omiye, MD[5]

[1]Thomas Jefferson High School for Science and Technology, Alexandria, Virginia, USA, [2]San Mateo High School, San Mateo, California, USA, [3]Homestead High School, Cupertino, California, USA, [4]Palo Alto Senior High School, Palo Alto, California, USA, and [5]Stanford Centre for Biomedical Informatics Research, Stanford University, Stanford, California, USA.

*These authors contributed equally
Corresponding author: Jesutofunmi A. Omiye, Stanford Centre for Biomedical Informatics Research, Room X-235, Medical School Office Building, 1265 Welch Road, Stanford, CA 94305-5479; tomiye@stanford.edu

Abstract

Shelter in place (SIP) orders were instituted by states to alleviate the impact of the COVID-19 pandemic. However, states proceeded to reopen as SIPs were noted to be hurting the economy. We evaluated whether these reopenings affected COVID-19 hospitalizations. We collected public data on US state reopening orders and COVID-19 hospitalizations from March 8 to August 8, 2020. We utilized a doubling time metric to compare increase in hospitalizations in line with reopenings and proceeded to quantify the impact of reopening orders on cumulative hospitalizations. We found that some reopenings increased hospitalizations, and this varied by state. We also discovered that the most negatively impactful reopenings overall tended to be restaurants/bars (-92%) and houses of worship (-63.6%). Without data-backed guidance on reopening states, the healthcare burden from COVID-19 will likely persist. State governments should use data to understand the potential effects of these reopenings to guide future policies.

Introduction

COVID-19, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first reported in Wuhan, China on December 1, 2019[1]. Over eight months, the disease spread to 213 countries and territories in the world with 17 million confirmed cases and over 687,000 deaths as of August 1, 2020[2]. The first reported case in the United States was in the state of Washington on January 21, 2020, and the first documented deaths were in February. Since then, COVID-19 has spread to all 50 states with over 5 million cases and 160,000 deaths[3].

As there was no effective treatment or vaccine available, public health officials and policymakers proceeded to deploy significant mitigation strategies which included, but were not limited to, Shelter in Place (SIP) orders, public health education on handwashing, and use of face masks. The US was one of the many countries to adopt SIP orders. However, they were enforced heterogeneously, with several states and counties implementing different orders at different rates throughout March and April, with no coordinated central government effort[4,5]. Studies done on SIP orders were noted to have reduced SARS-CoV-2 transmission in several states within the US, and other countries like China and Italy[4]. For example, Tian and colleagues found delayed case transmission corresponding to SIP measures[6]. Also, Badr et. al, found social distancing measures correlated with a decreased COVID-19 infection rate[4].

In evaluating the effects of COVID-19, case count is noted to be an inaccurate estimate of health systems burden, especially in the absence of state-wide testing, and is more an index of the testing capacity of states[7]. Death count, another metric also lags behind disease spread and is not useful for policy strategies towards alleviating COVID-19 burden. Hospitalization rates have emerged as the most useful index of healthcare systems burden, and they are less likely to be dependent on testing capacity[8]. This finding is buttressed by a recent study by Kashyap et. al., which noted hospitalization rates to have slowed with SIP, even as cases continued to rise[9].

Despite the beneficial effects of SIP on COVID-19 mitigation, SIP worsened unemployment, interest rates, and consumer spending[10]. As a result, many state governments proceeded to reopen their economies via reopening orders (ROs) in phases of 1-3, as they deemed safe. Similar to SIP orders, ROs were not centrally coordinated and were varyingly implemented across and within states. Without analysis of the effects of the reopening, many states started to encounter significant rises in cases and hospitalizations, as they reopened. In many cases, ROs were deemed responsible for this change leading to reversals of ROs in several states. It is clear that reopening as many businesses

as possible while suppressing transmission of SARS-CoV-2 and rate of COVID-19 is necessary to protect the economy and the people.

To provide data derived guidance in achieving this balance, we analyze the effects of reopening orders issued by state governments on COVID-19 hospitalizations. This study, therefore, examines the changes in trends of hospitalizations across US states, in light of reopening policies. Our goal is to highlight what types of ROs correspond to the most increases in hospitalizations per US state. We, further, use exponential regression analysis to predict the outcome of reopening orders and possibly advise future policy choices.

Methods

Data Collection

We obtained two main forms of data for our study: cumulative COVID-19 hospitalizations and reopening orders across all US states. Data on hospitalizations was retrieved from The Atlantic COVID Tracking Project: a publicly available data source that collects data from 56 US state and territory public health authorities, including official statements from state officials. The data is further verified by a team of human volunteers and updated daily between 5 and 6 pm eastern time. Data on racial demographics, testing, and patient outcomes are provided on a state level, and a data quality grade is assigned[11]. Our data collection process was automated via the Atlantic COVID Tracking Project API for data on COVID-19 confirmed cases, tests, cumulative hospitalizations, current hospitalizations, and deaths[11]. Since hospitalizations provide the most consistent indicator of disease spread, we focused on collecting hospitalization data. This was further narrowed to cumulative hospitalizations for the 41 states that provided them. Cumulative hospitalization was the chosen metric as it was easier to derive new hospitalizations from these data, evaluate changes in spread over time while noting recoveries or deaths. It is also useful in assessing the impact on health systems and deriving the doubling time metric used in our study.

For the nine states and the District of Columbia that provided only current hospitalizations, we developed an algorithm to convert current hospitalization counts to cumulative hospitalizations. For this, we hypothesized an average length of hospital admission for patients infected with COVID-19 across all states. Wang et. al., and Wu et. al., alongside other sources, report a significant variation in the length of stay, with studies averaging between 10-16 days[12–17]. Due to this variation, we employed a trial and error algorithm to calculate the most reflective hospital stay time, by evaluating states that report both current and cumulative hospitalization count. For all states which recorded both forms of data, we used this estimation algorithm with stay times ranging from 5 to 25 days and found the mean squared error between the estimated result and actual cumulative hospitalizations. Through this process, we found 14 days to be the most optimal hospital stay time, and its robustness was evaluated via our sensitivity analysis. For practicality purposes, we assumed that patients will be admitted for an integer number of days. Also, since none of the states had data dating back to the first day COVID-19 impacted them, this led to actual cumulative hospitalizations being consistently greater than the calculated figures. To limit the effect of this, we took all the states with both data sets dating back to before their closing dates and found that cumulative hospitalization was approximately 1.8 times greater than current hospitalizations with a 0.1 times standard deviation. This was additionally implemented to convert current hospitalizations to cumulative hospitalizations.

Data on reopening orders were collected from the New York Times (NYT) “See How All 50 States Are Reopening (and Closing Again)” website. This is also publicly available data that provides daily updated state-issued reopening orders. The website aggregates data on ROs from state governments, executive orders, and local news reports[18]. We found it to be the most consistent public data source on ROs for the 50 states and the District of Columbia. We chose to gather data on state-issued reopening orders, in place of county orders as they were readily available, accurate, and consistent. Since the NYT webpage only reflected the most current reopening policies and did not archive changes, we used the Internet archive’s Wayback Machine to access historical versions of the webpage. Specifically, for each day from May 1 to July 28, 2020, we web scraped information from the version archived closest to 12 pm EST. To document changes in reopening policy, we compared information from consecutive days. We further cleaned up the NYT data, accounted for reopening reversals, and aggregated vague categories into more defined ones.

We ended up with fourteen total reopening categories for all states (Table 1). These categories were chosen based on the most common comma-separated phrases on the NYT website (Table 1). The format of our data dictionary implies that, at any given time and within a given state, a category is either completely open or completely closed. Thus, to avoid ambiguity over a “partial” reopening, we split overly general categories into more specific subcategories. For

example, a preliminary analysis of when certain comma-separated phrases were added to the article revealed that bars tended to reopen well after indoor restaurant dining, which likewise tended to open after outdoor restaurant dining. As a result, we made three distinct subcategories for restaurant dining and bars, instead of having a single "Restaurants and bars" category. On the other hand, we grouped related businesses into single categories if they were frequently reopened on close or identical dates. For example, our preliminary analysis revealed that museums, casinos, movie theaters, and several other businesses all tended to reopen on the same day, so they were grouped into the "Indoor entertainment" category. In the rare situation where one kind of indoor entertainment business reopened before another, only the earliest reopening was considered and used in our study.

Table 1: Reopening categories used for analysis and the frequency of the NYT comma-separated phrases

Reopening category	Supercategory	NYT comma-separated phrase	Frequency
Outdoor dining only	Food and drink	"retail stores"	60
Indoor dining, no bars	Food and drink	"gyms"	48
Indoor dining with bars	Food and drink	"barbershops"	43
Limited retail (curbside pickup, etc.)	Retail	"restaurant dining"	40
Full retail	Retail	"tattoo parlors"	33
Hair salons and barbershops	Personal care	"salons"	31
Non-hair personal care	Personal care	"hair salons"	29
Gyms and fitness centers	None	"museums"	29
Indoor entertainment	None	"movie theaters"	25
Office environments	None	"retail stores open to curbside pickup"	25
Houses of worship	None		
Construction	None		
Beaches	None		
Campgrounds and state parks	None		

Data Analysis

To evaluate both datasets and find the impact of ROs on COVID-19 hospitalizations, we utilized the doubling time of hospitalization rate, a vital measure of hospitalization growth, which is comparable across states and can be used to model subsequent hospitalizations in a single-variable form. It is also expected to increase with effective mitigatory measures[19]. The utility of doubling time is evident when looking at the spread in the entire United States during June, in which infections, hospitalizations, and deaths began to increase significantly in most parts of the country, with lower rates in some northeastern states. By observing the rate of change of doubling time, we can largely eliminate noise caused by previous outbreak history to understand current and project future trends.

The formula for doubling time used was: $t * log(2) \,/\, log(\frac{h_2}{h_1})$ Where t is the window for doubling time, and as a result, h_1 and h_2 are the numbers of cumulative hospitalizations at the start and end of the window respectively. The window chosen was 7 days as it ensured that weekday reporting fluctuations would not remarkably affect doubling time. After using a state's cumulative hospitalization to calculate doubling time, ROs were then overlaid on this data.

The next step was determining the period in which ROs would be expected to have an effect, as any change in hospitalizations resulting from a given reopening event would not be observed immediately. We assumed 14 days, which was further tested, alongside other days, through our sensitivity analysis. Once this period was determined, the average rate of change of doubling time was calculated from the reopening to 14 days after, and then 14-28 days after the reopening. These two values were subsequently compared to calculate a percentage difference before and after the reopening, to determine if the reopening had a negative effect on disease spread by decreasing the doubling time's rate of change. Negative values are interpreted as either changing the doubling time curve from a sharp increase to a lesser rate of increase or even forming a peak and starting to decrease again. If reopening orders for a particular state had negative values, they would be used to estimate the added cumulative hospitalizations of the respective state.

By using doubling time before reopening and attempting to continue the current trend, prediction curves can be generated for the case in which a certain reopening had not occurred using exponential regressions. Once these were calculated, we compared the non-reopening trajectory with the least ultimate hospitalizations to the true total, and this is the difference between if none of the negative reopenings had occurred to the ground-truth value. We showcase this sequential process with the state of Florida below (Figure 1).

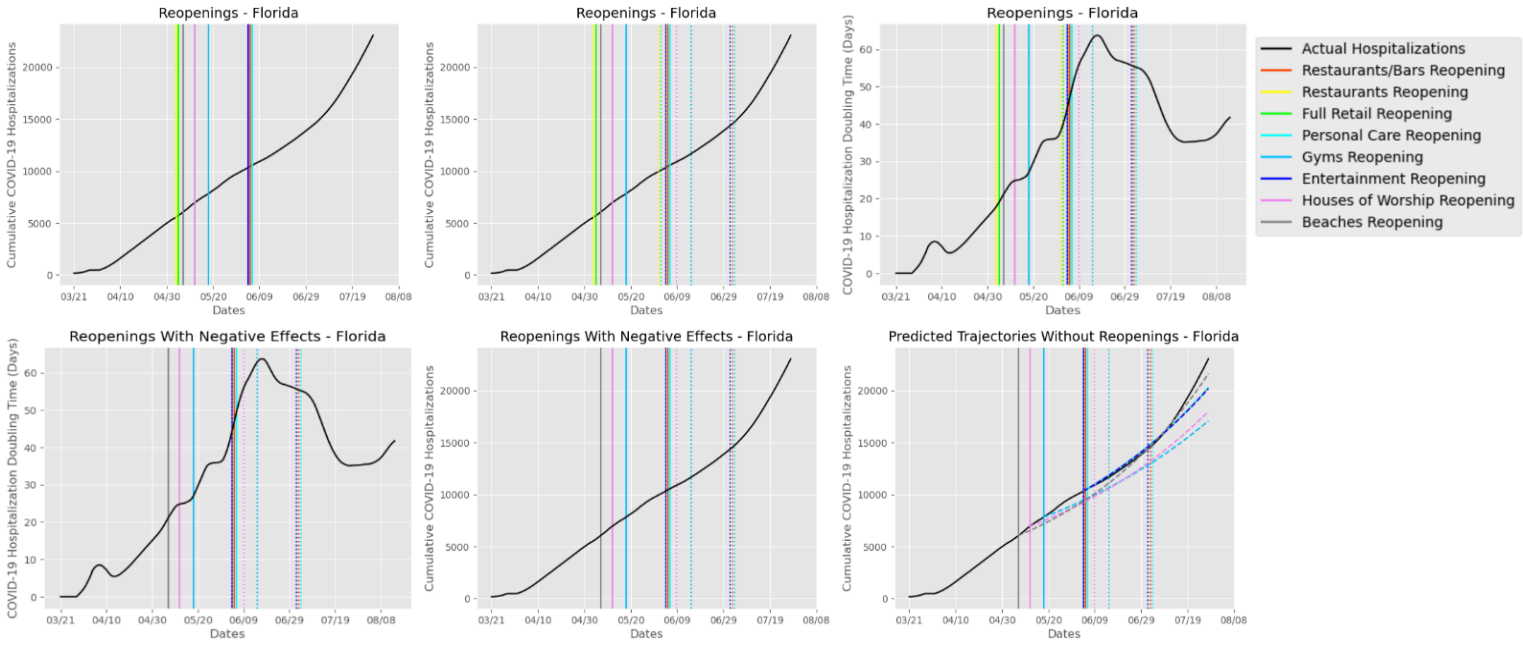


Figure 1. Visualization of the analysis process for the state of Florida, with reopenings overlayed on graphs of cumulative hospitalizations and doubling times, and only reopenings with negative effects are shown in the bottom 3 charts

Sensitivity Analysis

We evaluated the accuracy of our assumptions using sensitivity analyses. Our first sensitivity analysis was on the assumption of the length of hospital stay, which we used to calculate cumulative hospitalizations from current hospitalizations when unavailable. We tested a range of 10-16 days retrieved from various studies and evaluated each day with states that had data for both current and cumulative hospitalizations[12,13,15]. We found that 14 days provided the least error when comparing our calculated cumulative hospitalization counts to the actual values. Another sensitivity analysis was done on the most representative value of the doubling time window in light of the incubation period of COVID-19. The 7-day window was noted to decrease reporting error impact, reduce weekday fluctuations, and not overly smoothen the data.

Ultimately, in determining the ROs' effect on cumulative hospitalizations, we hypothesized an effect time of 14 days, accounting for the incubation period, disease spread, and symptom worsening. By varying this lag time through 7, 21, and even 28 days, our conclusions remained largely consistent. Although, higher values resulted in increasingly negative effects on hospitalizations, which could have been attributed to possibly more confounding variables associated with larger windows.

Results

All 50 states and the District of Columbia implemented at least one reopening order during the period of May, 28 to August 01, 2020**,** and had hospitalization data available. Our focus was on measuring hospitalization trends and determining which ROs had the greatest negative effects on COVID-19 hospitalizations across the country. Although our analysis was on all 50 states and the District of Columbia, four states are presented as case studies in this paper for brevity purposes. The states are Florida, California, South Carolina, and Arkansas. These four states were selected,

due to their increasing number of COVID-19 hospitalizations—Florida and California, and their low number of COVID-19 hospitalizations—South Carolina and Arkansas, as at the time of writing this paper. Cumulative hospitalization trajectory was superimposed on particularly impactful reopening categories, and this is shown in (Figure 2).
The most negatively impactful reopenings varied by state. For example, the most negatively impactful reopenings for Florida were restaurants/bars, entertainment, and personal care, while for California, these were houses of worship and hair salons, followed by restaurants/bars, gyms, and entertainment. For South Carolina, personal care, hair salons, and gyms had the most negative effects on hospitalization and for Arkansas, the reopenings of gyms, hair salons, personal care, indoor dining, entertainment, and houses of worship saw the greatest increase in hospitalizations.

Using these calculations, estimates for the number of hospitalizations resulting from reopenings in a state can also be measured. For example, for ROs with the most negative reopening effects in Florida state, 7,100 hospitalizations are estimated to have resulted – an additional 33%. California reopenings resulted in an additional 6,700 hospitalizations (21%), while South Carolina's reopenings are associated with an additional 2,000 hospitalizations (55%), and Arkansas reopenings are estimated to have resulted in an additional 1,500 hospitalizations (99%).

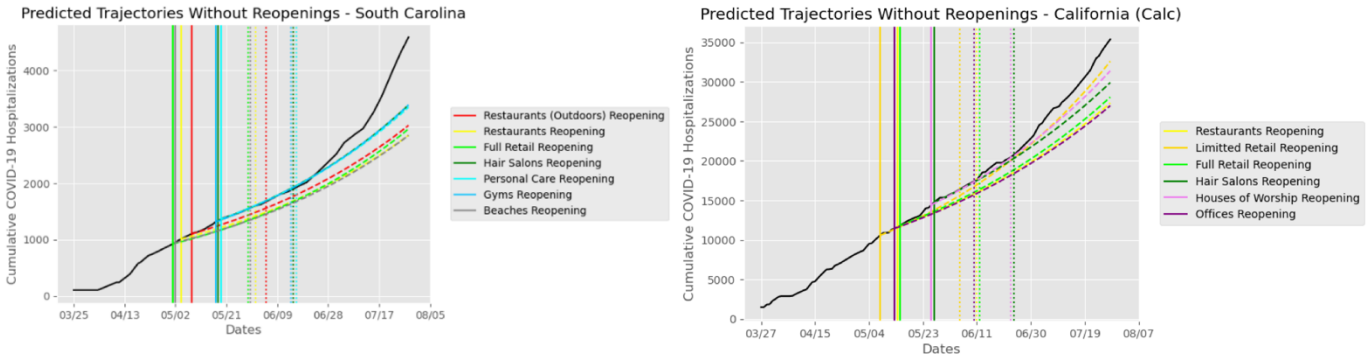


Figure 2. Predicted trajectories without reopenings for states of South Carolina and California

Impact of ROs
To further evaluate the impact of these ROs, percentage hospitalization changes during the period of reopening were derived for all states. This is distribution is shown in (Figure 3). To draw conclusions about which reopening types truly had the most negative effect, it was essential to balance out the effects of joint reopenings in states as well as individual outlier events such as protests or funerals. Looking at data from all states together ensured that individual reopening combinations did not affect the results and prevented unrelated events from significantly skewing the data.

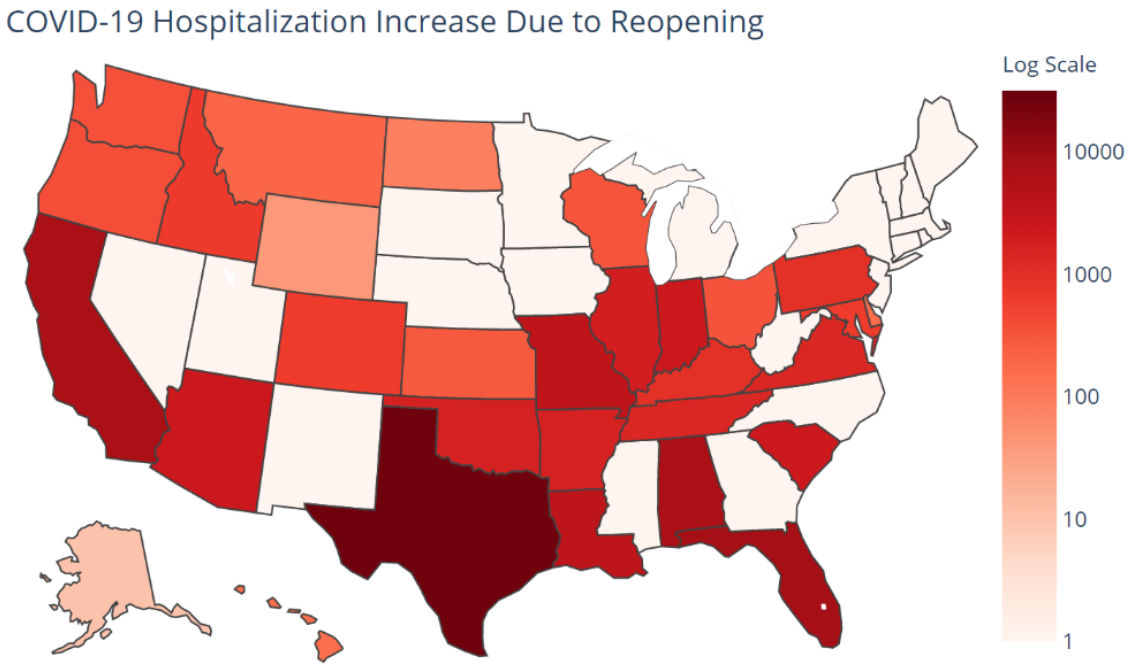


Figure 3. COVID-19 hospitalization changes across US states during reopenings

Comparative analysis impact of ROs

Among individual ROs, we proceeded to determine which had the most negative effects, for the US as a whole. The metric used to determine the effect of these reopenings was the percentage difference in the rate of change of doubling time. Based on previously described assumptions, this meant observing the rate of change of doubling time 0-14 days after a reopening occurred, and then 14-28 days after the reopening to calculate the percent difference between the two. It is important to note that a negative percentage shows that a reopening had a negative overall effect, while positive ones suggest that the reopening largely did not have a negative effect. We took the median of this value across all states to ensure that data was not skewed by outliers.

The results are shown in (Figure 4), in which it is evident that the reopenings of indoor dining with bars (-92.0%) and houses of worship (-63.6%) had the greatest negative effects of the reopening categories studied, while indoor entertainment (-24.4%), indoor retail (-23.7%), and hair salons (-22.4%) also increased community disease spread, as indicated by a change in hospitalization rate. Combined, these reopenings added an estimated 72,000+ COVID-19 hospitalizations nationwide. We observe a key pair here: restaurants/bars and restaurants (outdoors). These reopenings are very similar and often considered in a sequential manner. By observing doubling time rates of change across the country before and after each category of reopening, we can see that restaurants and bars had a much more negative effect than restaurants with outdoor seating and no bars. Some of the seemingly safer reopenings included outdoor retail (+183.8%), construction (+98.3%), parks (+72.1%), and outdoor dining (+65.2%).

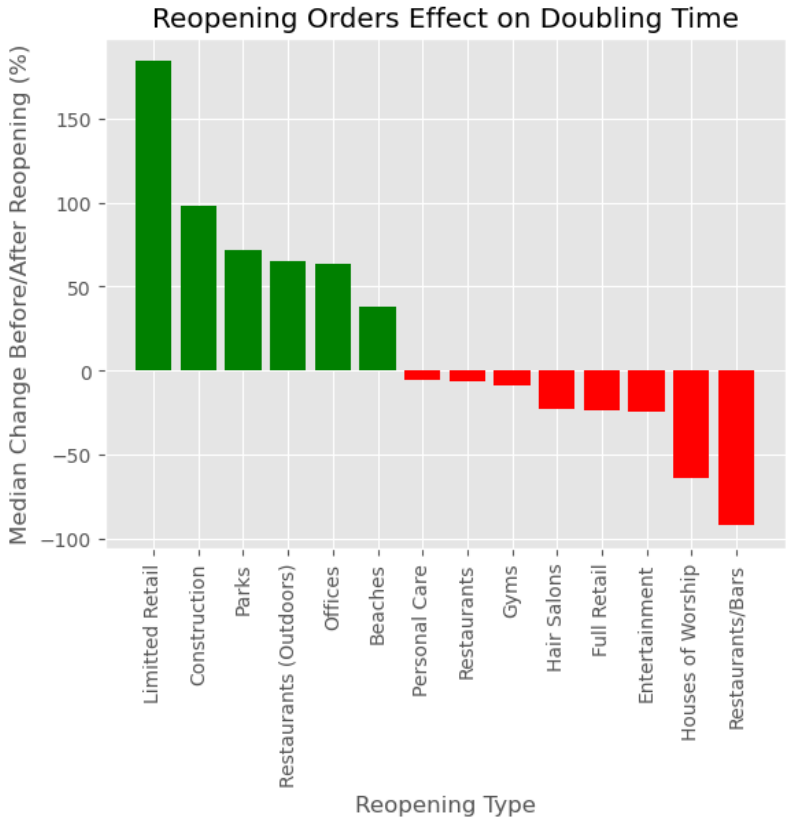


Figure 4. The median effect of reopening orders on the doubling time rate of change across all 50 states

Sensitivity Results Analysis

We hypothesized four variations of 7-day lag times, as follows: 7, 14, 21, and 28 days on the ROs effect on doubling time, and proceeded to analyze if the same reopenings will consistently have negative effects on doubling time. Results from this analysis are shown in (Figure 5) below, and it can be seen that restaurants/bars, houses of worship, entertainment, full retail, hair salons, and gyms appear to have the most negative effects across the 4 lag-times tested.

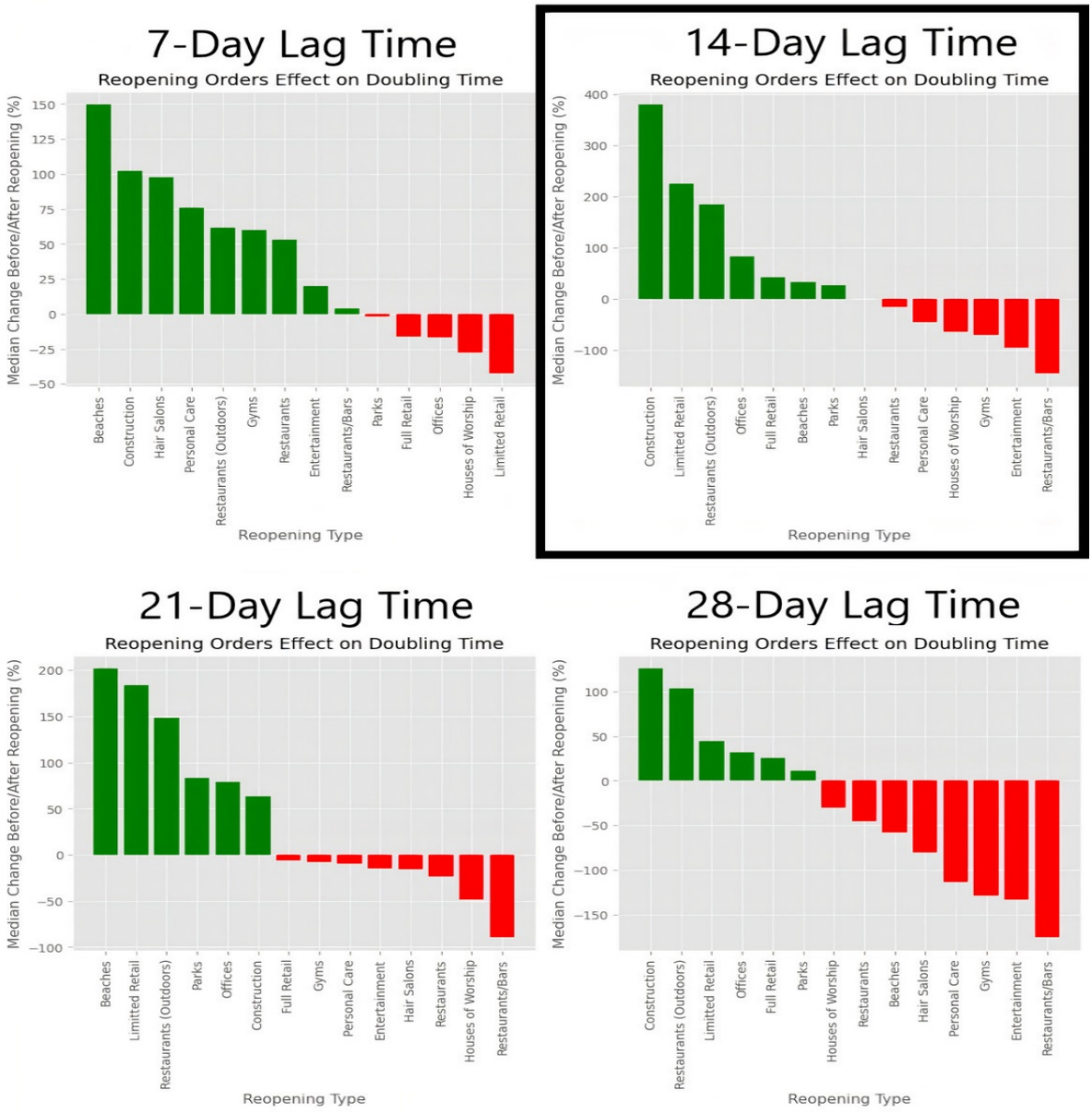


Figure 5. Sensitivity analysis showing comparison of the four different lag times used to evaluate the effect on doubling time

Discussion

Since the onset of the COVID-19 pandemic in the USA, government responses have been varied and heterogeneous. Variation in recommended mitigatory actions ranged from issuing SIP orders to encouraging non-pharmacological interventions like the use of facemasks. Subsequently, when governments proceeded to reopen their economies due to the socioeconomic impact of SIP, it was not surprising that these orders were also implemented very differently within and across US states.

In this study, we demonstrate that state ROs were generally associated with negative effects on COVID-19 hospitalizations, and doubling time, our index of disease spread. We further show that particular categories of

reopening are associated with more negative effects, compared to others: entertainment, restaurant/bars, houses of worship, full retail, hair salons, gyms. We estimate that these reopenings have led to an additional 57,000 COVID-19 hospitalizations nationwide. Using Florida, a state that, as at the time of this writing has become the epicenter of the pandemic, as a case study, one can better contextualize the effects of ROs, including timing, changes in hospitalization rate doubling time, and resultant hospitalizations. As shown in (Figure 1), reopenings that occurred before May 11 - 28 had no negative effects on the hospitalization rate doubling time, while reopenings from after May 28 up until June 14, had significant negative effects.

The results from our study show strong associations between particular categories of reopening and COVID-19 hospitalizations. Also, our study shows that the effects of these ROs are not likely to show significant effects until at least 14 days after enactment, and could be longer in some cases. This lag time likely indicates the period for transmission and worsening of symptoms enough to warrant hospitalization. Our study also provides a template for modeling effects of particular reopenings, and this can be applied to model similar reopenings for territories outside the US. The findings from our study suggest deleterious effects in accelerating reopenings that include restaurants, bars, gyms, full retail, and hair salons. This study provides one of the first datasets on the effects of ROs. Our findings and dataset can be used to further advise policymakers on possible future ROs and what it could mean for COVID-19 burden. It is essential to note that different states have different demographics, GDP per capita, and health security index which could impact how ROs affect hospitalizations.

This study is not without limitations. First, we relied on data from the Atlantic COVID Tracking Project and the New York Times, which tracks data from US state public health authorities. This data is of varying quality across all states, with minor inconsistencies and data lag. Our analyses also relied on assumptions that could potentially impact our findings. For example, the length of COVID-19 hospitalization is highly variable, depending on the presence of comorbid renal, respiratory, and cardiovascular diseases, which could complicate recovery from COVID-19[12,15]. Thus, our method to convert current hospitalizations to cumulative hospitalizations may have introduced errors in our downstream analyses. Also, while we tracked state-level policies, the authors recognize that some counties had different policies within a state and could be more useful to further track county reopenings. Besides, mobility data could have provided more information on people's behavior and answered questions on specific movements of people to documented places e.g. restaurants, and bars, as soon as they reopened. Research on association with mobility data and SIP showed individual behaviors sometimes occurring ahead of state policies[4]. Again, the socioeconomic impact of COVID-19 cannot be overemphasized and this has led to a rise in health disparity, loss of health insurance, further translating to decreased hospital utilization and overall hospitalization count. Also, varying COVID-19 hospitalization criteria across different states will potentially affect hospitalization data. It is important to note that we are examining data from different states, with varying public health strengths and the implementation of multiple non-pharmacological interventions that are hard to quantify which could have ultimately affected baseline hospitalizations. i.e. facemasks, and handwashing. Additional confounders, like the protests that occurred, could have also impacted the findings from our study.

Conclusion and Possible Future research

Our results support the conclusion that reopening aggravated the burden of the COVID-19 pandemic in the US, from June to August 2020. Some reopening choices such as reopening bars, restaurants, and places of worship, have clearly shown to have a worse effect on hospitalization rates than others. While it is paramount to balance the deleterious socio-economic effects of the pandemic, it is necessary to quantify the additional hospitalizations that result from the different choices. Our analysis attempts to provide that quantification. As there is no current effective pharmacological intervention available to curb this pandemic, carefully planned data-backed reopening guidance is necessary. Further studies will be needed to utilize mobility, states' demographics, and possibly contact tracing data to evaluate more specific transmission trends that correlate with reopenings, alongside hospitalizations. In addition, to promote future research on related projects, all our work, including data and results for all the states, is publicly available here: https://covid-reopenings.herokuapp.com/.

Acknowledgments

We acknowledge Prof. Nigam Shah, from the Stanford Centre for Biomedical Informatics Research, for assembling the team, outlining the potential project, providing insightful contributions, and reviewing the manuscript of this study. We also thank Dr. Alison Callahan for reviewing the draft of this paper.

References

1. WHO. Novel Coronavirus (2019-nCoV) Situation Report- 1 [Internet]. WHO. [cited 2020 Jul 28]. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4
2. Worldometer. Coronavirus Update (Live): 17,990,226 Cases and 687,690 Deaths from COVID-19 Virus Pandemic. [Internet]. [cited 2020 Aug 1]. Available from: https://web.archive.org/web/20200801235959/https://www.worldometers.info/coronavirus/
3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Vol. 20, The Lancet Infectious Diseases. Lancet Publishing Group; 2020. p. 533–4.
4. Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. Lancet Infect Dis. 2020;3099(20):1–8.
5. Auger KA, Shah SS, Richardson T, Hartley D, Hall M, Warniment A, et al. Association between Statewide School Closure and COVID-19 Incidence and Mortality in the US. JAMA - J Am Med Assoc. 2020;45229.
6. Tian H, Liu Y, Li Y, Wu C-H, Chen B, G Kraemer MU, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. 2020.
7. Sen S, Karaca-Mandic P, Georgiou A. Association of Stay-at-Home Orders with COVID-19 Hospitalizations in 4 States. JAMA - J Am Med Assoc. 2020;323(24):2522–4.
8. Miller IF, Becker AD, Grenfell BT, Jessica Metcalf CE. Disease and healthcare burden of COVID-19 in the United States. Nat Med. 2020;26:1212–7.
9. Kashyap S, Gombar S, Yadlowsky S, Callahan A, Fries J, Pinsky BA, et al. Measure what matters: Counts of hospitalized patients are a better metric for health system capacity planning for a reopening. J Am Med Informatics Assoc. 2020;00(0):1–6.
10. Ihrig J, Weinbach GC, Wolla SA. COVID-19's Effects on the Economy and the Fed's Response. Page One Econ.
11. The COVID Tracking Project. The COVID Tracking Project | The Atlantic [Internet]. 2020 [cited 2020 Jul 20]. Available from: https://covidtracking.com/
12. Guan W, Ni Z, Hu Y, Liang W, Ou C, He J, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med. 2020;382(18):1708–20.
13. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507–13.
14. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497–506.
15. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA - J Am Med Assoc. 2020;323(11):1061–9.
16. Wu C, Chen X, Cai Y, Xia J, Zhou X, Xu S, et al. Risk Factors Associated with Acute Respiratory Distress Syndrome and Death in Patients with Coronavirus Disease 2019 Pneumonia in Wuhan, China. JAMA Intern Med. 2020;180(7):934–43.
17. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med. 2020;8(5):475–81.
18. Lee J, Mervosh S, Avila Y, Harvey B, Matthews A. See How All 50 States Are Reopening (and Closing Again) - The New York Times [Internet]. [cited 2020 Aug 1]. Available from: https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html
19. Patel S, Patel P. Doubling Time and its Interpretation for COVID 19 Cases. Natl J Community Med. 2020 Mar;11(3):141–3.

Assessing the Impact of Imputation on the Interpretations of Prediction Models: A Case Study on Mortality Prediction for Patients with Acute Myocardial Infarction

Seyedeh Neelufar Payrovnaziri, MS[1], Aiwen Xing, BS[1], Shaeke Salman, MS[1], Xiuwen Liu, PhD[1], Jiang Bian, PhD[2], Zhe He, PhD[1,*]
[1]Florida State University, Tallahassee, Florida, USA;
[2]University of Florida, Gainesville, Florida, USA

Abstract

Acute myocardial infarction poses significant health risks and financial burden on healthcare and families. Prediction of mortality risk among AMI patients using rich electronic health record (EHR) data can potentially save lives and healthcare costs. Nevertheless, EHR-based prediction models usually use a missing data imputation method without considering its impact on the performance and interpretability of the model, hampering its real-world applicability in the healthcare setting. This study examines the impact of different methods for imputing missing values in EHR data on both the performance and the interpretations of predictive models. Our results showed that a small standard deviation in root mean squared error across different runs of an imputation method does not necessarily imply a small standard deviation in the prediction models' performance and interpretation. We also showed that the level of missingness and the imputation method used can have a significant impact on the interpretation of the models.

Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide [1] and account for 1 in 3 deaths of adults in the United States every year.[2] CVDs cause a heavy toll on the health and economy all over the world.[3] Among various CVDs, acute myocardial infarction (AMI) is the most severe form of coronary artery disease and a fatal CVD responsible for the death of millions of people annually around the world.[4] Thus, prediction of mortality risk among AMI patients is important for early interventions or advising preventive strategies to high-risk patients, which will save lives and costs.

The wide adoption of electronic health records (EHR) systems in the United States is the result of a series of government initiatives[5] and led to a large amount of clinical data accumulated in digital forms.[6] EHR data is a rich source of patient information for predictive analysis in healthcare.[7] Predictive analysis in healthcare and clinical decision-support is not a new topic.[8] Nevertheless, in recent years, there is an increasing demand for using routinely collected real-world data (RWD) such as EHRs, administrative claims, and billing data to generate real-world evidence (RWE) that informs regulatory decisions and clinical care.[9] On the other hand, the emergence and efficient implementation[10] of state-of-the-art machine learning,[11] especially deep learning methods,[12] as well as increasingly powerful computing infrastructure make predictive analysis using EHR data more possible than ever.

Nevertheless, using EHR data for predictive analysis with machine learning and deep learning methods is still challenging. One major issue is the quality of EHR data due to incompleteness.[13] The existence of missing values in EHR data is multi-fold, including human errors such as the lack of collection (e.g., the medical expert did not perform an evaluation) and the lack of documentation (e.g., the medical expert did not document an evaluation result).[14] Thus, a significant body of literature has attempted to approach the missing value issue by imputation, rather than eliminating records with missing data entirely (i.e., as it reduces the sample size).[15] Mean or median value imputation is a common approach for imputing missing values in EHR data mainly due to its ease of implementation.[16] Researchers have also proposed multiple imputation by chained equations (MICE)[17] or its variations to deal with missing values in EHR data.[14] There are also machine-learning-based imputation methods such as MissForest[18] and K-nearest neighbors (KNN)-based imputation.[19] A few recent studies have also used deep learning methods such as generative adversarial networks (GANs)[20] and autoencoders for missing value imputation in EHR data.[19]

Depending on the reason for a value to be missing in a dataset, there are three main categories of missingness mechanism, including missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).[21] Characterizing the missingness mechanism in EHR data can be an indicator of choosing an appropriate imputation method. This impact has been explored previously[19] and is not within the scope of this study.

* Corresponding author: Zhe He. Email: zhe@fsu.edu

On the other hand, predictive analysis using complex machine learning methods (e.g., deep learning), which yields superior prediction accuracy, usually results in black-box models that are not interpretable by the end-users. Despite their promising performance, using such complex predictive models in healthcare and clinical decision-making process is quite challenging. Medical professionals need to understand the rationale behind the predictive models' predictions.[22] Thus, they prefer models that are less complex such as logistic regression for clinical decision-making.[12] Researchers in the field have taken different approaches to address the interpretability of machine learning models, for instance, feature interaction and importance, attention mechanism, data dimensionality reduction, knowledge distillation and rule extraction.[23] Nevertheless, there are still some fundamental issues that need to be addressed such as fidelity of the post-hoc interpretation methods to the reference model, evaluation of the interpretation methods, and design biases due to focusing on the intuition of researchers rather than real end-users' (medical professionals in this context) needs. For more reading on this, we refer the interested audience to a recent systematic review on the explainable AI models using EHR data.[23]

For example, in a logistic regression model for a binary outcome, the coefficients of the features (predictors) can be readily transformed into odds ratios and can be easily understood as feature importance. Nevertheless, as these coefficients are estimated from the input data when the data points were replaced with different imputation techniques, the extent to which missing data points are extrapolated has a certain impact on the interpretation of these coefficients. Further, in real-world applications of EHR and in the absence of a complete dataset (knowing the ground truth), it would not be straightforward to choose the best imputation method.[19] Although missingness has been recognized as a major data quality issue of EHR,[24] the impact of imputation methods on the interpretations of machine learning models has not been well explored and warrants more investigation. Thus, in this study, we examine (1) the performance of imputation methods on missing values in EHR data, (2) the impact of different imputation methods on the performance, and (3) the interpretations of predictive models, using all-cause mortality among AMI patients as a case study.

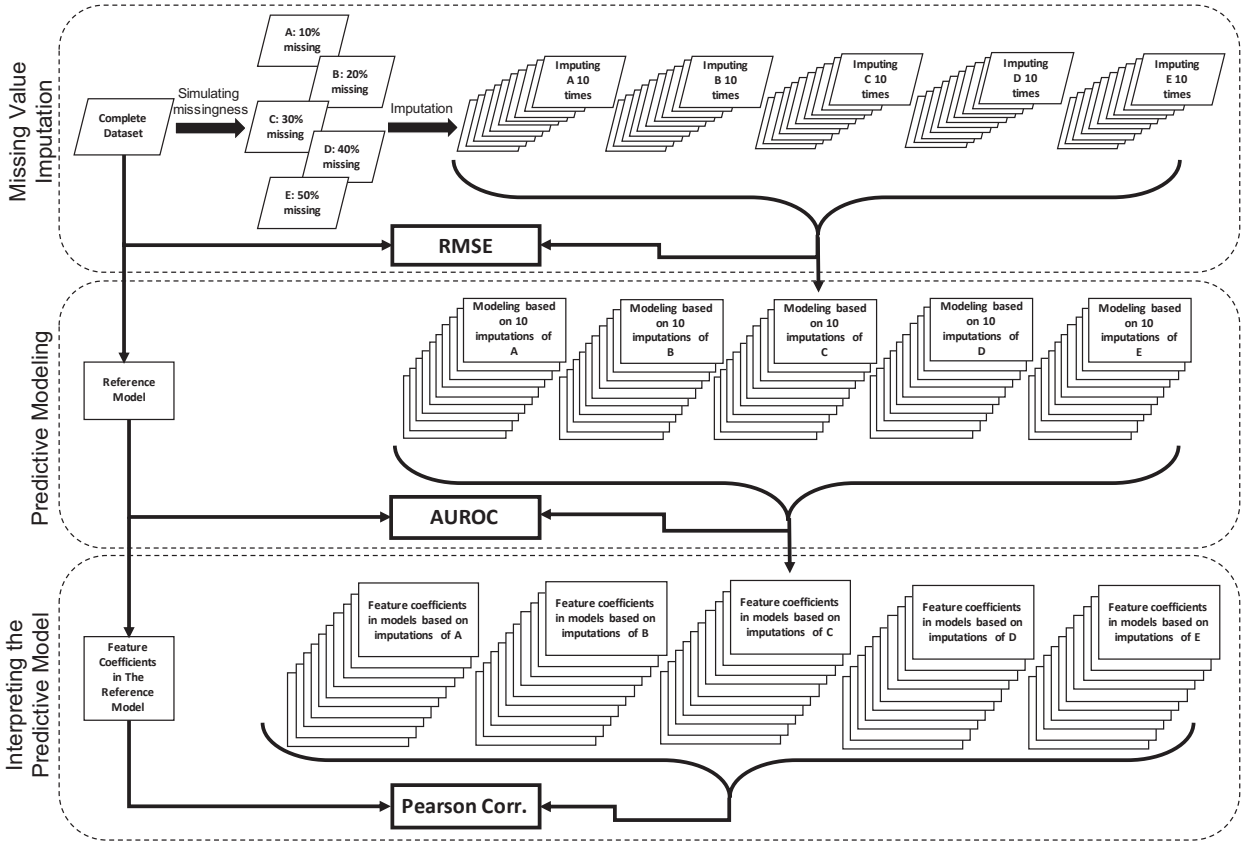


Figure 1. The workflow of this study. For simplicity, this figure only illustrates the workflow for one imputation method and one prediction model.

Methods

Figure 1 illustrates the overall workflow of this study. First, we identified patients with AMI from the Medical Information Mart for Intensive Care (MIMIC-III) dataset and created a complete dataset without missing values as the baseline. We then introduced different levels of missingness (i.e., from 10% to 50%) through simulations. Then, we applied different statistical and machine learning-based imputation methods including mean (mode for two categorical variables), MICE, MissForest, and a KNN-based method, as well as Generative Adversarial Imputation Networks (GAIN)[20] - a novel imputation method based on neural networks. Then, we compared these imputation methods' performance in terms of root mean square error (RMSE), which measures the difference between the imputed values (in the datasets with missing values that were imputed) and the actual values (in the complete dataset).[25] Further, we built shallow machine learning models that are intrinsically interpretable and preferred by medical experts, such as logistic regression, linear support vector machine (SVM), and decision tree. We compared the performance of the models that were based on the imputed datasets in terms of area under the receiver operator characteristic curve (AUROC) against the performance of the reference model (i.e., the model based on the complete dataset). For simplicity, we refer to the former as the "imputed-data models" and the latter as the "reference model" throughout the paper. Finally, we compared the feature importance derived from the imputed-data models against the feature importance derived from the corresponding reference model using Pearson correlation analysis. When the underlying data changes (i.e., from the complete dataset to one with imputed values), the resulting model characteristics might change as well as the produced feature importance. Since we observed variance in the performance of the shallow models, we employed the DeepConsensus[26] algorithm to investigate the impact of consensus mechanism among deep models on reducing performance variance. We elaborate more on this in the "Predictive Modeling" subsection.

Data Source

The Medical Information Mart for Intensive Care (MIMIC-III) database is an integration of de-identified and comprehensive EHR data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts.[27] This dataset is freely available and contains patient information spanning over more than a decade. Considering the importance of studying the all-cause mortality risk among patients with cardiovascular diseases, especially AMI, in this study, we focused on patients with AMI and post myocardial syndrome (PMS). The International Classification of Diseases, 9th revision, Clinical Modification (ICD-9-CM) codes considered for this study are 410.0 to 411.0. For each admission, we aggregated the laboratory and chart values and considered the average value of each of the 19 numerical features. We also considered two categorical features with few missing values including gender and initial emergency room diagnosis of AMI. Since our purpose for this study was to examine how the ranking of feature importance would change under different imputation methods and level of missingness, we did not consider the longitudinal dimension of features. To build a reference dataset with no missing values, we excluded features with more than 50% missing values and applied listwise deletion to the rest of the original dataset. The resulting complete dataset has 3054 observations and 21 features. The binary outcome was all-cause mortality within one year of admission. Features description and summary statistics are reported in Table 1.

Table 1. Summary statistics of the variables in the reference complete dataset with 3054 instances.

Variable Name	Description	Variable Type	Min	Max	Median	Mean	Standard Deviation
Diastolic BP	Diastolic blood pressure	Numerical	18.31	134.69	51.2	52.2	11.28
Systolic BP	Systolic blood pressure	Numerical	37.97	484.12	104.96	105.44	22.74
Heart Rate	Heart rate	Numerical	42	139.48	83.81	84.13	11.97
Resp Rate	Respiratory rate	Numerical	9	42.7	19.33	19.62	3.28
Bicarbonate	Bicarbonate	Numerical	8	41.88	24.92	24.61	3.65
Calcium	Calcium	Numerical	5.6	13.95	8.43	8.44	0.59
Chloride	Chloride	Numerical	80.42	125.61	104	104.01	4.55
Potassium	Potassium	Numerical	1.87	6.9	4.14	4.18	0.36
Sodium	Sodium	Numerical	118.18	158.5	138.72	138.67	3.47
Glucose	Glucose	Numerical	65.67	543	131.76	141.25	39.93
Hematocrit	Hematocrit	Numerical	21.11	50.61	30.97	31.53	3.6
Hemoglobin	Hemoglobin	Numerical	6.4	16.27	10.43	10.63	1.34
WBC	White blood cell count	Numerical	0.45	107.68	10.93	11.75	5.19
ALT	Alanine aminotransferase	Numerical	2	5509	30.33	89.71	270.59

AST	Aspartate aminotransferase	Numerical	2	13511.7	46	142.36	486.22
ALP	Alkaline phosphatase	Numerical	19	1147.92	80	102.61	83.6
Albumin	Albumin	Numerical	1.2	5	3.2	3.2	0.6
Bilirubin	Bilirubin	Numerical	0.1	31.14	0.6	0.97	1.77
Admit Age	Age at admission	Numerical	21.22	97.52	72.55	70.89	12.96
InitialERDiagnosisMI (0 = No, 1 = Yes)	Initial emergency room diagnosis was AMI or rule out AMI (#0s = 2050 , #1s = 1004)	Categorical	Not applicable				
Gender (0 = Female, 1= Male)	Gender (#0s = 1202, #1s = 1852)	Categorical					

Missingness Mechanisms

In this study, our focus is not on studying missingness mechanisms, rather it is on exploring the impact of imputation methods on models' performance, and more importantly, the derived interpretations. We acknowledge that MCAR is not the only possible missingness mechanism in RWD such as EHRs. To give an example, medical professionals might less likely order fasting glucose tests for healthier patients in comparison to those with risk factors of diabetes. Thus, for healthier patients, in this case, there might be more missing glucose values. However, covering MAR and NMAR missingness mechanisms is out of the scope of this paper. Considering all possible combinations of missingness mechanisms with other criteria experimented in this paper would exponentially expand the scope of the paper and increase the complexity of reporting. We simulated random missingness on the complete dataset by removing random 10%, 20%, 30%, 40%, and 50% of values. Any value in the data was as likely to be missing as any other value. Thus, the missingness mechanism in this study was MCAR.

Imputation Methods

In this study, we evaluated five different imputation methods, including (1) mean value imputation, (2) MICE, (3) K-nearest neighbors (KNN)-based, (4) MissForest, and (5) GAIN. For implementation purposes, we used available packages in R to implement methods (1) to (4). The implementation code of GAIN in Python is made available by its authors on GitHub (https://github.com/jsyoon0823/GAIN). The performance of different imputation methods was compared using RMSE. The details of these methods are described as follows.

Mean value imputation: The easiest to implement and most conventional approach to impute missing values in EHR data is mean value imputation. However, this simplicity might result in ignoring the underlying statistical information in data and introduce unintentional biases in the subsequent analyses.[14]

MICE: MICE is one of the most popular methods for imputing missing values in EHR data. The main reason resides in its ability to impute different types of variables that might be present in the EHR data. Using MICE, each variable with missing observations is regressed on all the remaining variables in the dataset. The missing values are replaced with the predicted value, and this imputation process is repeated sequentially until all missing values are imputed.

KNN-based: KNN is a machine learning method that can be used for imputing missing values in EHR data.[19] In this approach, missing values are replaced with the mean value of k most similar complete observations. A distance function (e.g., Euclidean) is used to measure this similarity.

MissForest: MissForest is a promising imputation method for missing values in EHR data.[18] In this method, first, mean imputation (or any other imputation method) is performed as an initial guess for the missing values. Then, variables in the dataset are sorted based on the number of missing values they have with the one with the fewest missing values ordered first. Further, for each variable x, a random forest [28] model is fitted on all other variables' observed values and the outcome variable being the observed values of variable x. Then, the trained model is used to predict the missing values of x. This process is repeated until a stopping criterion is met.

GAIN: Recently, GAIN,[20] a neural network-based imputation method was introduced for missing value imputation. This imputation method is based on the generative adversarial networks (GAN) framework. In the framework, corresponding to a minimax two-player game, two models are trained simultaneously, a generative model and a discriminative model. The generative model captures the data distribution while the discriminative model estimates the probability of a sample being from the training data or from the generative model. The objective of the generative model is to make the discriminative model make more mistakes. In GAN, the generative and discriminative models are defined based on multilayer perceptron (feedforward neural networks). GAIN is an imputing GAN framework in

which the goal of the generative model is to accurately impute the missing values in data, while the goal of the discriminative model is to predict the probability of a value being from the original dataset or from the generative model (observed or imputed component). The objective of the discriminative model in GAIN is to minimize the error loss (on guessing if the elements in the generative model's output are produced by the generative model or from the original data) while the generative model's goal is to maximize the discriminative model's mistakes. The authors of GAIN have reported superior imputation performance of GAIN in comparison to autoencoders and other statistical and conventional machine learning-based imputation methods. For more information on GAIN, we refer the interested audience to the original paper.[20]

Predictive Modeling

To compare the prediction performance and feature importance ranking of different prediction models with different levels of missingness, we performed predictive analysis using three popular shallow machine learning methods in predictive modeling with EHR data,[29] namely logistic regression, SVM, and decision tree. To keep consistency, the same configurations (default) were used across all models based on the imputed datasets.[30] Hypothetically, an effective imputation method should approximate values that are very close to the original values of the missing data. Thus, to understand how the performance and interpretations change under different imputation methods, we kept the configurations consistent from the reference model to the imputed-data models with a gradually increasing number of missing values. Further, we captured feature coefficients (importance) in each imputed-data model to compare to the same in the reference model of its own kind. For comparison, we used Pearson correlation coefficients. A higher correlation means closer results from the imputed-data model (in terms of feature importance) to the reference model based on the complete dataset. Also, we built a deep learning model (i.e., DeepConsensus) to investigate if it can reduce the variance of the models' classification performance. The binary prediction task was patient all-cause mortality within one year after admission. The dataset was divided to separate training and testing sets at the ratio of 0.9 to 0.1 respectively. The dataset is imbalanced with 65 (negative class) to 35 (positive class) ratio. For implementation purposes, we used Python programming language with Tensorflow, NumPy, Pandas, and Sklearn packages. We give a brief description of DeepConsensus in the following.

DeepConsensus: The main idea behind DeepConsensus is that since different deep neural networks tend to classify training samples accurately, they generate similar linear regions. Thus, these models should behave similarly in classifying training samples. Such behavior enables multiple models to agree with each other on classifying valid inputs and filtering out adversarial examples, while individual models are sensitive to those examples. Using consensus among different models helps to capture the underlying structure of data. It is shown that consensus helps to differentiate extrinsically classified samples (i.e., classified under extrinsic factors such as randomness of weight initialization) from consistently classified samples (i.e., samples that are classified in the same class with high probability by multiple models). Thus, such a consensus mechanism among multiple models can reduce the variance caused by extrinsic factors. The effectiveness of this method is demonstrated in the reference paper.[26] Note that in this study, DeepConsensus was only employed to investigate the impact of consensus mechanism on the variance of the classification performance. In terms of model interpretations, to keep consistency across all experiments, we only focused on less complex machine learning models that are intrinsically interpretable (DeepConsensus utilizes a post-hoc interpretation approach).[26] Five individual deep models were trained using the same 90% of the records and evaluated on the remaining 10%. All models consisted of 4 hidden dense layers. The implementation details of these models are provided in Table 2.

Table 2. Implementation detail of five individual models in DeepConsensus.

Specification	Model 1	Model 2	Model 3	Model 4	Model 5
#Neurons in each layer	105	130	130	140	105
Activation function	ReLU	Tanh	ReLU	SeLU	ReLU
Optimization	Adagrad	Adamax	RMSprop	Adam	Adagrad
Bias	Random_Uniform	Zeros	Constant	Zeros	Random_Uniform
Weights	Random_Normal	Glorot_Uniform	Random_Normal	Random_Normal	Random_Normal

Results

Imputation Performance

We compared five different imputation methods including MICE, MissForest, KNN-based, Mean (mode for categorical variables), and GAIN, on 10% to 50% missing datasets. We ran each experiment 10 times and computed the average RMSE along with its standard deviation. The results are reported in Table 3.

Averaging the RMSE of different imputation methods across all different levels of missing values, GAIN showed the best performance. MissForest was the second-best performing imputation method following GAIN showing only 0.012 difference in RMSE on average. Mean came next, showing quite monotonic behavior across all datasets with different amount of missing values. KNN-based and MICE showed similar performance reporting the highest RSME (worst performance).

Table 3. Average RMSE from different imputation methods on 10% to 50% missing data (10 runs).

Missingness level / **Imputation Method**	**10%**	**20%**	**30%**	**40%**	**50%**
MICE	0.2254±0.0025	0.2249±0.0020	0.2292±0.0013	0.2343±0.0014	0.2325±0.0014
MissForest	0.1935±0.0017	0.1950±0.0013	0.1997±0.0018	0.2051±0.0018	0.2062±0.0011
KNN-based	0.21447	0.2219	0.2241	0.2267	0.2228
Mean\Mode	0.2038	0.2046	0.2052	0.2066	**0.2059**
GAIN	**0.1757±0.0049**	**0.1763±0.0064**	**0.1838±0.0039**	**0.1963±0.0114**	0.2088±0.0100

Prediction Performance

Next, to narrow down the required experiments, we focused on the datasets imputed with the best performing imputation methods in terms of RMSE: GAIN and MissForest. First, we built the reference models that served as the benchmark for our comparisons. Then, we built models based on datasets with varying percentage of missingness that were imputed using GAIN and MissForest (through 300 experiments = 5 levels of missingness * 2 imputation methods* 3 ML methods * 10 runs each). The performance of these models based on GAIN imputed datasets and MissForest imputed datasets in terms of AUROC is depicted in Figure 2. AUROC is a classification performance measure that illustrates how models perform in terms of discriminating between the classes. The performance of SVM and logistic regression models based on MissForest imputed datasets (Figure 2) showed a lower variance in comparison to the same models based on GAIN imputed datasets. The variance in decision tree models based on the datasets imputed by both GAIN and MissForest is relatively high, making those models' performance unstable even under low levels of missingness. By increasing the number of missing values in the datasets from 10% to 50%, models based on the datasets imputed by MissForest showed a more stable behavior in comparison to GAIN across all models. A closer look at SVM performance (based on AUROC) from 10% to 50% missing values imputed by MissForest showed an increase of standard deviation from 0.003 to 0.010 with an average of 0.006. This measure was 0.002 to 0.025 for GAIN with an average of 0.014. Also, a big jump in standard deviation was not observed in the case of MissForest until 40% of missingness. This jump occurred at 20% of missingness in the case of GAIN. The standard deviation of AUROC of multiple runs for logistic regression models based on GAIN showed an increasing trend from 0.005 to 0.019 with an average of 0.010. However, in the case of MissForest, the trend is increasing from 10% to 30% and then decreasing from 30% to 50% with an average of 0.006. We hypothesized that the variance among models on the same missing level of missing data that is imputed with the same method 10 times can be reduced by using the consensus mechanism among multiple deep models. Applying DeepConsensus on the complete dataset (baseline) showed a significant increase in performance as expected: from 0.721 AUROC in logistic regression and 0.713 in SVM to 0.809 in DeepConsensus. Further, we narrowed down our experiments to 10 datasets that were the result of imputing 10% missingness (lowest missing rate) 10 times with MissForest (imputation method with less variance on machine learning models). The average AUROC across 10 experiments showed an increase of performance to 0.790 (from 0.675 in SVM and 0.683 in logistic regression). However, the variance in performance still persists at 0.0344.

Feature Importance Ranking Comparison

The Pearson correlation of feature importance resulted from each of the imputed data models with that of the corresponding reference models is reported in Table 4. The goal of this analysis was to investigate how the reported feature importance would change under different levels of missingness and different imputation techniques. In other

words, how similar the feature importance in each imputed-data model is to the corresponding reference model. These results showed a generally lower performance for decision tree models across all datasets. Focusing on SVM and logistic regression with higher performance, averaging coefficients as a result of 10 runs for each level of missingness showed a correlation coefficient of more than 0.99 with statistically significant results on the imputed datasets of 10% missing value with GAIN and MissForest. However, this trend on models based on an increasing level of missingness on average showed a decreasing correlation with the reference model across all machine learning methods and imputation methods.

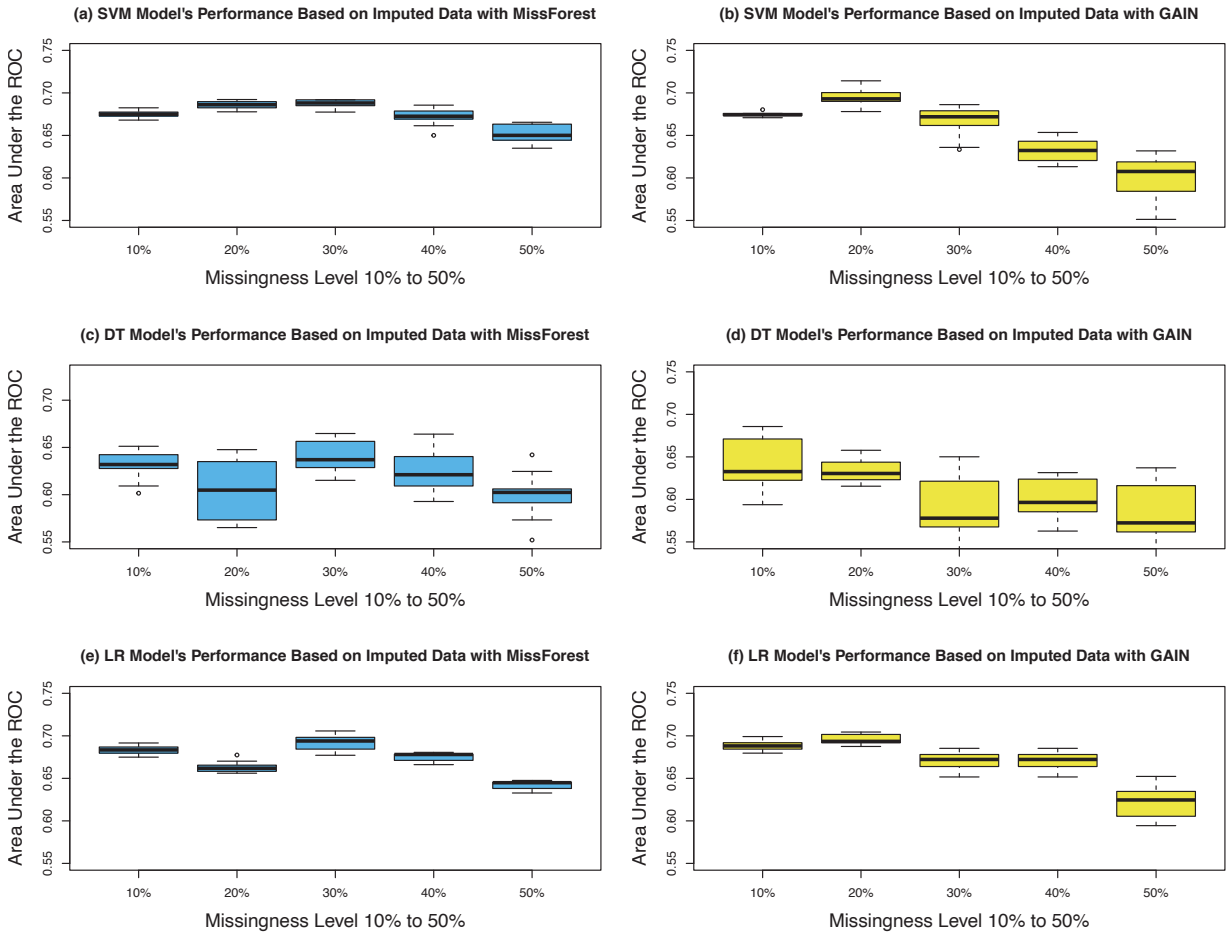


Figure 2. Comparing the performance of models based on MissForest (in blue) and GAIN (in yellow): (a) support vector machine (SVM) based on the imputed datasets with MissForest, (b) SVM based on the imputed datasets with GAIN, (c) decision tree (DT) based on the imputed datasets with MissForest, (d) DT based on the imputed datasets with GAIN, (e) logistic regression (LR) based on the imputed datasets with MissForest, (f) LR based on the imputed datasets with GAIN, under gradually increasing missingness level. The performance is reported on the area under the receiver operating characteristic curve (AUROC).

Table 4. Pearson correlation coefficients and p-values of feature importance comparison between the imputed-data models and the reference models.

Machine Learning method	Imputation method	Missingness %	Pearson correlation coefficient	p-value	Imputation method	Missingness %	Pearson correlation coefficient	p-value
Decision Tree	GAIN	10%	0.966	1.24E-12	MissForest	10%	0.944	1.23E-10
		20%	0.959	7.09E-12		20%	0.914	6.56E-09
		30%	0.917	4.84E-09		30%	0.906	1.49E-08
		40%	0.842	1.64E-06		40%	0.874	2.24E-07
		50%	0.903	2.05E-08		50%	0.770	4.34E-05

SVM		10%	0.992	8.57E-19		10%	0.994	7.71E-20
		20%	0.986	1.80E-16		20%	0.983	1.12E-15
		30%	0.949	5.54E-11		30%	0.985	5.31E-16
		40%	0.834	2.51E-06		40%	0.975	5.09E-14
		50%	0.793	1.73E-05		50%	0.911	8.83E-09
Logistic Regression		10%	0.995	6.80E-21		10%	0.996	2.56E-22
		20%	0.988	5.14E-17		20%	0.987	1.28E-16
		30%	0.962	3.62E-12		30%	0.985	4.24E-16
		40%	0.895	4.00E-08		40%	0.978	1.53E-14
		50%	0.854	8.39E-07		50%	0.958	7.44E-12

Discussion

Comparing the imputation methods' RMSE reported in Table 3 implies that (1) GAIN performs better than MissForest, and (2) the standard deviation between different runs of the same method on the same dataset with missing values is small. However, our experiments confirmed that choosing the best imputation method may not always be a straightforward process. Although GAIN surpassed all imputation methods in terms of RMSE on all datasets, MissForest imputation yielded more stable results (smaller standard deviation on average) in the presence of a gradually increasing number of missing values. Also, comparing the performance of models based on datasets with different percentage of missingness reveals the fact that higher performance does not necessarily indicate more similar interpretations to the reference model. We observed that on average a relatively small standard deviation of RMSE across all levels of missingness yielded a bigger standard deviation in models' performance and a lower correlation of feature importance between the reference models and the imputed-data models. Also, the dilemma of bias/variance is well understood regarding neural networks that are hyperparameterized.[31] Training neural networks requires a larger number of training samples to achieve acceptable performance and less variance. Thus, although using the consensus of deep models did not resolve the issue of variance in this study, we hypothesize that using a bigger dataset (with more samples and more features) could potentially yield a more stable consensus of deep learning models and result in less variance.

These observations might not be generalizable to other datasets or imputation methods. There is no universally optimal approach for missing data imputation or predictive modeling using EHR data. However, these experiments showed that the way we approach missing values in EHR data impacts not only the model performance but also the interpretations of the models' predictions. In the real-world predictive analysis of EHR data, it is usually not possible to obtain a dataset with no missing values. However, in cases where the interpretations of predictive models matter, in order to choose the best imputation method, just relying on RMSE or model performance measures may not be sufficient. In these cases, we suggest running extensive experiments on a smaller complete-case version of the dataset first, evaluate the impact of different imputation methods on the interpretations in comparison to the complete-case, and then apply the best performing method on the original dataset with missing values. In cases where it is not possible to have the complete-case dataset, researchers should be aware of this potential impact, use different imputation methods for predictive modeling, and discuss the resulting interpretations with medical experts or compare to the medical knowledge when choosing the imputation method that yields the most reasonable interpretations. Also, more in-depth analyses of data with methods such as principal component analysis (PCA) can be used to investigate the redundancy in datasets and determine the maximal allowed missing value rate. In our case, for instance, A further PCA on the complete dataset showed that a linear model could capture between 90.1% and 92.8% of the statistical information using 14 to 15 features out of a total of 21 (~66% to 71% of the complete dataset). Thus, even if the other 30% of the data was missing still more than 90% of the statistical information would be preserved.

Limitations and Future Opportunities

A potential limitation of this study was the relatively small and imbalanced dataset (65:35). Although the findings in this study are robust, future studies could be done on datasets with more balance and more samples to investigate how these results would change. Also, MIMIC-III is an ICU EHR. In comparison to other EHR types such as inpatient or ambulatory, MIMIC-III data might be less noisy and more standardized. Further, we acknowledge that there are different missingness mechanisms, missing value imputation methods, predictive modeling approaches, and interpretability enhancement techniques that are not covered in this study. However, implementing all the possible combinations of these criteria is expensive. The main intention behind this case study was to demonstrate the potential impact of missing value imputation on derived interpretations of predictive models. Thus, we encourage future in-depth theoretical and applied studies on this topic. We believe the interpretability issue in predictive modeling is of

the same importance, if not more, as performance when it comes to medical applications. Also, the interpretability enhancement of predictive models based on longitudinal EHR data is inherently challenging and has not been extensively explored. However, we believe investigating the potential impact of missing value imputation in longitudinal EHR data and its interpretations is an important future direction and deserves more investigation.

Conclusions

In this study, we simulated 5 levels of missingness (10% to 50%) on a complete EHR dataset of 21 features for 3054 patients with AMI from MIMIC-III database. We examined different statistical and machine learning-based imputation methods such as mean, MICE, MissForest, and KNN-based, as well as GAIN-a novel imputation method based on GAN. Our experiments showed that GAIN and MissForest yielded the best performance in terms of RMSE and small standard deviations across all levels of missingness. However, further predictive modeling based on each of these datasets revealed the fact that the variance in their performance (in terms of AUROC) gradually grows with more missingness. Also, Pearson correlation analysis showed that the similarity of feature importance of models based on the imputed datasets to the feature importance of baseline models gradually decreases, a trend that could not initially be inferred by just looking at the performance of imputation and predictive modeling in terms of RMSE and AUROC respectively.

Acknowledgments

This study was supported in part by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R21AG061431; and the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences under award number UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. al-Aiad A, Duwairi R, Fraihat M. Survey: Deep Learning Concepts and Techniques for Electronic Health Record. In: 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA). 2018. p. 1–5.
2. Curry SJ, Krist AH, Owens DK, Barry MJ, Caughey AB, Davidson KW, et al. Risk Assessment for Cardiovascular Disease With Nontraditional Risk Factors: US Preventive Services Task Force Recommendation Statement. JAMA. 2018 Jul 17;320(3):272–80.
3. Ford ES, Capewell S. Coronary Heart Disease Mortality Among Young Adults in the U.S. From 1980 Through 2002: Concealed Leveling of Mortality Rates. Journal of the American College of Cardiology. 2007 Nov 27;50(22):2128–32.
4. Reed GW, Rossi JE, Cannon CP. Acute myocardial infarction. The Lancet. 2017 Jan 14;389(10065):197–210.
5. Blumenthal D. Implementation of the Federal Health Information Technology Initiative. New England Journal of Medicine. 2011 Dec 22;365(25):2426–31.
6. ANSI I. ISO/DTR 20514: Health informatics–electronic health record—definition, scope and context. ISO; 2005.
7. Milenkovic MJ, Vukmirovic A, Milenkovic D. Big data analytics in the health sector: challenges and potentials. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies. 2019;24(1):23–33.
8. Parikh RB, Kakad M, Bates DW. Integrating Predictive Analytics Into High-Value Care: The Dawn of Precision Delivery. JAMA. 2016 Feb 16;315(7):651–2.
9. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence — What Is It and What Can It Tell Us? New England Journal of Medicine. 2016 Dec 8;375(23):2293–7.
10. Erickson BJ, Korfiatis P, Akkus Z, Kline T, Philbrick K. Toolkits and Libraries for Deep Learning. J Digit Imaging. 2017 Aug 1;30(4):400–5.
11. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLOS ONE. 2018 Aug 31;13(8):e0202344.
12. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics. 2018 Sep;22(5):1589–604.
13. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit on Translat Bioinforma. 2010 Mar 1;2010:1–5.

14. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC). 2013;1(3):1035.
15. Little RJ, Rubin DB. Statistical analysis with missing data. Vol. 793. John Wiley & Sons; 2019.
16. Salgado CM, Azevedo C, Proença H, Vieira SM. Missing Data. In: MIT Critical Data, editor. Secondary Analysis of Electronic Health Records [Internet]. Cham: Springer International Publishing; 2016 [cited 2020 Feb 20]. p. 143–62.
17. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2010;1–68.
18. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;28(1):112–8.
19. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. JMIR medical informatics. 2018;6(1):e11.
20. Yoon J, Jordon J, Schaar M. GAIN: Missing Data Imputation using Generative Adversarial Nets. In: International Conference on Machine Learning [Internet]. PMLR; 2018 [cited 2020 Dec 9]. p. 5689–98. Available from: http://proceedings.mlr.press/v80/yoon18a.html
21. Scheffer J. Dealing with missing data. 2002
22. Ahmad MA, Eckert C, Teredesai A. Interpretable Machine Learning in Healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics [Internet]. Washington, DC, USA: Association for Computing Machinery; 2018 [cited 2020 Feb 21]. p. 559–560. (BCB '18).
23. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc. 2020 Jul 1;27(7):1173–85.
24. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144–51.
25. Duy Le T, Beuran R, Tan Y. Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare. In: 2018 10th International Conference on Knowledge and Systems Engineering (KSE). 2018. p. 247–51.
26. Salman S, Payrovnaziri SN, Liu X, Rengifo-Moreno P, He Z. DeepConsensus: Consensus-based Interpretable Deep Neural Networks with Application to Mortality Prediction. In: 2020 International Joint Conference on Neural Networks (IJCNN). 2020. p. 1–8.
27. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016 May 24;3(1):1–9.
28. Breiman L. Random Forests. Machine Learning. 2001 Oct 1;45(1):5–32.
29. Taslimitehrani V, Dong G, Pereira NL, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. Journal of Biomedical Informatics. 2016 Apr 1;60:260–9.
30. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recogn. 2008 Dec 1;41(12):3692–3705.
31. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural computation. 1992;4(1):1–58.

Deep EHR Spotlight: a Framework and Mechanism to Highlight Events in Electronic Health Records for Explainable Predictions

Thanh Nguyen-Duc, MSc[1,2]; Natasha Mulligan, MSc[1]; Gurdeep S. Mannu, MBBS, MRCSEd, DPhil[3]; Joao H. Bettencourt-Silva, MSc, PhD[1]
[1]IBM Research Europe, Dublin, Ireland; [2]Monash University, Melbourne, Australia; [3]Nuffield Department of Population Health, University of Oxford, Oxford, UK

Abstract

The wide adoption of Electronic Health Records (EHR) has resulted in large amounts of clinical data becoming available, which promises to support service delivery and advance clinical and informatics research. Deep learning techniques have demonstrated performance in predictive analytic tasks using EHRs yet they typically lack model result transparency or explainability functionalities and require cumbersome pre-processing tasks. Moreover, EHRs contain heterogeneous and multi-modal data points such as text, numbers and time series which further hinder visualisation and interpretability. This paper proposes a deep learning framework to: 1) encode patient pathways from EHRs into images, 2) highlight important events within pathway images, and 3) enable more complex predictions with additional intelligibility. The proposed method relies on a deep attention mechanism for visualisation of the predictions and allows predicting multiple sequential outcomes.

1 Introduction

Electronic health records (EHR) are essential in supporting healthcare practitioners track their patients and deliver services. Details from patients' encounters are routinely collected in EHRs and, despite advances in machine learning and statistical modeling, these data are heterogeneous and difficult to reuse. The expectation is that routinely collected data together with deep learning techniques may help drive personalised medicine and improve the quality of healthcare service delivery. Typical analytic tasks include disease classification or prediction of clinical events and a number of different deep learning architectures has been identified[1].

Most analytic architectures rely on EHR data which has undergone substantial pre-processing steps to transform or model these data before it is later ingested for analysis.

EHR data has been previously modeled as pathways[2] describing patient trajectories through time, and studies have also combined multi-modal information (e.g. static demographic variables with patient vitals and text notes) for prediction tasks with some degree of explainability[3]. Other studies have represented EHR information as a matrix of ICD9 codes versus number of visits, and used convolutional neural networks (CNNs) for prediction[4].

The way in which data is modeled and represented can help support both the prediction of clinical events and the interpretability of the results, the latter being particularly important in healthcare. Bringing models into real-world use requires understanding the mechanisms by which models operate and this level of transparency is currently challenging to achieve[1]. Attention mechanisms are a common way to provide visual explanations for natural images in neural networks by highlighting the most relevant events in terms of contributions to model prediction or classification. Attention mechanisms can be divided into *soft* attention and *hard* attention and their associated weights represent the degree in which the model is paying attention to certain regions of an image. While *soft* attention uses differentiable functions to produce attention weights, *hard* attention involves non-differentiable functions to generates binary weights. For example, Xu et al. exploited *soft* attention for natural image captioning [5] and several improvements have been made to *hard* attention for the model to be differentiable (e.g. REINFORCE[6]).

In this paper we propose to represent EHR data as an image-like 2D matrix in order to predict a sequence of clinical events while providing some level of intepretation of the results. We extract features from EHRs using CNN techniques and predict sequences of clinical events using a Recurrent Neural Network (RNN) and attention techniques. RNNs are a well-known method for predicting sequential data. However, they do not perform well in practice, especially because of the vanishing gradient and long-term dependency problems[7]. Long short-term memory (LSTM)[8] is a special kind of recurrent neural network that aims to overcome these drawbacks. LSTM consists of different gates (such as input,

output and forget gate) that allow it to learn to preserve important information and discard other information in order to effectively predict long sequences. However, vanilla LSTM does not consider the previously predicted output from a previous state in order to improve the prediction of current state. A teacher enforcing technique may be used to improve performance by keeping constrains between outputs (i.e. passing the previous prediction to the current state). LSTMs using teacher enforcing have been successfully applied to images and used in natural language processing such as image captioning[5], as well as in sequence to sequence problems[9].

Previous work encoding patient EHR information using autoencoders, RNNs and attention have been applied to improve model accuracy but have not relied on a combination of 2D representation while exposing attention as a mechanism to visualize important areas within a patient's EHR[4, 10–12].

This paper proposes *Deep EHR Spotlight*, a framework for predicting and highlighting important clinical events from pathways based on electronic health records. In particular, this paper introduces:

- a 2D pathway representation which can be used with two dimensional CNN techniques to improve visual interpretation
- a novel deep spotlight model which can highlight important events to help interpret the model's predictions, which may include sequential events, using attention, LSTM with teacher enforcing.

Experiments were conducted on a real world dataset MIMIC-III[13] and an evaluation was carried out based on performance metrics and a domain expert review.

2 Methods

This section describes the proposed framework in detail over two parts as illustrated in Fig. 1: a) EHR Data Transformation and b) the Deep Spotlight Model. The first part is an EHR data transformation module which transforms the heterogeneous raw data from an EHR dataset into image-like representations named pathway images x with associated height and width denoted by $h \times w$, respectively. We also propose a Deep Spotlight Model, described in detail in section 2.2, which uses an attention mechanism that takes as input a pathway, x, to predict a sequence of targets (e.g., a sequence of conditions or diseases) $\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_L\}, \tilde{y}_i \in \mathbb{R}^K$, where K is the number of possible events to classify at $\tilde{y}_i$ and L is the maximum length of the sequence). The attention mask produced can highlight ('spotlight') the important events that significantly contribute to the model's decision using an attention mechanism described in section 2.2.2. Each prediction $\tilde{y}_i$ corresponds to specific highlighted events in the pathway, described in section 2.1.

The Pathway Composition module takes the heterogeneous EHR data to produce pathway images consisting of transformed EHR data in a 2-D image-like representation (defined in Section 2.1). The pathway image x inputs to the Deep Spotlight model to predict a sequence of targets $\tilde{y}$ that are pushed closer to the ground truth $y = \{y_1, y_2, ..., y_L\}, y_i \in \mathbb{R}^K$ (i.e. the labelled data used for training) by minimizing cross-entropy loss and its attention masks corresponding to each prediction $\tilde{y}_i$. The model and approach are described in detail in the next sections.

2.1 Pathway Data Representation

Let $\mathbb{D}$ represent a data model consisting of medical codes, where the i-th entry has code $m_i \in \mathbb{M}$ where $1 \leq i \leq N$ in a total of N possible codes. Codes may be extracted and mapped directly to terminologies such as ICD or LOINC and each code has an associated dimension, $h_1, ..., h_6$, of six pre-defined dimensions that group events of similar nature together (e.g. procedures, observations and medications) as shown in Figure 2.

A pathway describes a patient's hospital admission as a series of events $E = (r, t, m, h)$ where:

- r is the patient identifier.
- t is the sequence in which events occurred, which can be calculated using the time in days since the day of primary diagnosis for an admission recorded for patient r. Events without an associated time have $t = 0$.

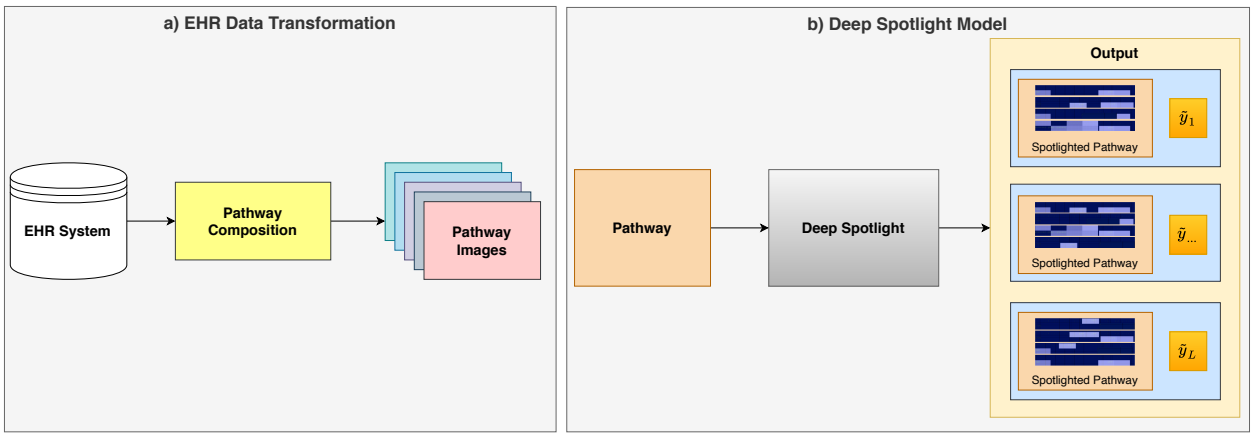


Figure 1: Overview of the framework and its two parts: a) EHR Data Transformation used to compose pathways and b) Deep Spotlight Model used to predicting a sequence of events y and highlight areas that explain predictions $\tilde{y}_i$.

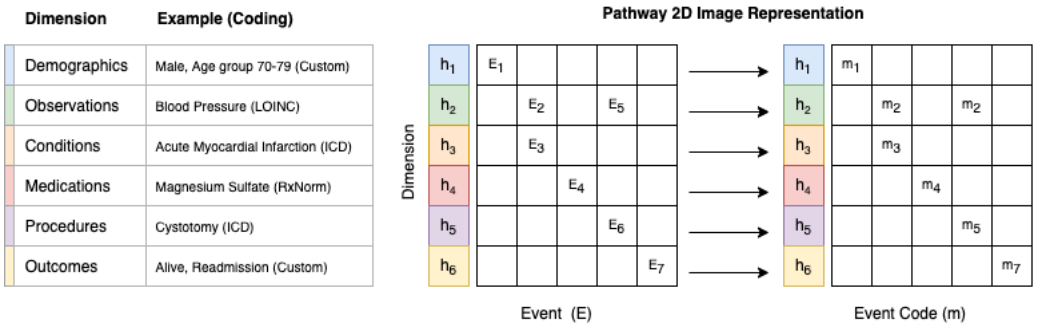


Figure 2: 2D-image representation for a given pathway. The y-axis represents the pre-defined dimensions h and the x-axis the pathway events from which codes and values are used.

- $m \in \mathbb{M}$ is an event code, such as a specific ICD9, LOINC code, or other, as described in Figure 2.
- h is the dimension associated with medical code m as described in Figure 2.

Thus, a pathway for a patient r is defined as an ordered set of events $\mathbb{P} = \{E_1, E_2, ..., E_M\}$ where:

1. E_i is of the form (r, t_i, m_i, h_i) for $1 \leq i \leq M$,
2. $t_i \leq t_i + 1$ for $1 \leq i \leq M$,
3. M is total number of events in a pathway.

Subsequently, a pathway x can be described as a 2-D image by arranging dimensions h across the y-axis and events E across the x-axis as illustrated in Figure 2. For each event in the pathway image x, the code m_i may be concatenated, resulting in a single value equivalent to each pixel in an image. This representation enables a spatial correlation for feature extraction using two dimensional CNN techniques (e.g., medications and procedures in Fig. 2) and may provide better visualization in practice. The above pathway model is inspired by previous work[2] and may be remodelled to, for example, include values associated with each event code m.

2.2 Deep Spotlight Model

Deep Spotlight model is designed to predict a sequence of output events and to highlight input events which contribute to the prediction. We denote the three main modules: feature extractor f, attention module s and long short-term

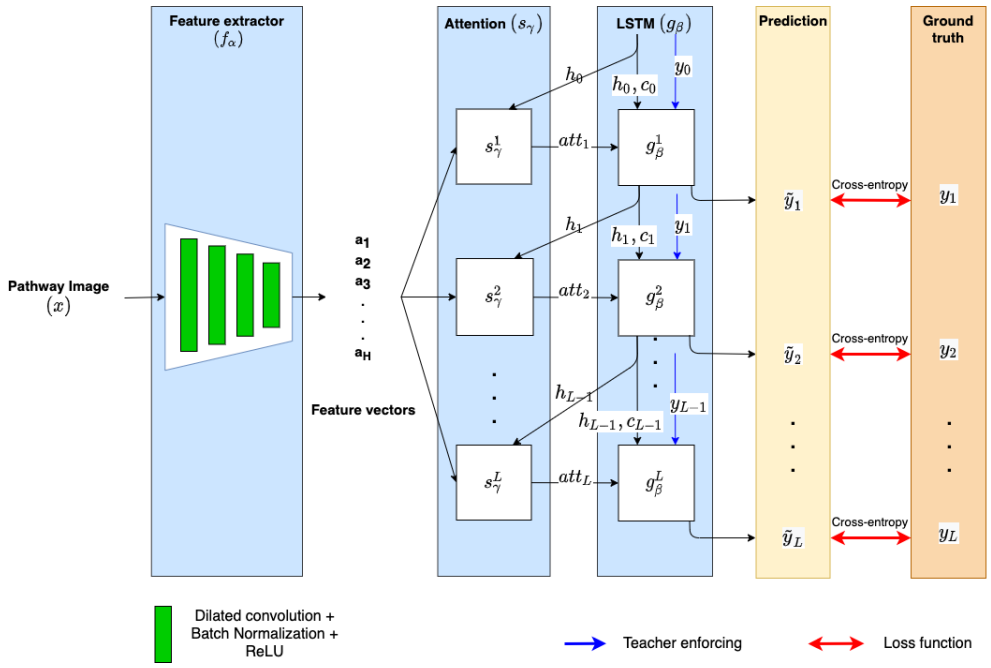


Figure 3: Training process of proposed deep spotlight model consisting of encoder f, attention module s and LSTM g parameterized by α, γ and β respectively. The prediction $\tilde{y}$ is pushed close to ground truth y by minimizing cross-entropy loss and current prediction $\tilde{y}_i$ is also enforced during training process by label y_{i-1} called teacher enforcing[14].

Input : x a single pathway image
Output : $\tilde{y}$ probabilities of predicted conditions
p attention masks of predicted conditions

$a \leftarrow f_\alpha(x)$ // a:feature vectors extracted by dilated convolution nets
$h_0 \leftarrow 0$ // h_0:initial hidden state of LSTM
$c_0 \leftarrow 0$ // c_0:initial cell state of LSTM
$\tilde{y}_0 \leftarrow 0$ // $\tilde{y}_0$:initial output at step 0 used as teacher forcing
$i \leftarrow 1$ // i:step
repeat
  $p_i, att_i \leftarrow s^i_\gamma(a, h_{i-1})$ // att_i:attention features for predicting condition i
  $\tilde{y}_i, h_i, c_i \leftarrow g^i_\beta(att_i, \tilde{y}_{i-1}, h_{i-1}, c_{i-1})$ // $\tilde{y}_i$:predicted condition i
  $\tilde{y} \leftarrow y.append(\tilde{y}_i)$ // save predicted condition i
  $p \leftarrow p.append(p_i)$ // save attention mask of predicted condition i
  $h_{i-1} \leftarrow h_i$
  $c_{i-1} \leftarrow c_i$
  $\tilde{y}_{i-1} \leftarrow \tilde{y}_i$
  $i \leftarrow i + 1$
until $i \geq L$ *or* $terminate\_condition$;
return $\tilde{y}, p$

Algorithm 1: Predicting process of a sequence of medical events $\tilde{y}$ and attention masks p corresponding to each $\tilde{y}_i$ given an input pathway image x.

memory module g which are parameterized by α, γ and β respectively. The feature extractor uses convolutional neural networks f_α (CNN) to extract features a from input pathway image. The attention module s_γ produces attention masks p with values from [0; 1] in order to highlight important events that significantly contributed to the predictions. This

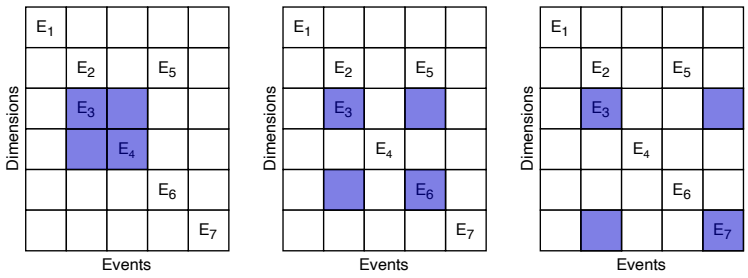


Figure 4: Receptive field expansion of a 2 $\times$2 convolution filter at different dilation factors on pathway image.

is done by increasing the weights for important feature regions of a and decreasing for unimportant ones. The LSTM module is used to generate the predicted sequence targets $\tilde{y}$ by taking the output of the attention module, as shown in Fig 3.

Note that we take the dimension (row) from the pathway image which contains a sequence of events that represents our training ground-truth y. Therefore, the input image size is $h - 1 \times w$. For example, in oder to predict conditions, the condition dimension in Fig. 2 may be taken to become ground-truth y.

2.2.1 Feature extractor using dilated convolution (f_α)

Vanilla convolutional layers[15] do not perform well in capturing the global context from images because small blocks of pixels can only be influenced by their filter size[16], especially in sparse signals; however, using larger filter sizes requires more learnable parameters which lead to computational expensive and the *data hungry* problem. Specifically, we observed that pathways extracted from MIMIC-III are highly sparse as empty elements can be 14$\times$ more frequent than encoded events. Therefore, to overcome these drawbacks our feature extractor f_α efficiently extracts features from the input pathway image using dilated convolution layers[16]. The motivation for feature extractor architecture is based on the fact that the dilated convolutions support exponentially expanding receptive fields, which is the implicit area captured on the input, while the number of parameters associated with the layer are identical. In simple intuition in Fig. 4, the dilated convolution layer is just a convolution layer applied to input with defined gaps. Moreover, various well-known deep learning architectures use batch normalization[17] to tackle the internal covariate shift problem in deep neural networks, and ReLU[18] to avoid the vanish gradient problem during training (e.g. ResNet[19], DenseNet[20]). Thus, our feature extractor consists of layers where each contains a dilated convolution layer, batch normalization and ReLU, as shown in Fig. 3. The output from the dilated convolution networks is a set of feature vectors a:

$$a = \{a_1, a_2, ..., a_H\}, a_i \in \mathbb{R}^F, \tag{1}$$

where $H = h' \times w'$ is the number of flatten output feature vectors from the last convolutional layer which corresponds to spatial locations $h - 1 \times w$ of the input pathway and F is the number of convolutional filters at the last layer of f_α. By using dilated CNN, we can leverage the pathway image representation by efficiently encoding spatial correlations to feature vectors a which is the input to the attention module (see Section 2.2.2).

2.2.2 Attention mechanism (s_γ) to highlight events

In this paper, we use *soft* attention[21] to produce attention mask p in order to highlight important events in the pathway image that significantly contribute to the model's prediction. The attention mechanism mimics human behavior when focusing on the most informative regions to make decisions. It can also be efficiently trained using popular gradient-based methods in current deep learning frameworks. Our attention module s_γ leverages fully connected neural networks which take feature vectors a and hidden state h_{i-1} from the LSTM (defined in Section 2.2.3) as input in order to generate $score_i$ as in step i in Eq. 2.

$$score_i = s^i_\gamma(a, h_{i-1}). \tag{2}$$

The attention mask p_i, used for visualizing the highlighted areas in Fig.1, is then calculated by passing $score_i$ through the softmax function that produces attention weights from [0;1], as shown in Eq. 3.

$$p_i = softmax(score_i). \tag{3}$$

Finally, attention features att_i are computed using element-wised multiplication between a and p_i to select important (spotlighted) regions.

$$att_i = a \odot p_i. \tag{4}$$

Attention mask p_i gives a score for each extracted feature vector a which corresponds to a location in the pathway image. The higher the value, the more important the events that contributed to the model's prediction.

2.2.3 Long short-term memory network (g_β) for sequence prediction

An LSTM network is used in our framework to improve the performance of predicting sequences and has memory states, hidden state h_i, and cell state c_i at step i. Specifically, during training, the LSTM module takes the most informative extracted features (weighted attention features) att_i from the *soft* attention module's output, previous state (h_{i-1} and c_{i-1}) and previous ground-truth y_{i-1} to predict a target (e.g., a condition or disease) $\tilde{y}_i$ at step i and output information for next prediction (h_i and c_i). Moreover, using previous ground-truth as an input, we can improve the constrains of the LSTM with the teacher enforcing technique[14], as shown in Fig. 3. The LSTM predicts elements one by one $\tilde{y}_i$ in $\tilde{y}$ at every step i until reaching terminate conditions. Note that we do not have previous ground-truth y_{i-1} during the prediction state; therefore, a new prediction $\tilde{y}_i$ is enforced by previous prediction $\tilde{y}_{i-1}$ that improves the predicted sequence's accuracy, as shown in Alg. 1.

3 Results and Discussion

This section first describes the results of the transformations to compose pathway images using the MIMIC-III dataset. The performance results of the proposed Deep EHR Spotlight framework are described in section 3.2 and an evaluation with a domain expert is described in section 3.3.

3.1 Creating Pathway Images from the MIMIC-III Dataset

Pathway images were produced using the MIMIC-III dataset and based on the definitions provided in 2.1. ICD9 codes were reclassified into a smaller number of disease groups of similar codes based on a re-classification system by *Rassekh et. al.*[22]. Other approaches may be used to group ICD codes together for specific use cases, however, this was considered sufficient to train the models presented in this paper and to support building the proposed framework. Selecting a large enough amount of training data was also constrained on the length of the pathways (i.e. the x-axis where event codes are displayed). Overall 58,976 pathway images were produced across 102 different diagnosis groups and 56% of those had a length of 400 or under. The selection of diagnosis groups was based on balancing the most frequent conditions (diagnosis groups) with the lengths of the pathways in order to provide a large enough training set. Furthermore, the pathways selected also have a sequence of conditions y (based on the diagnosis codes groups) which include at most $L = 2$ conditions. Specifically, sequence y begins with one of the three selected main conditions $\{$*Birth Outcome, Cerebrovascular Disease, Ischemic Heart Disease*$\}$, and is followed by an optional second condition from the 99 remaining most frequent conditions. The second condition was also selected based on frequency, as shown in Table 1. A total of 11,400 pathway images (i.e. the first three rows in Table 1) were selected and split 80% for training and 20% for testing. The task undertaken to demonstrate the developed framework involves predicting a first main condition in a pathway followed by a second condition. For this reason the pathway image dimension h associated with conditions was removed from the training set.

Condition (Diagnosis Group)	Abbreviation	Pathways (N)	Precision	Recall	F1
Birth Outcome	BO	6675	0.999	0.996	0.997
Cerebrovascular Disease	Ce	1893	0.916	0.903	0.909
Ischemic Heart Disease	IH	2832	0.933	0.948	0.941
BO → Elective Surgery	BO-NH	616	0.623	0.494	0.551
BO → Perinatal Condition	BO-PC	3931	0.912	0.797	0.850
Ce → Arrhythmia	Ce-Ar	133	0.529	0.187	0.277
Ce → Neurological Disorder	Cr-Ne	310	0.914	0.348	0.504
IH → Cardiomyopathy & HF	IH-Ca	450	0.463	0.221	0.299
IH → Hypertension	IH-Hy	321	0.583	0.219	0.318

Table 1: Performance of the proposed framework (precision, recall and F1) for predicting a given condition or a sequence of two conditions. This table also shows the number of pathway images for each condition. The diagnoses codes used for each condition were selected based on an ICD9 re-classification system[22].

		Predicted Condition								
		BO	BO-NH	BO-PC	Ce	Ce-Ar	Cr-Ne	IH	IH-Ca	IH-Hy
True Condition	BO	**99.86**	0	0	0	0	0	0.13	0	0
	BO-NH	0	**62.32**	26.08	0	0	0	0	0	0
	BO-PC	0	4.58	**91.06**	0	0	0	0	0	0
	Ce	1.40	0	0	**91.54**	0	0	7.041	0	0
	Ce-Ar	0	0	0	5.88	**52.94**	5.88	0	0	0
	Cr-Ne	0	0	0	0	5.714	**91.42**	0	0	0
	IH	0	0	0	6.68	0	0	**93.31**	0	0
	IH-Ca	0	0	0	0	0	0	4.87	**46.341**	14.63
	IH-Hy	0	0	0	0	0	0	14.70	**61.76**	

Table 2: Confusion matrix whose vertical axis shows ground-truth conditions and horizontal axis illustrates predicted conditions (%).

3.2 Performance Evaluation

The proposed framework was first evaluated by computing peformance metrics: precision, recall and F1 score, as shown in Table 1. Whilst these metrics are computed by comparing predicted $\tilde{y}_0$ and labeled y_0, another metric (Intersection over union (IoU), described later) can be calculated to evaluate the predicted sequence $\tilde{y}$ against a ground-truth (labelled) sequence y. With respect to precision and recall, the proposed method achieved adequate scores (over 90% F1 scores) for predicting the main condition alone. However, due to the small amount of data and their heterogeneity, F1 scores for the sequence of two conditions were significantly lower. For example, as seen in Table 1, there are very few training samples for *Cerebrovascular Disease → Arrhythmia* (133 pathways, F1 27%) and Ischemic Heart Disease → Hypertension (312 pathways, F1 31%). It is expected that the proposed method may reach better performance scores for predicting sequences of conditions with sufficiently larger amounts of training data. For example, *Birth Outcome → Perinatal Condition* (3931 pathways) shows an F1 score of 85%. Moreover, a confusion matrix was calculated and is shown in Table 2. The proposed approach shows good performance on *Birth Outcome, Cerebrovascular and Ischemic Heart Disease* reaching precisions 99.86%, 91.54% and 93.31%, respectively. We also observed that the performance gradually drops for *Ischemic Heart Disease → Cardiomyopathy* and *Ischemic Heart Disease → Hypertension* due to a highly imbalanced dataset.

Intersection over union (*IoU*) is a metric for evaluating the differences between predicted sequence $\tilde{y}$ and ground-truth sequence y. *IoU* is measured by overlapping $\tilde{y}$ and y as shown in Eq. 5

$$IoU = 2\frac{\tilde{y} \cap y}{\tilde{y} \cup y}, \tag{5}$$

where IoU is equal to 1 if two sequences are perfectly matched. The calculated average IoU metric across all pathways

Predicted Condition	Specifically Related	Related	Not Related
Birth Outcome (BO)	14	6	0
Cerebrovascular Disease (Ce)	6	14	0
Ischemic Heart Disease (IH)	15	5	0
(BO $\rightarrow$) Elective Surgery	17	3	0
(BO $\rightarrow$) Perinatal Condition	18	2	0
(Ce $\rightarrow$) Arrhythmia	15	4	1
(Ce $\rightarrow$) Neurological Disorder	16	4	0
(IH $\rightarrow$) Cardiomyopathy & HF	7	12	1
(IH $\rightarrow$) Hypertension	8	12	0

Table 3: Evaluation of the top 20 events highlighted by the attention mask for each predicted condition where $(\cdot)$ shows the first predicted condition.

was 0.75, showing agreement between the predicted sequence against the ground-truth.

4 Domain Expert Evaluation

The framework was further evaluated with respect to the attention mask and whether it is highlighting important events across all pathways. While the attention mask p_i values range between 0 and 1, a threshold was set at 0.9 to obtain the top 20 events highlighted by the mask in all pathways of the testing set. Several events may be highlighted in each pathway image, however, only those that contributed most significantly to the predicted condition $\tilde{y}_i$ were selected (i.e., corresponding to the locations of threshold=0.9 in p_i). With the help of a domain expert, the top 20 events were inspected and a determination was made on whether the event codes were: specifically related to the predicted condition(s), generally related (not necessarily specific to the predicted condition), or not related, as shown in Table 3. Despite this evaluation being carried out in a small sample of the highlighted events and by a single observer, it provides reassuring results that most events highlighted as important were indeed related to the predicted condition. One of the limiting factors in this evaluation is the absence of values for the event codes. For example, a large proportion of events highlighted are LOINC codes referring to blood tests and data about the results of the blood tests was not included in our experiments. As described in section 2.1, adding values to each event code is possible, however, that would introduce additional sparseness and more training data would be required to test the developed framework. Figure 5 shows an example of a pathway image and the prediction results provided by the proposed framework as highlighted areas. Pathway image A) in Fig. 5 shows the highlighted areas that are most predictive of *Cerebrovascular* alone, and pathway image B) shows highlighted areas predicting *Cerebrovascular* $\rightarrow$ *Neurological Disorder* for the same pathway. Fig. 5 also shows the pathway image when zoom is applied. A selected segment of the highlighted area was further expanded into text for readability and shows an example of the event codes.

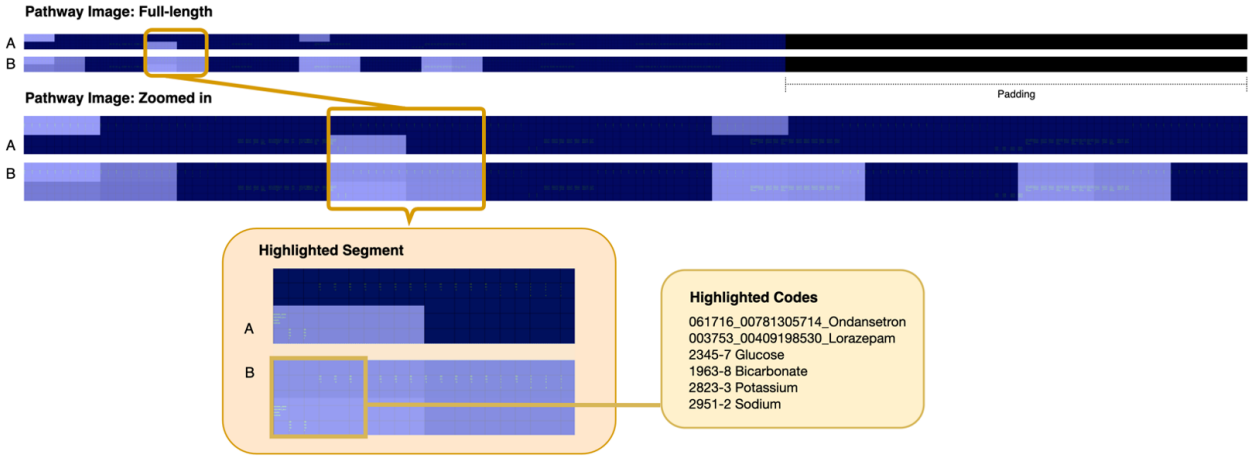


Figure 5: Example of a pathway image (output of the proposed framework) showing the highlighted areas that predicted the pathway's conditions: Cerebrovascular disease alone (A) and Cerebrovascular disease followed by a Neurological Disorder (B). The full-length image shows a darker area denoting padding.

5 Conclusion

This paper proposes a new framework for predicting and highlighting important clinical events from electronic health records (EHRs). In particular, this paper proposes to transform EHR data into pathways and 2D pathway images, which can then be used with two dimensional CNN techniques to support visual interpretation. The proposed Deep EHR Spotlight framework can highlight regions in pathway images which are particularly important for the predictions. In this paper we used the MIMIC-III dataset, which produced highly sparse pathway images. The performance results were adequate (i.e. F1 scores $> 90\%$) when a larger number of training data was available. The top events appearing in highlighted masks were also evaluated by a domain expert and found to be mostly related and specific to the predicted conditions. These are reassuring results that demonstrate the value in the proposed approach. However, further work is needed to substantiate these results, improve the methods for a specific use case and compare with other techniques and approaches. As future work we are planning to test the proposed framework on significantly larger datasets as well as remodeling the pathway images and their dimensions for more specific clinical use cases. This will allow predicting individual conditions and taking into account events with associated codes (e.g. haemoglobin test) and respective values (e.g. 20 g/dL). Further work is also needed to continue to evaluate the events highlighted by the attention mask for different thresholds.

References

1. C. Xiao, E. Choi, and J. Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 10 2018.

2. Joao H Bettencourt-Silva, Jeremy Clark, Colin S Cooper, Robert Mills, Victor J Rayward-Smith, and Beatriz de la Iglesia. Building data-driven pathways from routinely collected hospital data: A case study on prostate cancer. *JMIR Medical Informatics*, 3(3):e26, Jul 2015.

3. Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. volume 68 of *Proceedings of Machine Learning Research*, pages 322–337, Boston, Massachusetts, 18–19 Aug 2017. PMLR.

4. Q. Suo, F. Ma, Y. Yuan, M. Huai, W. Zhong, J. Gao, and A. Zhang. Deep Patient Similarity Learning for Personalized Healthcare. *IEEE Transactions on NanoBioscience*, 17(3):219–227, 07 2018.

5. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceeding of ICML*, pages 2048–2057, 2015.

6. Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

7. Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

8. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

9. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proceeding of NIPS*, pages 3104–3112, 2014.

10. Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

11. Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceeding of NIPS*, pages 3504–3512, 2016.

12. Qiuling Suo, Fenglong Ma, Ye Yuan, Mengdi Huai, Weida Zhong, Jing Gao, and Aidong Zhang. Deep patient similarity learning for personalized healthcare. *IEEE Transactions on NanoBioscience*, 17(3):219–227, 2018.

13. A. E. Johnson, T. J. Pollard, L. Shen, L. W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, May 2016.

14. Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.

15. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

16. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

17. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

18. Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceeding of ICML*, 2010.

19. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR, pages=770–778, year=2016*.

20. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of CVPR*, pages 4700–4708, 2017.

21. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceeding of ICLR*, 2015.

22. S. R. Rassekh, M. Lorenzi, L. Lee, S. Devji, M. McBride, and K. Goddard. Reclassification of ICD-9 Codes into Meaningful Categories for Oncology Survivorship Research. *Cancer Epidemiol*, 2010:569517, 2010.

Understanding Clinical Trial Reports: Extracting Medical Entities and Their Relations

Benjamin E. Nye, MS[1], Jay DeYoung, MS[1], Eric Lehman, BS[1], Ani Nenkova, PhD[2], Iain J. Marshall, MD, PhD[3], Byron C. Wallace, PhD[1]
[1]Northeastern University, Boston, MA; [2]University of Pennsylvania, Philadelphia, PA; [3]King's College London, London

Abstract

The best evidence concerning comparative treatment effectiveness comes from clinical trials, the results of which are reported in unstructured articles. Medical experts must manually extract information from articles to inform decision-making, which is time-consuming and expensive. Here we consider the end-to-end *task of both (a) extracting treatments and outcomes from full-text articles describing clinical trials (entity identification)* and*, (b) inferring the reported results for the former with respect to the latter (relation extraction). We introduce new data for this task, and evaluate models that have recently achieved state-of-the-art results on similar tasks in Natural Language Processing. We then propose a new method motivated by how trial results are typically presented that outperforms these purely data-driven baselines. Finally, we run a fielded evaluation of the model with a non-profit seeking to identify existing drugs that might be re-purposed for cancer, showing the potential utility of end-to-end evidence extraction systems.*

1 Introduction

Currently, Randomized Controlled Trials (RCTs) pertaining to specific clinical questions are manually identified and synthesized in *systematic reviews* that in turn inform guidelines, health policies, and medical decision-making. Such reviews are critically important, but onerous to produce. Moreover, reliance on these manually compiled syntheses means that even when a systematic review relevant to a particular clinical question or topic exists, it is likely that new evidence will have been published since its compilation, rendering it out of date[1]. Language technologies that make the primary literature more *actionable* by surfacing relevant evidence could expedite evidence synthesis[2] and enable health practitioners to inform care using the totality of the available evidence.

Results from individual trials are disseminated as generally unstructured text but will contain descriptions of key components: The enrolled *Population* (e.g., diabetics), the *Intervention* (e.g., beta blockers), the *Comparator* treatment (e.g., placebos), and finally the *Outcomes* measured. Collectively, these are known as the PICO elements. The key findings in an RCT relate these elements by reporting whether a specific intervention yielded a significant difference with respect to an outcome of interest, as compared to a given comparator. These results — reported results for ICO triplets — are what we aim to extract from trial reports.

Prior work has considered the task of automatically extracting snippets describing PICO elements from articles describing RCTs[3–7]. Other efforts have focused on identifying and analyzing scientific claims[8,9]. More directly of relevance, prior efforts have also considered the inference problem of extracting the reported finding in an article for a given ICO triplet[10].

Extracting a semantically meaningful structured representation of the evidence presented in journal articles describing results of RCTs is a critical task for enabling a wide range of interactions with the medical literature. Patients may, e.g., want to know which side effects are associated with a particular medication; clinicians may wish to know which health outcomes are likely to be affected by a given treatment; and policy makers need to know which healthcare strategies are most efficacious for a particular disease. However, a usable, end-to-end system must both identify ICO elements within a trial report and infer the findings concerning these: This is the challenge we address in this work.

We design, train, and evaluate systems that extract ICO triplets (specifying which interventions were assessed, and with respect to which comparators and outcomes), *and* infer the corresponding reported findings (Figure 1), directly from the abstract text. This poses difficult technical challenges — e.g., grouping mentions of the same underlying intervention — that we aim to address. While methods and systems that address the sub-components of this task have been previously proposed, as far as we are award this is the first attempt to design a system for the *end-to-end* evidence extraction task, including ICO identification and inference.

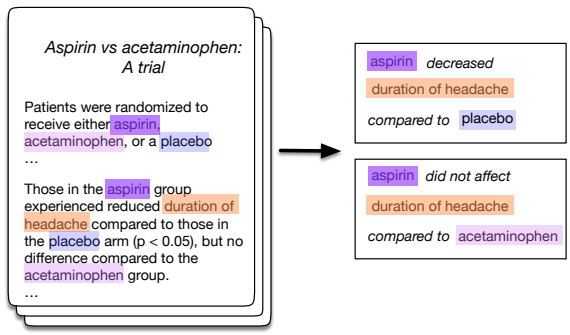


Figure 1: We propose models to extract key clinical entities (interventions, comparators, and outcomes) from reports of randomized controlled trials (RCTs) as well as the reported findings concernig these. Prior work has considered these tasks only in isolation.

Our contributions in this work are as follows.

- We introduce the challenging task of *clinical evidence extraction*, and we provide a distantly supervised training dataset along with a new directly supervised test set for the task.
- We propose a novel approach informed by language use, and compare it against current state-of-the-art joint end-to-end NLP models such as DyGIE++[11].
- We evaluate these models both quantitatively and qualitatively. We ablate components and modes of supervision. And we find that while F-scores across models considered appear low in absolute terms (which is unsurprising given the difficulty of the task), domain experts nonetheless find model outputs for an example application — identifying candidate drugs that might be repurposed for cancer — useful.

1.1 End-to-End Evidence Extraction

The two core subtasks inherent to evidence extraction are identifying ICO elements and inferring reported relationships between them. These tasks may be viewed as instances of Named Entity Recognition (NER) and relation extraction (RE), respectively. These general tasks have been extensively studied in the prior work that we build upon here. Traditional approaches to relation extraction use a pipeline approach that entails first extracting relevant entity mentions, and then passing them forward to a relation extraction module. Recent work has proposed performing joint extraction of entities and relations, allowing information to be shared between these related tasks[12–14].

When performed jointly with NER, relation extraction is typically treated as a sentence-level task in which interactions are only evaluated between entity mentions that co-occur within a limited range. This limitation is crippling in the abstracts of RCT articles; conclusion sentences conveying the relation of interest only explicitly mention the primary intervention 28% of the time,[1] often using an indirect coreferent expression such as "The propofol consumption was similar in the four groups", or omitting it entirely as in: "There was no difference in the use of inotropes, vasoconstrictors or vasodilators."

Further, the directionality of reported results (i.e., whether the intervention *significantly increased*, *significantly decreased*, or induced *no significant difference* relative to the comparator, with respect to an outcome) are conventionally reported with respect to an implicit primary intervention. In a statement such as "The consumption of both propofol and sevoflurane significantly decreased", the provided evidence for a relation requires knowing which trial arm is the

[1]Most trials investigate a particular, often new, intervention of interest and compare this against a placebo or existing standard of care; this is reflected in the framing of the trial. We refer to this intervention as the 'primary' intervention, for want of a better term.

primary intervention, and which the comparator. This information, especially at the abstract level, is typically available only at the beginning of the text. To extract these relations, we must therefore draw upon context derived from the entire document.

Recent work on document-level relation extraction has shown promising results[11], but work on medical texts so far has been limited in scope to subareas with comparatively standardized entities (e.g., chemicals or genes) that can be reliably identified and linked to a structured vocabulary via synonym matching[15]. By contrast, the space of medical interventions is vast, ranging from pharmacological treatments to prescribed animal companions.

We operationally define *clinical entities* as concepts describing: trial participants (and their conditions); treatments (interventions) that participants were randomized to receive, and; outcomes measured (including measurement scales) to determine treatment efficacy. These clinical entities collectively describe the key characteristics of trials and provide the context for interpretation of the reported statistical results. In evidence extraction we aim to identify N-ary relations that capture interactions between treatments and outcomes (Figure 1).

We say that (*intervention*, *comparator*, *outcome*) triplets exhibit a relation if there is a reported measurement for the outcome with respect to the intervention and comparator. We derive candidate relation labels from the Evidence Inference corpus[16]: The relative effect of the intervention can be *increased*, *decreased*, or *not statistically different*. We also consider a relaxed version of this task that considers only binary relations between (*interventions* and *outcomes*. This simplification (in which the comparator is implicit) permits direct comparison to existing models that only consider binary relations. One may interpret this as asking "what is the comparative effect of this intervention with respect to this outcome, as compared to whatever was used as the baseline".

Each entity is grounded to the abstract as a list of mentions. Although entity-focused tasks in related domains often link entities to structured vocabularies such as the Unified Medical Language System (UMLS), such a mapping is difficult in the highly variable setting of general clinical interventions and outcomes. Outcomes in particular are often complex combinations of several concepts, such as "Duration of pain after opening the tourniquet." For this reason we eschew explicit entity linking for evaluation, and instead say that a predicted relation between extracted mentions constitutes a prediction for the corresponding entities.

To evaluate systems that attempt to perform this task, abstracts must be annotated with: All unique intervention and outcome entities; Mentions (spans) corresponding to each of these, and; The directionality of reported findings for reported comparisons. Independent corpora exist for extracting intervention and outcome spans[6] and for identifying reported findings concerning these clinical entities[16]. However, EBM-NLP does not include groupings of mentions into unique entities, and documents in the Evidence Inference assume ICO entities are *given* as a "prompt", and these only sometimes are taken verbatim from the corresponding article. Furthermore, the Evidence Inference corpus only contains annotations for *some* of the evaluations for which results were reported, i.e., these are non-exhaustive.[2]

Because we do not have direct supervision for this task, we relied on *distant supervision*[17], i.e., noisy automatically derived 'labels'. We derived this from existing corpora. Specifically, we used the EBM-NLP corpus to train a model that we then use to identify (all) entity mentions in the Evidence Inference dataset. For development and testing data, we collected new *exhaustive* annotations from domain experts (medical doctors) on 60 and 100 abstracts, respectively.[3]

2 Methods

2.1 Data

As mentioned above, corpora exist for the constituent tasks of extracting interventions and outcomes[6], and for inferring results for a given intervention and outcome[16], but not for the proposed end-to-end task. We therefore use existing datasets and heuristics to derive a relatively large, distantly supervised training set (Section 2.1). We additionally collect relatively small development and train sets explicitly annotated by domain experts (Section 2.1).[4] We will make all data available upon publication

[2]Evidence Inference includes annotations over full-texts, but here we work with an abstract-only subset of this data.

[3]We will publicly release this data alongside publication.

[4]This is expensive, as we acquire *exhaustive* annotations from individuals with medical degrees.

	Train		Dev		Test	
Abstracts	1,772	(1.00)	60	(1.00)	100	(1.00)
Relations	4,565	(2.58)	200	(3.33)	289	(2.29)
Entities	12,556	(7.09)	531	(8.85)	808	(8.08)
Mentions	29,908	(16.88)	1,163	(19.38)	1,788	(17.88)

Table 1: Total count (and average per-document) for data types in the distantly supervised train set and expert-labeled dev and test sets.

Distant Supervision

Documents in the Evidence Inference corpus include triplets comprising: (i) Text describing an intervention; (ii) Text describing a comparator; and (iii) Text describing an outcome. For each such "ICO" triplet, a label that indicates the directionality of the reported result is provided, along with a *rationale* span extracted from the source text that provides evidence for this conclusion. For instance, in the illustrative example depicted in Figure 1, the label for (*aspirin*, *placebo*, *duration of headache*) would be *decreased* and the supporting rationale for this would be the snippet "Those in the aspirin group experienced reduced duration of headache compared to those in the placebo arm ($p<0.05$)".

To identify mentions of the entities, we follow Nye *et al.*[6] in training a BiLSTM-CRF sequence tagging model on the EBM-NLP data. To encode tokens for the NER model we use SciBERT, which is currently state-of-the-art for this task[3]. Predicted mentions are then assigned to the provided entity that has the highest cosine similarity with respect to the embeddings induced by SciBERT. Any mention that exceeds a maximum distance threshold is assigned to a new entity that does not instantiate any relations. We tune this distance threshold over the development set.

Evaluation Data

To accurately evaluate performance for this task we collected exhaustive manual annotations over 160 abstracts. For this we hired personnel with medical degrees via Upwork.[5] Annotation was completed in two phases. The first step entailed identifying the relations; for this we followed[16], except that we requested annotators exhaustively identify *all* relations reported in abstracts (rather than only a subset of them in full-texts).

In the second step annotators marked all intervention and outcome mentions in the abstract, and then grouped these into distinct entities. Annotators were instructed to only highlight explicit mentions and to ignore coreferent spans that lacked meaning without the head mention (e.g., "Group 1"). Mentions were grouped only when directly interchangeable in the context of the RCT.

We hired and trained five expert annotators to perform the second phase, but we only retain annotations from the two most reliable annotators due to the difficulty of the task. On a multiply-annotated subset of the development set, inter-annotator agreement for identifying and grouping mentions were calculated using B^3, *MUC*, $CEAF_e$[18]. Overall scores were 0.40, 0.46, and 0.42 respectively. Each abstract took between 6 and 37 minutes (average 17) to annotate, depending on complexity. The total cost to annotate 160 documents exhaustively was $\sim$\$600.

Across the expert-labeled data the average trial contains 2.62 different treatment arms, necessitating identification of intervention and comparator entities for each evidence claim rather than at the document level. An additional complication is that sentences presenting a result only make explicit mention of the relevant intervention and comparator arms 37% of the time, instead using coreferent mentions (e.g. "Headaches were reduced in the treatment group") 40% of the time, or implicit mentions (e.g. "Overall prevalence of adverse effects was decreased") 31% of the time.

2.2 Modeling

We now describe the models we evaluate for the proposed task: Existing, transformer-based systems that perform extraction and relation extraction jointly and a new approach that we propose informed by how trials tend to report results.

[5] http://www.upwork.com

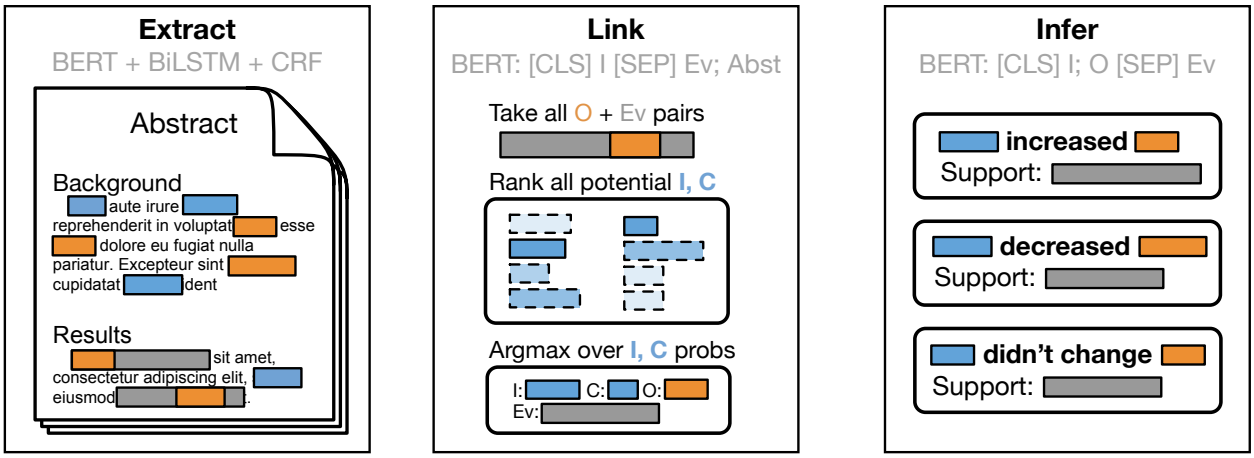


Figure 2: In our proposed Extract, Link, Infer (ELI) method, we first Extract all snippets describing treatments and outcomes, and evidence-bearing sentences. Outcome snippets found within an evidence sentence are then Linked to the most probable abstract-level intervention. The direction of the finding for the selected (intervention, outcome) pair is then Inferred from the evidence-bearing context in which the outcome appeared.

Joint Models

Models that perform entity recognition, coreference resolution, and relation extraction jointly may benefit from sharing information across tasks. Such models have recently achieved state-of-the-art performance on benchmark tasks, but they can be tricky to train, and it is difficult to incorporate prior knowledge for specific domains into such data-driven models. For the present task we evaluate two recently proposed models that most closely match our task setting.

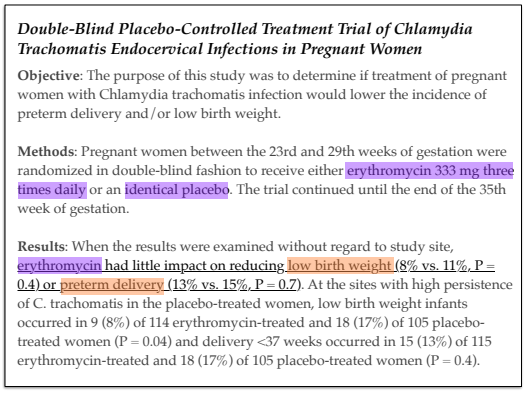
The first candidate we consider reports the best known results on the Biocreative V CDR dataset[19], for which the task is to identify relations between chemicals and diseases in scientific abstracts. The Bi-affine Relation Attention Network (BRAN) proposed by[15] jointly learns to predict entity types and relations across full abstracts, as well as aggregating across the mentions of each entity. This model relies on existing NER labels which are easily obtained for chemicals and diseases, but more challenging in the broader domain of our task.

We also compare to DyGIE++, which recently achieved top results on several scientific extraction datasets, including SciERC, GENIA, and ACE05[11,20]. This model performs joint NER, coreference, and relation extraction, but does so at the sentence level by iterating over all possible mention pairs. This strategy does not readily scale up to processing full abstracts due to the significant increase in the number of mentions. In light of training and data constraints, we disable the coreference module and additional propagation layers in DyGIE++.

Our Approach: Extract, Link, Infer

Motivated by observations concerning how results tend to be described in trial reports, we propose a new approach that works by first independently identifying (i) spans describing interventions and (ii) snippets that report key results (i.e., that report the observed comparative effectiveness between two or more treatments, with respect to any outcome). Trial reports may present findings for multiple intervention comparisons, with respect to potentially many outcomes. In a second step we therefore link the identified evidence-bearing snippet to the extracted outcome and intervention to which it most likely pertains. This Extract, Link, Infer (ELI) approach (Figure 2) therefore effectively works backwards, first identifying evidence statements and then working to identify the clinical entities that participate in these reported findings.

As an illustrative example, consider Figure 3. The main findings are reported in the underlined snippet: "erythromycin had little impact on reducing low birth weight (8% vs. 11%, P = 0.4) or preterm delivery (13% vs. 15%, P = 0.7)". There are two outcomes here: "low birth weight" and "preterm delivery". We need to link these findings to the primary intervention that they concern. The two interventions discussed in this abstract are: "erythromycin 333 mg



Double-Blind Placebo-Controlled Treatment Trial of Chlamydia Trachomatis Endocervical Infections in Pregnant Women

Objective: The purpose of this study was to determine if treatment of pregnant women with Chlamydia trachomatis infection would lower the incidence of preterm delivery and/or low birth weight.

Methods: Pregnant women between the 23rd and 29th weeks of gestation were randomized in double-blind fashion to receive either erythromycin 333 mg three times daily or an identical placebo. The trial continued until the end of the 35th week of gestation.

Results: When the results were examined without regard to study site, erythromycin had little impact on reducing low birth weight (8% vs. 11%, P = 0.4) or preterm delivery (13% vs. 15%, P = 0.7). At the sites with high persistence of C. trachomatis in the placebo-treated women, low birth weight infants occurred in 9 (8%) of 114 erythromycin-treated and 18 (17%) of 105 placebo-treated women (P = 0.04) and delivery <37 weeks occurred in 15 (13%) of 115 erythromycin-treated and 18 (17%) of 105 placebo-treated women (P = 0.4).

Figure 3: An example abstract. Intervention snippets are highlighted in purple, outcomes in orange. The main evidence-bearing snippet is underlined.

three times daily" and "identical placebo"; the former is the treatment of interest. Finally, we can infer the direction of the reported finding for the primary treatment for extracted outcomes: erythromycin yielded *no significant difference* in both outcomes, versus placebo.

Following this illustration, in ELI we decompose evidence extraction into independent components that operate in two phases over inputs. Using independent components brings drawbacks: We cannot borrow strength across tasks, and errors will cascade through the system. But it also allows us to explicitly capitalize on domain knowledge about how results are reported in such texts — as in the preceding example — in order to simplify training and render the task more tractable. Note that in purely data-driven systems such as BRAN and DyGIE++, there is no natural means to operationalize the intuitive strategy just outlined.

The initial stage of the ELI pipeline comprises two independent tasks. First, all mentions of interventions and outcomes are extracted using the sequence tagging model described in Section 2.1. Second, all sentences are classified as containing evidence-bearing snippets (or not). For this sentence classification model, we add a linear layer on top of SciBERT representations.

We construct training data for this using evidence spans from the Evidence Inference corpus[16]. We take as positive sentences any that overlap any annotated evidence spans; for negative samples we use sentences of similar length from the same document. This model realizes 0.97 recall and 0.53 precision on the Evidence Inference test set.[6]

These extracted spans are passed forward to a second stage, in which a model attempts to determine which clinical entities are referred to in a particular evidence span. We obviate the need for an explicit model to link outcomes to evidence spans by observing that reported conclusions almost always contain an explicit mention of the relevant outcomes within the same sentence as the stated result. Therefore, we only consider outcome mentions that occur inside one of the extracted evidence spans. In the development set, 87% of outcome entities are directly mentioned in an evidence statement.

We train a second sentence pair classification model that takes as input an extracted intervention span and an evidence sentence, and predicts if the given intervention is the primary treatment, the comparator, or is unrelated to the given evidence span. We train this model with the (intervention, comparator, evidence span) triplets provided in the Evidence Inference corpus, augmented with synthesized negatives selected to mimic the failure modes of the NER tagger — extraneous interventions from within the same document, compound phrases involving multiple interventions, and random spans from other locations in the document. This model is then used at test time to select the most probable intervention mention for a given evidence sentence, producing a pair of interacting intervention and outcome mentions linked to the corresponding evidence sentence.

Finally, a linear classification layer is fine-tuned on top of SciBERT that takes the assembled relation candidate and pre-

[6]The modest precision may reflect the fact that evidence snippets are not exhaustively labeled in the dataset.

Entity Extraction	P	R	F1
DyGIE++	0.45	0.47	0.46
ELI	0.46	0.69	0.55

Relation Inference	P	R	F1
BRAN	0.05	0.41	0.08
DyGIE++	0.24	0.13	0.17
ELI	0.33	0.31	0.32

Table 2: System performances on the intermediate entity extraction and the end-to-end inference tasks.

dicts the directionality of the finding with respect to the given intervention and outcome. As reported in prior work[16], if ground truth evidence spans are given, predicting the direction of the findings reported in these is comparatively easy: Models achieve an F1 score of 0.80 on this three-way classification task.

2.3 Experimental Details

All of our components operate over representations yielded from BERT[22], specifically the pretrained SciBERT[3] instance. We use the Adam optimizer[21], with learning rate $1e^{-3}$.

For DyGIE++, we use the default configuration, except: We increase the loss weight for relations from 1 to 10, we disable the coreference module, and we disable relation propogation. DyGIE++ fine-tunes BERT parameters, using the BertAdam optimizer[22]. We truncate inputs to 300 tokens (roughly the mean input length), discarding mentions beyond this.

3 Results

We present quantitative results for the proposed end-to-end task in Section 3.1, comparing ELI to modern transformer-based joint models that represent the SOTA for similar tasks. We then ablate the components of ELI in Section 3.2 to characterize where the system does well and where it fails, highlighting directions for improvement.

All systems we evaluate perform relatively poorly in absolute terms on this challenging task. To better characterize system performance, we therefore enlist domain experts from a non-profit organization to qualitatively assess the accuracy and potential utility of model outputs.

3.1 Main Task Results

We report results for both the intermediate aim of entity extraction and the final task of inferring relations between entities in Table 2. As expected, the task of mapping raw inputs to structured results relating clinical entities is quite challenging, yielding low absolute numbers for current state-of-the-art joint models. While DyGIE++ achieves moderate success (0.46 F1) with respect to identifying mentions of the relevant clinical entities, performance with respect to the final task of identifying the relations between them leaves much to be desired.

Our proposed method, ELI, fares considerably better than DYGIE++, with relative F1 increases of 20% and 88% for entity extraction and relation inference respectively. That said, in absolute terms the performance of ELI is also low (0.32 F1), highlighting the challenging nature of the task. However, we note that we note that even while performance seems poor here in absolute terms with respect to F1, our evaluation in which domain experts directly assess model outputs (Section 3.3) suggests that the ELI system is already practically useful for downstream applications.

One of the primary potential drawbacks of employing a series of independent models is that errors made in an early stage will propagate through the system, resulting in lower quality inputs and higher error rates for the subsequent models. In the next Section we ablate components of ELI to highlight potential means of improving this performance.

3.2 Ablations and Analysis

To investigate the sources of error that accumulate over the course of the ELI modeling pipeline, we evaluate each component model individually.

The extraction phase, which consumes raw unlabeled documents, produces per-token labels for intervention and outcome mentions. We achieve 0.55 F1 (0.69 recall) at the token level. However, a system does not need to extract every

Extraction	P	R	F1	**Linking**	Acc	**Inference**	P	R	F1
tokens	0.45	0.69	0.55	interventions	0.75	increased	0.64	0.90	0.75
entities	0.58	0.85	0.69	comparators	0.70	decreased	0.85	0.54	0.66
evidence	0.45	0.98	0.62	outcomes	0.78	no difference	0.93	0.94	0.93

Table 3: Performance of each ELI module on the test set, if given ground-truth inputs.

mention of an entity in order to arrive at the correct conclusions. As long as at least one of an entity's mentions are correctly extracted, it is still a candidate for participating in different relations. We therefore evaluate clinical extraction at the entity level as well, where an entity is marked as extracted as long as we identify at least one of its mentions. While this is more permissive from a recall perspective, we treat any extracted span that is not a mention of a ground-truth entity as a false positive. This produces a pessimistic view of the true precision, since the metric penalizes repeated extraction of spans that represent a single false entity. At the entity level we improve extraction scores to 0.69 F1, and more importantly recall increases to 0.85.

Extraction of evidence sentences is very high recall (0.98) with middling precision, and we rely on the linking phase to correctly reject evidence sentences that do not contain conclusive statements about specific outcomes.

We next isolate the performance of the linking phase by providing ground-truth evidence sentences and entity mentions as inputs. The model ranks all intervention and comparator entities, linking the most probable for each evidence sentence. We observe that 78% of outcome entities are contained within an evidence sentence (and therefore automatically linked), and the model reaches 75% and 70% accuracy for selecting the correct intervention and comparator, respectively.

The inference model is then given all ground-truth outcome spans and their corresponding evidence sentences. Instances for which *no significant difference* is reported are easy to classify (0.93 F1), while significant changes are more difficult (0.71 F1). The majority of mistakes come from mislabeling decreases as increases; in many cases this error is caused by cases in which a decrease in an undesirable outcome is reported positively (e.g. "adverse effects were improved in the treatment group").

3.3 Application Example: Helping RebootRX Find (Potential) Cancer Treatments

To assess the practical utility of the presented task and systems, a domain expert at `rebootrx.org`, a non-profit organization that seeks to identify previously studied drugs from the literature that might be repurposed to treat cancer, evaluated outputs from the pipeline system on 20 abstracts from RCTs investigating cancer treatments.

The clinical entities in this sub-domain are particularly challenging to extract and differentiate: Interventions typically consist of compound treatments with complicated dosage schedules, and similar outcomes are often measured repeatedly at different times. RebootRX seeks to identify all RCTs in which specific interventions were used to evaluate outcomes of interest; this is an information retrieval problem that maps directly on to our proposed task of evidence extraction.

The domain expert was asked to assess the per-document recall and precision of the extracted relations on a five point Likert scale, and to note which aspects of erroneous predictions were incorrect. Overall, the annotator scored the ELI system at 80.0% recall and 63.8% precision. They noted that a mistake in predicting the directionality of the results accounted for 47.6% of the extraction errors. Therefore, despite the ostensibly low absolute scores on the strict evaluation performed above using our test set, these results indicate that even current models for the proposed task are useful in a meaningful, real-world setting.

4 Discussion

Automating structured evidence extraction has the potential to provide better access to emerging clinical evidence. In the immediate future, this may reduce the manual effort needed to produce and maintain systematic reviews. Thinking longer term, if we can improve the accuracy of such end-to-end systems we may be able to eventually realize *living* information systems, where health professionals could draw upon real-time assessments of the evidence to inform

their decision-making.

Realizing this potential will require improving methods for automated evidence extraction, which in turn necessitates addressing NLP challenges relating to joint extraction and grouping of heterogeneous treatment and outcome mentions from relatively lengthy inputs, and inference concerning the reported relations between these. This domain and setting poses several interesting obstacles such as frequent use of coreference and implied mentions, and relations requiring document-level context.

A system for extracting rich, structured data from a completely unlabeled document has many opportunities to make mistakes, and this composition of complex tasks provides significant challenges for current state of the art systems. Our proposed model achieves strong results on each component task, and even so overall performance leaves room for improvement. Our hope is that this difficult, important task motivates innovations to meet these challenges.

One avenue of future research that this motivates is how one might integrate intuitions about language use in particular tasks into end-to-end, joint systems. Other promising leads include extending ELI to incorporate distantly supervised labels, as well as refinement of the supervision strategy.

5 Conclusion

We have proposed the end-to-end task of automatically extracting structured evidence — interventions, outcomes, and comparative results — from trial reports. This differs from prior work which has considered the tasks of data extraction and inferring results separately.

We introduced new data for this task (which we will make publicly available), including a large distantly supervised train set and expert annotated development and test sets. The latter feature *exhaustive* annotations, inclusive of all ICO triplets for which results have been reported in the corresponding abstract. Using this data, we evaluated state-of-the-art joint NLP models for this challenging task and proposed a new modular method — Extract, Link, Infer (ELI) — motivated by observations about how authors convey findings in trial reports. This model yielded superior performance on the task, presumably due to its implicit incorporation of domain knowledge concerning how trial reports are structured.

We performed a fielded evaluation in collaboration with a non-profit (RebootRX) that is interested in identifying results from previously conducted studies on drugs that might be used to treat cancer. This small study suggested that despite low absolute performance metrics, the proposed model can be useful in practice.

References

[1] Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9).

[2] James Thomas, John McNaught, and Sophia Ananiadou. 2011. Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1):1–14.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained language model for scientific text. In *EMNLP*.

[4] Di Jin and Peter Szolovits. 2018. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75.

[5] Grace E Lee and Aixin Sun. 2019. A study on agreement in PICO span annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1149–1152.

[6] Benjamin Nye, Junyi Jessy Li, Roma Patel, et al. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. *CoRR*, abs/1806.04185.

[7] Lena Schmidt, Julie Weeds, and Julian Higgins. 2020. Data mining in clinical trial text: Transformers for classification and question answering tasks. *arXiv preprint arXiv:2001.11268*.

[8] Catherine Blake. 2010. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of biomedical informatics*, 43(2):173–189.

[9] Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11.

[10] Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

[11] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *ArXiv*, abs/1909.03546.

[12] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *CoRR*, abs/1804.07847.

[13] Fei Li, Meishan Zhang, Guohong Fu, and Donghong Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1):198.

[14] Dat Quoc Nguyen and Karin Verspoor. 2018. End-to-end neural relation extraction using deep biaffine attention. *CoRR*, abs/1812.11275.

[15] Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, New Orleans, Louisiana. Association for Computational Linguistics.

[16] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. *arXiv preprint arXiv:1904.01606*.

[17] Mike Mintz and Steven Bills and Rion Snow and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. *Association for Computational Linguistics (ACL).*

[18] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

[19] Jiao Li, Yueping Sun, Robin J Johnson, et al. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

[20] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

[21] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Detection of Anomalous Patterns Associated with the Impact of Medications on 30-Day Hospital Readmission Rates in Diabetes Care

William Ogallo, RPh, PhD, Girmaw Abebe Tadesse, PhD, Skyler Speakman, PhD, Aisha Walcott-Bryant, PhD
IBM Research – Africa, Nairobi, Kenya

Abstract

Improving quality of care in diabetes requires a good understanding of variations in diabetes outcomes and related interventions. However, little is known about the impact of diabetes interventions on outcome measures at the subpopulation-level. In this study, we developed methods that combine causal inference techniques with subset scanning techniques to study the heterogeneous effects of treatments on binary health outcomes. We analyzed a diabetes dataset consisting of 70,000 initial inpatient encounters to investigate the anomalous patterns associated with the impact of 4 anti-diabetic medication classes on 30-day readmission in diabetes. We discovered anomalous subpopulations where the likelihood of readmission was up to 1.8 times higher than that of the overall population suggesting subpopulation-level heterogeneity. Identifying such subpopulations may lead to a better understanding of the heterogeneous effects of treatments and improve targeted intervention planning.

Introduction

Thirty-day hospital readmission is an important outcome measure for assessing the quality of care given to diabetes patients. Reducing readmission rates among diabetes patients could improve care and reduce care-related costs[1]. However, although diabetes patients have an increased risk of readmission, little research has been done on this subject[1]. Some of the key barriers to understanding the risk factors of readmissions in diabetes are complicated by the natural variations in diabetes outcomes and related interventions. For example, care providers may use different treatments for their patients, and patients may respond differently to the same treatments. To improve the quality of care in diabetes care, there is a need for robust approaches for investigating variations at the subpopulation level. This is particularly useful since methods that analyze individual patients may fail to identify subtle patterns that are discernable when groups of patients are considered collectively, while methods that generate aggregate statistics for entire populations may fail to detect small-scale patterns[2].

Traditional approaches to investigating subpopulation-level heterogeneity have relied on manual stratification of covariate profiles. For example, an outcome such as mortality can be stratified by age, gender, and ethnicity to identify high-risk subpopulations. However, this is limited to analyzing only a few features beyond which it becomes computationally infeasible. Furthermore, these approaches lack a data-driven knowledge discovery aspect as investigators must suggest a priori which features they would like to stratify across, and may also inadvertently lead to data manipulation as investigators attempt to produce desired p-values ('p-hacking'). Machine learning approaches have also been used to investigate heterogeneity. Techniques such as LASSO regression can be used to select important covariates, while decision tree regression can be used to recursively partition data. However, these approaches are either subject to several modeling assumptions and limitations or lack adequate interpretability[3]. Fortunately, recent advancements in the anomalous pattern detection literature enable the scalable and unsupervised discovery of specific subpopulations (subsets) that are anomalous. These subset scanning methods focus on identifying anomalous subsets of records in a multidimensional array that differ from expected behavior. Herein, anomalousness is quantified using a scoring function that is typically a log-likelihood ratio statistic[2]. The scoring function is maximized over the exponentially-many combinations of feature values to identify the subset with the highest score. This function must satisfy the linear time subset scanning (LTSS) property so that the search can be done in linear rather than exponential time[2].

Some of the key subset scanning techniques include Bias-Scan[4], treatment effect subset scanning (TESS)[3], and anomalous patterns of care (APC) Scan[5]. Bias-Scan focuses on the discovery of the subpopulation with the most divergence between the true outcomes and the predicted probabilities of a binary classifier. TESS discovers heterogeneous treatment effects by identifying the subpopulation in a randomized controlled trial that is most significantly impacted by the studied treatment. APC Scan extends TESS to enable anomalous pattern detection in observation data by incorporating multiple treatments and propensity score weighting to account for observable differences between treated

and untreated patients. While these techniques can be applied in health and research informatics domains, certain limitations have to be addressed to improve their utility. For example, Bias-Scan is primarily used for the assessment of bias in predictive binary classifiers and although it analyzes binary outcomes, it has not been adopted for use in the assessment of heterogeneous effects of treatments. On the other hand, TESS and APC Scan are primarily designed for scalar outcomes and are currently limited to the discovery of heterogeneous effects of single interventions in randomized controlled trials (i.e., TESS) or multiple interventions with the assumption of temporal independence between the interventions (i.e., APC).

The overarching goal of our research is to extend and generalize the application of anomalous pattern detection techniques from subset scanning literature to enable the efficient discovery of anomalous patterns in large-scale observational health data such as electronic health records. The objectives of this study were threefold. First, we proposed an approach for selecting the least biased propensity score (PS) model among multiple PS models to overcome treatment selection bias in observational studies. Second, we developed algorithms combining causal inference and anomalous pattern detection techniques to discover the heterogeneous effects of interventions on binary outcomes. Third, we demonstrated the application of the developed techniques to discover anomalous patterns associated with the impact of diabetes medications on 30-day hospital readmission in diabetes care.

Methods

Propensity score based bias smoothing

Causal inference and anomalous pattern detection on observational studies require smoothing of the treatment assignment bias that accounts for observable differences (bias) between treated and untreated groups. Propensity score models, which are often used for such tasks, model the probability of receiving a treatment (p_s) conditioned on observed baseline covariates, i.e., $p_s(\mathbf{X}) = p(Z = 1|\mathbf{X})$, where X is a set of covariates for a given sample, and Z is a particular treatment assigned.

Once the propensity score for each sample in a study is computed, several propensity-based smoothing techniques could be applied[6]. These include the following: (a) *Inverse Propensity of Treatment Weighting (IPTW)* that uses weights based on propensity scores to generate synthetic samples such that the distribution of covariates is independent of the treatment; (b) *Propensity Score Matching* that matched sets of treated and untreated subjects who share a similar value of the propensity score; (c) *Stratification on the Propensity Score* that stratifies subjects into mutually exclusive subsets (e.g. quintiles) based on their propensity scores.; and (d) *Covariate Adjustment Using the Propensity Score* in which the outcome variable is regressed on an indicator variable denoting treatment status and the propensity score.

Among these approaches, we selected IPTW due to its effectiveness compared to the others as it does not require matching or stratification that might result in discarding unmatched samples or suboptimal stratification. IPTW uses weights based on propensity scores to generate synthetic samples such that the distribution of covariates is independent of the treatment. These weights include: Average Treatment Effects (w_{ATE}); Stabilized Average Treatment Effects (w_{ATE_stab}) and Average Treatment Effect on the Treated (w_{ATT}), which could be computed as follows:

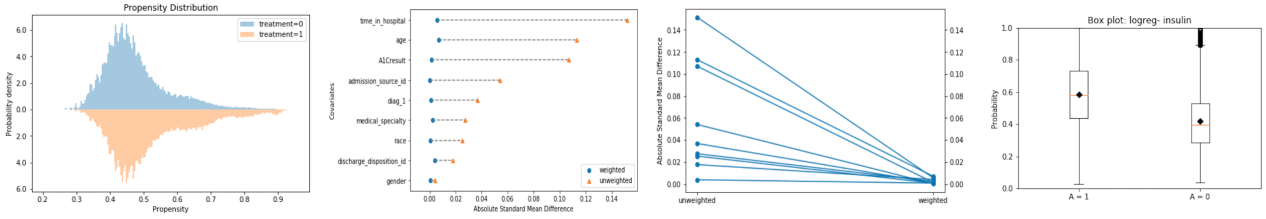
$$w_{ATE} = \frac{Z}{p_s} + \frac{1-Z}{1-p_s}, \qquad w_{ATE\_stab} = \frac{Zp(Z=1)}{p_s} + \frac{1-Z(p(Z=0))}{1-p_s}, \qquad w_{ATT} = Z + p_s\frac{1-Z}{1-p_s}.$$

Balance diagnosis and evaluation of positivity violation

Propensity score-based bias smoothing could be achieved using different binary classification algorithms, such as logistic regression and random forest. Thus, it is important to evaluate the balance of the treatment bias achieved by the propensity model. To this end, existing balance diagnosis methods could be grouped into two: *graphical* and *quantitative* methods.

Graphical methods provide visualizations to qualitatively evaluate the balance between treated and untreated subsets. Examples of graphical balance diagnosis methods include the Propensity Score Distribution, Covariate Balance Love Plot, Covariate Balance Slope Plot, and Covariate Balance Box Plot as shown in Fig. 1. Propensity score distribution presents the probability density of treated and untreated groups across propensity score. Overlapping between

the two density functions suggests balancing where a horizontal deviation of these distributions signals the lack of balancing and hence positivity assumption violations. Covariate Balance Love Plot provides the absolute standard mean difference for each covariate between the treated and untreated group, before and after weighting is applied. A smaller deviation (< 0.1) signals good balancing. The Covariate Balance Slope Plot is another visualization of the absolute mean differences and an alternative to the Covariate Balance Love Plot. Finally, the Covariate Balance Box Plot shows the mean propensity score for the treated and untreated groups, where good balance is depicted from similar mean propensity scores of the two groups.



(a) propensity score distribution **(b)** covariate balance love plot **(c)** covariate balance slope plot **(d)** covariate balance box plot

Figure 1: Examples of graphical methods to assess the balance of observed covariates in treated and comparison groups using propensity scores and inverse probability of treatment weighting.

Quantitative methods provide quantifiable values regarding the balancing by the propensity score-based weights, which could then also help to evaluate positivity assumption violations[7]. Examples of quantitative methods include the *standardised difference* (s_d)[7], the *Kolmogorov-Smirnov test statistic* (k_s)[8], and *overlapping index* (o_i)[9]. Given a covariate profile, x, which becomes x_t for treated group and x_c for the comparison group, these quantitative values could be obtained as follows:

$$s_d(x) = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{s_{xt}^2 + s_{xc}^2}}, \qquad k_s = Max|p_{st}(x) - p_{sc}(x)|, \qquad o_i = \int (min[p_{st}(x), p_{sc}(x)]dx)$$

where $\bar{x}_t$ and $\bar{x}_c$ represent mean of treated (x_t) and comparison (x_c) groups, respectively; s_{xt} and s_{xc} represent the standard deviation of x_t and x_c, respectively. Similarly, $p_{st}x$ and $p_{sc}x$ represent the propensity score for x_t and x_c, respectively.

Unified Performance Index

The quantitative approaches used to diagnose both the imbalance of the treated and comparison groups as well as positivity assumption violations are often used separately, yet it is important to have a unified evaluation framework that takes into consideration the different quantitative evaluation techniques. To this end, we introduced the Unified Performance Index (UPI) that takes into account the overlapping index, mean stabilized ATE weights, KS statistic, and mean of the weighted standardized mean difference. Better balancing and lesser positivity assumption violation diagnosis correspond to a higher UPI score. The UPI maximizes the overlapping index o_i, minimizes the absolute mean standardized differences for covariates s_d, minimizes the KS test statistic k_s and minimizes the deviation of mean stabilized ATE weights w_{ATE_stab} from unit value. Thus, UPI is formulated to reflect these proportionality characteristics as follows:

$$UPI = \frac{o_i}{s_d + k_s + abs(1 - w_{ATE\_stab})}$$

Here, o_i has a domain of $[0, 1]$ such that $o_i = 0$ indicates that the propensity score distributions among the treated and untreated groups are completely separated. Conversely $o_i = 1$ indicates that the two distributions the same. The s_d has a domain of $[0, 1]$ such that $s_d = 0$ implies that there is no difference in the mean or prevalence of variables between the treated and untreated groups. k_s also has a domain of $[0, 1]$ such that $k_s = 0$ if the cumulative distributions of

propensity scores among the treated and untreated groups are identical, and $k_s = 1$ if the distributions are completely distinct. Lastly, w_{ATE_stab} has a domain of $[-\infty, +\infty]$ such that mean stabilized weights that are further from one are indicative of higher degrees of the violation of the positivity assumption.

Subset scanning for anomalous pattern detection

In subset scanning literature, the pattern detection problem can be framed as a search over all subsets in a multidimensional array that spans any combination of feature values to identify the most anomalous subset, i.e. the subset with the most evidence of divergence from expected behavior. Scanning is achieved by maximizing a scoring function, $F(S)$, over all subsets to identify the highest-scoring subset, $S^* = \arg\max_S F(S)$. This approach can, therefore, be used to reveal hidden anomalous subsets that may not be obvious when inspecting individual features manually. The scoring functions exploit a mathematical property, the Linear Time Subset Scanning (LTSS) property[2], which proves that the values of a given discrete/discretized feature can be ordered optimally using a priority function such that scanning is done without requiring an exhaustive search and is guaranteed to be completed in linear time ($O(n)$) rather than in exponential time ($O(2^n)$).

As previously highlighted, several subset scanning techniques have been developed. In this study, we specifically extend the Bias-Scan methodology. The goal of Bias-Scan is to discover the subpopulation with the most divergence between the true outcomes and the predicted probabilities of a binary classifier[4]. Given tabular data with discrete/discretized covariates, a binary outcome, y_i, and predictions generated by a binary classifier, $\hat{p_i}$, Bias-Scan maximizes a Bernoulli likelihood ratio scoring statistic, $score_{bias}(S)$, that quantifies bias in a given subgroup. The algorithm identifies the subgroup that has the most evidence of having the expected odds differing from the predicted odds. Here, the null hypothesis is that prediction odds are correct across all subgroups, $H_0 : odds(y_i) = \frac{\hat{p_i}}{1-\hat{p_i}}$; while the alternative hypothesis assumes a constant multiplicative increase in the prediction odds for some given subgroup, $H_1 : odds(y_i) = q\frac{\hat{p_i}}{1-\hat{p_i}}$ where $q > 1$. The scoring function in Bias-Scan is:

$$score_{bias}(S) = \max_q log(q) \sum_{i \in S} y_i - \sum_{i \in S} log(1 - \hat{p_i} + q\hat{p_i})$$

Consequently, subsets in which records have larger numbers of $y_i = 1$ but smaller corresponding p_i will have higher scores. To detect the anomalous subgroups, Bias-Scan uses the Multi-Dimensional Subset Scanning (MDSS) algorithm[10]. Given a multi-dimensional array of discrete/discretized features, MDSS optimizes Bias-Scan's likelihood ratio statistic over all subsets of values of each feature conditioned on the current subset of all other features in the multi-dimensional array. To do so efficiently and exactly, MDSS satisfies the LTSS property[2] with a priority function computed as the ratio of the observed odds and the expected odds. This priority function[3] ranks the values of a given feature and then select the highest-scoring subset as the subset consisting of the "top-k" priority values for some $k \in [1, \ldots, J]$. MDSS iterates over all features in the multidimensional array until convergence to a local maximum is found. The global maximum is subsequently optimized using multiple random restarts.

Anomalous subgroup detection for binary outcomes

Our study combines propensity score techniques with the Bias-Scan to discover anomalous patterns associated with medication classes used in diabetes care. Here, we specifically proposed three algorithms: Conditional Automated Stratification Scan (CASS), Matched Conditional Automated Stratification Scan (mCASS), and Weighted Conditional Automated Stratification Scan (wCASS). These algorithms analyze tabular data with discrete/discretized covariate profiles X, a single binary treatment $Z \in \{0, 1\}$, and a single binary outcome $Y \in \{0, 1\}$. The key steps in the algorithms are described in Table 1 and discussed in detail below.

1. Conditional Automated Stratification Scan (CASS)

The CASS algorithm represents our simplest extension of the Bias-Scan methodology to enable the estimation of heterogeneous effects of interventions on a binary outcome. In CASS, the anomalousness of any given subpopulation S of treated subjects is quantified as $E[Y_i(1) = 1 | X_i \in S] < E[Y_i(0) = 1]$ for under-risked subpopulations (i.e.

Table 1: Algorithms for anomalous subgroup detection for binary outcomes

Conditional Automated Stratification Scan (CASS)	Matched Conditional Automated Stratification Scan (mCASS)	Weighted Conditional Automated Stratification Scan (wCASS)
1. Get treatment group data $Data\|_{Z=1}$ 2. Get comparison group data $Data\|_{Z=0}$ 3. For each subject i in $Data\|_{Z=1}$, estimate the counterfactual outcome as mean outcome in $Data\|_{Z=0}$, i.e. $\hat{Y}_i = E[Y(0) = 1]$ 4. Apply Bias-Scan(X, Y, $\hat{Y}$) using $Data\|_{Z=1}$ 5. Estimate statistical significance of identified subpopulation using boot-strapped randomization testing	1. Get the treatment's propensity scores from the best propensity score model 2. Get the treatment's logit of the propensity score 3. Get treatment group data $Data\|_{Z=1}$ 4. Get comparison group data $Data\|_{Z=0}$ 5. For each subject i in the $Data\|_{Z=1}$ (a) Identify nearest neighbors in $Data\|_{Z=0}$ as those within 0.2SD of the logit of the propensity score (b) Estimate the counterfactual outcome $\hat{Y}_i$, as the average outcome among the identified nearest neighbors 6. Apply Bias-Scan(X, Y, $\hat{Y}$) using $Data\|_{Z=1}$ 7. Estimate statistical significance of identified subpopulation using boot-strapped randomization testing	1. Get the treatment's propensity scores from the best propensity score model 2. Compute the average treatment effect on the treated (ATT) weights (w_{ATT}) 3. Get treatment group data $Data\|_{Z=1}$ 4. Get comparison group data $Data\|_{Z=0}$ 5. For each subject i in $Data\|_{Z=1}$, estimate the counterfactual outcome $\hat{Y}_i$ as the ATT-weighted mean expected outcome in $Data\|_{Z=0}$ 6. Apply Bias-Scan(X, Y, $\hat{Y}$) using $Data\|_{Z=1}$ 7. Estimate statistical significance of identified subpopulation using boot-strapped randomization testing

lower than expected outcomes), and $E[Y_i(1) = 1|X_i \in S] > E[Y_i(0) = 1]$ for over-risked subpopulations (i.e. higher than expected outcomes). Here, $Y_i(1)$ denotes the occurrence the an outcome for a treated subject, $E[Y_i(1) = 1|X_i \in S] = \frac{1}{N_1}\sum_{i=1}^{N_1} i \in S\, Y_i$ is the probability of the outcome in the subpopulation S, and $E[Y_i(0) = 1] = \frac{1}{N_0}\sum_{i=1}^{N_0} Y_i$ is the marginal probability of occurrence of the outcome among the comparison group subjects. Therefore, CASS assumes that the counterfactual outcome for each treated subject is $E[Y(0) = 1]$ and that the average unit-level causal effect of the treatment for each treated subject is $Y(1)_i - E[Y(0) = 1]$. CASS then searches for a specific most anomalous subpopulation S^* as the subpopulation in which the probability of the outcome conditioned on belonging to this subpopulation has the most evidence of being divergent from the marginal probability of the outcome among all the comparison group subjects. Procedurally, the key steps in CASS are described in Table 1.

2. Matched Conditional Automated Stratification Scan (mCASS)

The mCASS algorithm combines propensity score matching with Bias-Scan to discover heterogeneous treatment effects across subpopulations. In mCASS, the counterfactual outcome for each treated subject is determined by propensity score matching between pairs of treated and comparison group subjects who have similar propensity scores. Consequently, in mCASS, the average treatment effect is estimated as the average of the differences between the actual versus counterfactual outcome within each pair. While several propensity score matching techniques can be used in mCASS, we specifically use the nearest neighbor caliper matching[6]. In this approach, matching is done on the logit of the propensity scores using calipers of width 0.2 standard deviation of the logit of the propensity score as a threshold for matching[6]. The key steps in mCASS are described in Table 1.

3. Weighted Conditional Automated Stratification Scan (wCASS)

The wCASS algorithm uses the IPTW derived from propensity scores. Specifically we use the previously described average treatment effect on the treated (ATT) weights, w_{ATT}. When using wCASS, the anomalousness of a subpopulation S is quantified as $E[w_iY_i(1)|X_i \in S] < E[w_iY_i(0)]$ for under-risked subpopulations and as $E[Y_i(1)|X_i \in S] > E[Y_i(0)]$ for over-risked subpopulations. Here, $w_iY_i(1)$ denotes the weighted outcome for a treated subject, $E[w_iY_i(1)|X_i \in S] = \frac{1}{N_1}\sum_{i=1}^{N_1} i \in S\, w_iY_i$ is the weighted probability of the outcome in the subpopulation S, and

$E[w_i Y_i(0)] = \frac{1}{N_0} \sum_{i=1}^{N_0} w_i Y_i$ is the weighted marginal probability of occurrence of the outcome among the comparison group subjects. Procedurally, wCASS works as described in Table 1.

Randomization Testing

All the 3 algorithms, CASS, mCASS, and wCASS use randomization testing to estimate the statistical significance of the detected anomalous subgroups. To do so, we draw 1000 bootstrapped random subpopulations and compute the subset score for each subpopulation drawn, and then compare each score to the score for the detected anomalous subpopulation. We compute the empirical p-value as $\frac{(r+1)}{(n+1)}$ where r is the number of scores greater than or equal to that actual score and n is the total randomization scores for bootstrapped samples.

Experiment and Results

To validate the algorithms we developed, we used a case study whose goal was to detect the anomalous patterns associated with the impact of medications on 30-day hospital readmission in diabetes. Figure 2 illustrates the pipeline used in the study.

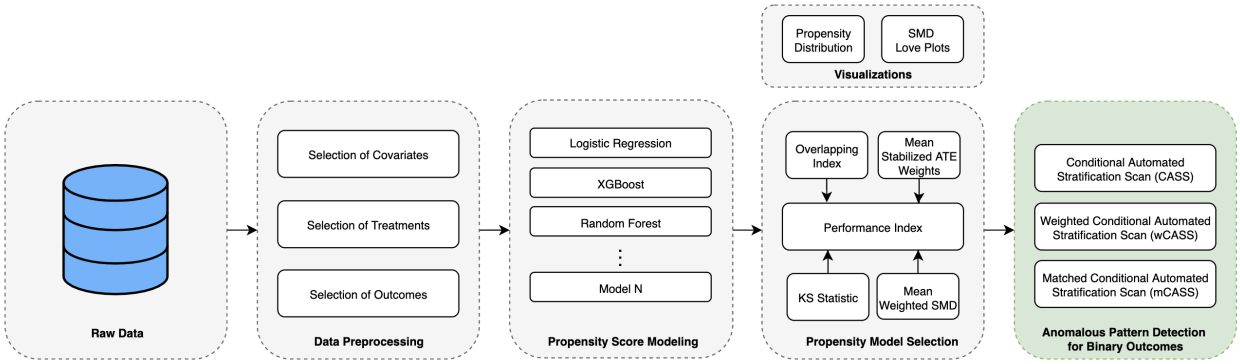


Figure 2: Pipeline for anomalous pattern detection associated with binary Outcomes in healthcare

The Diabetes Cohort and Dataset

We used a de-identified diabetes dataset extracted from the Health Facts database (Cerner Corporation, Kansas City, MO). The study population consisted of 69,990 patients with inclusion criteria defined as follows: (1) having an initial inpatient encounter (a hospitalization), (2) having a diagnosis of diabetes made at the initial encounter, (3) the length of stay at least 1 day and at most 14 days, (4) laboratory tests were performed during the encounter, and (5) medications were administered during the encounter[11]. The data preprocessing is described in more detail by Strack et al.[11] who also made it publicly available as supplementary material accessible at `http://dx.doi.org/10.1155/2014/781670`. We conducted additional preprocessing to generate a final dataset that consisted of a binary outcome (Readmission within 30 days), 10 discrete/discretized covariates (race, gender, age group, admission type, discharge disposition, admission source, primary diagnosis, secondary diagnosis, tertiary diagnosis, and HbA1C), and 4 intervention drug classes (Biguanides, Insulin, Sulfonylureas, and Thiazolidinediones). These preprocessing steps included dropping features with a large proportion of missing values, remapping categorical features e.g., age groups and race, mapping specific ICD9-CM codes to more general ICD9-CM codes, mapping treatment interventions to binary pharmacological classes, and mapping the readmission outcome to a binary variable. Table 2 describes the features and the distribution of the feature values in the final dataset.

Propensity Score Modeling

For each of the 4 treatment classes (biguanides, insulins, sulfonylureas, and thiazolidinediones) in the final dataset, we trained 3 propensity score models for predicting the likelihood of a patient subject receiving the treatment given

Table 2: Distribution of feature values in the final dataset

Feature	Feature Value	Overall (%)	Readmitted (%)	Not Readmitted (%)
n		69990	6285 (9.0)	63705 (91.0%)
Age	0-29	1808 (2.6)	112 (1.8)	1696 (2.7)
	30-59	21871 (31.2)	1574 (25.0)	20297 (31.9)
	60-99	46311 (66.2)	4599 (73.2)	41712 (65.5)
Gender	Female	37239 (53.2)	3365 (53.5)	33874 (53.2)
	Male	32751 (46.8)	2920 (46.5)	29831 (46.8)
Race	AfricanAmerican	12627 (18.0)	1095 (17.4)	11532 (18.1)
	Caucasian	52305 (74.7)	4807 (76.5)	47498 (74.6)
	Missing	1919 (2.7)	141 (2.2)	1778 (2.8)
	Other	3139 (4.5)	242 (3.9)	2897 (4.5)
Primary diagnosis	Circulatory	21390 (30.6)	2070 (32.9)	19320 (30.3)
	Diabetes	5748 (8.2)	524 (8.3)	5224 (8.2)
	Digestive	6488 (9.3)	520 (8.3)	5968 (9.4)
	Genitourinary	3441 (4.9)	309 (4.9)	3132 (4.9)
	Injury	4696 (6.7)	507 (8.1)	4189 (6.6)
	Musculoskeletal	4064 (5.8)	341 (5.4)	3723 (5.8)
	Neoplasm	2538 (3.6)	230 (3.7)	2308 (3.6)
	Respiratory	9491 (13.6)	693 (11.0)	8798 (13.8)
	Other	12134 (17.3)	1091 (17.4)	11043 (17.3)
Specialty of admitting physician	Cardiology	4208 (6.0)	303 (4.8)	3905 (6.1)
	Family/GeneralPractice	4978 (7.1)	485 (7.7)	4493 (7.1)
	InternalMedicine	10641 (15.2)	1039 (16.5)	9602 (15.1)
	Surgery	3751 (5.4)	297 (4.7)	3454 (5.4)
	Other	12758 (18.2)	1051 (16.7)	11707 (18.4)
	Missing/Unknown	33654 (48.1)	3110 (49.5)	30544 (47.9)
Admission source	Emergency Room	37273 (53.3)	3452 (54.9)	33821 (53.1)
	Referral	22793 (32.6)	1973 (31.4)	20820 (32.7)
	Other	9924 (14.2)	860 (13.7)	9064 (14.2)
Discharge disposition	Home	44322 (63.3)	3079 (49.0)	41243 (64.7)
	Other	25668 (36.7)	3206 (51.0)	22462 (35.3)
A1Cresult	Normal	3741 (5.3)	323 (5.1)	3418 (5.4)
	7 to 8	2866 (4.1)	247 (3.9)	2619 (4.1)
	>8	6239 (8.9)	509 (8.1)	5730 (9.0)
	No Test	57144 (81.6)	5206 (82.8)	51938 (81.5)
Time in hospital, mean (SD)		4.3 (2.9)	4.8 (3.1)	4.2 (2.9)
Time in Hospital (Categorical)	<=3 days	35146 (50.2)	2647 (42.1)	32499 (51.0)
	>3 days	34844 (49.8)	3638 (57.9)	31206 (49.0)
Biguanides	0	54628 (78.1)	5009 (79.7)	49619 (77.9)
	1	15362 (21.9)	1276 (20.3)	14086 (22.1)
Insulins	0	34268 (49.0)	2843 (45.2)	31425 (49.3)
	1	35722 (51.0)	3442 (54.8)	32280 (50.7)
Sulfonylureas	0	49228 (70.3)	4336 (69.0)	44892 (70.5)
	1	20762 (29.7)	1949 (31.0)	18813 (29.5)
Thiazolidinedione	0	60092 (85.9)	5416 (86.2)	54676 (85.8)
	1	9898 (14.1)	869 (13.8)	9029 (14.2)

his/her covariate profile. The predictor variables in each model consisted of race, gender, age, discharge disposition, admission source, time in hospital (continuous), primary diagnosis, hbA1C result, and the medical specialty of the admitting physician. Each response variable was a binary treatment class. These variables are described in Table 2. The trained propensity score models included a logistic regression model, a gradient boosting decision tree model (XGBoost), and a random forest model. Table 3 describes the modeling performance results. We note that, based on the UPI score, XGBoost consistently performed well across the treatment classes considered. Logistic regression also performs relatively well, while the random forest model performed worst across all the treatments despite having relatively better Area Under Curve results for training and test sets. The overlapping indexes for the propensity scores generated from the random forest model were lower than those of the other two models. These findings confirm a known observation that the best propensity score models are not necessarily those that are good at prediction, but those that provide better overlapping between the propensity scores of the treated versus untreated subjects.

Characteristics of the most anomalous subpopulations discovered

Table 4 shows the anomalous pattern detection results for the different treatments analyzed and algorithms used. Each algorithm was able to identify heterogeneous treatment effects by discovering the most anomalous subgroup

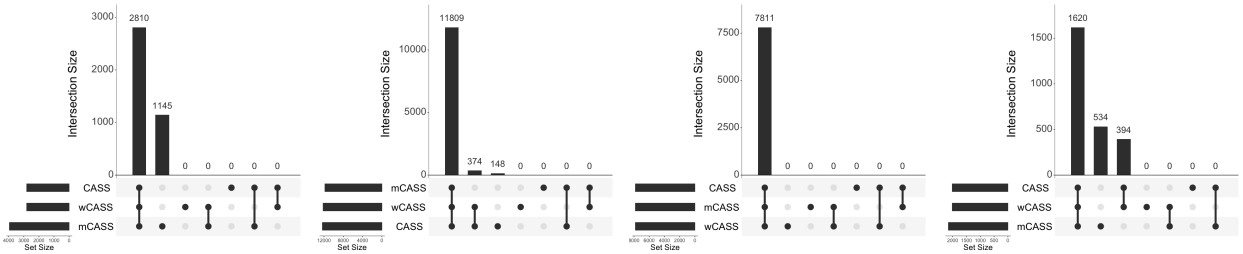
Table 3: Propensity Score Modeling Performance Results

Treatment	Model	Prob of Treatment	Train AUC	Test AUC	Mean Std Diff Unweighted	Mean Std Diff Weighted	KS Test Statistic	KS Test p-value	Overlapping Index	Mean ATE Stab Weight	UPI Score
	XGBoost	0.219	0.626	0.616	0.218	0.015	0.175	<0.001	0.707	0.998	3.692
Biguanides	Logistic Regression	0.219	0.606	0.607	0.218	0.014	0.152	<0.001	0.743	1.319	1.529
	Random Forest	0.219	0.691	0.615	0.218	0.039	0.24	<0.001	0.616	1.271	1.12
	Logistic Regression	0.51	0.624	0.62	2.533	0.014	0.172	<0.001	0.708	1.001	3.787
Insulins	XGBoost	0.51	0.663	0.642	2.533	0.015	0.221	<0.001	0.64	0.995	2.65
	Random Forest	0.51	0.679	0.639	2.533	0.05	0.241	<0.001	0.613	0.974	1.938
	XGBoost	0.297	0.613	0.607	1.099	0.013	0.158	<0.001	0.73	0.998	4.218
Sulfonylureas	Logistic Regression	0.297	0.605	0.606	1.099	0.008	0.15	<0.001	0.741	1.166	2.284
	Random Forest	0.297	0.67	0.601	1.099	0.035	0.223	<0.001	0.637	1.137	1.615
	XGBoost	0.141	0.598	0.57	0.362	0.027	0.128	<0.001	0.778	0.998	4.962
Thiazolidinediones	Logistic Regression	0.141	0.578	0.571	0.362	0.01	0.105	<0.001	0.815	1.514	1.295
	Random Forest	0.141	0.689	0.569	0.362	0.035	0.233	<0.001	0.625	1.454	0.866

conditioned on patients receiving a given treatment. This heterogeneity is described in terms of risk difference, relative risk, and odds ratio in the overall population of treated subjects versus the most anomalous subpopulation identified. We observe that for each treatment, the discovered anomalous subpopulations are relatively similar in sizes across the CASS, mCASS, and wCASS algorithms. We also observe that for each treatment, the algorithms resulted in similar measures of effect across the identified anomalous subpopulations. This observation could be explained by the high degrees of overlaps in the subpopulations identified by the different algorithms as illustrated in Figure 1.

Table 4: Anomalous Pattern Detection Results

		Anomalous Subpopulation			Risk Difference		Relative Risk		Odds Ratio	
Treatment	**Algorithm**	**Size**	**Score**	**P-Value**	**Population**	**Subpopulation**	**Population**	**Subpopulation**	**Population**	**Subpopulation**
	CASS	2810 (18.3%)	39.1	0.004	0	0.1	0.9	1.6	0.9	1.7
Biguanides	mCASS	3955 (25.7%)	41.1	0.003	0	0	0.9	1.5	0.9	1.6
	wCASS	2810 (18.3%)	44	0.003	0	0.1	0.9	1.6	0.9	1.7
	CASS	12331 (34.5%)	180.5	0.002	0	0.1	1.2	1.6	1.2	1.7
Insulins	mCASS	11809 (33.1%)	169.5	0.002	0	0.1	1.1	1.6	1.2	1.7
	wCASS	12183 (34.1%)	158.8	0.002	0	0	1.1	1.6	1.1	1.7
	CASS	7811 (37.6%)	79.3	0.002	0	0	1.1	1.5	1.1	1.6
Sulfonylureas	mCASS	7811 (37.6%)	61.7	0.002	0	0	1	1.4	1	1.5
	wCASS	7811 (37.6%)	65.7	0.002	0	0	1	1.4	1	1.5
	CASS	2014 (20.3%)	37.9	0.003	0	0.1	1	1.7	1	1.8
Thiazolidinediones	mCASS	2154 (21.8%)	34.5	0.002	0	0.1	1	1.6	1	1.7
	wCASS	2014 (20.3%)	39.1	0.003	0	0.1	1	1.7	1	1.8



(a) Biguanides **(b)** Insulins **(c)** Sulfonylureas **(d)** Thiazolidinediones

Figure 3: Overlaps between anomalous subpopulations identified by CASS, mCASS, and wCASS

Interestingly, the subgroups with the largest measures of effect pertained to the use of thiazolidinediones, suggesting a stronger heterogeneous treatment effect of this class of drugs on 30-day hospital readmission. By way of example, the subsets discovered by CASS and wCASS for thiazolidinediones were identical and suggests that patients who use thiazolidinediones and are aged 60 years or older; and are Caucasian or African American or have their race information missing; and have a primary diagnosis that is not musculoskeletal; and were admitted by specialists who are not cardiologists; and had HbA1C >8 or had no HbA1C test conducted at the time of admission; and were discharged to destinations other than their homes, were 1.8 times more likely to be readmitted within 30 days than the average non-treated population. This subpopulation had a 30-day readmission rate that differed the most from the

expected 30-day readmission rate determined as the global mean among untreated subjects.

Discussion

This study aimed at developing and demonstrating the application of techniques for assessing the causal heterogeneous effects of binary interventions on binary outcomes in observational health data such as electronic health records. To this end, the study proposed a unified performance index (UPI) for choosing the best propensity score model among multiple propensity score models and describes how algorithms from the causal inference and anomalous pattern detection literature could be leveraged to discover anomalous patterns of care. Furthermore, the demonstrates how the developed algorithms can be used to detect the heterogeneous treatment effects captured in electronic health records.

We discovered highly similar subpopulations among which the use of anti-diabetic medication classes (biguanides, insulins, sulfonylureas, and thiazolidinediones) was associated with an increased likelihood of being readmitted within 30 days after the index inpatient admission. Interestingly, thiazolidinedione therapy has previously been associated with a higher risk of hospital readmissions on average[12–14]. The findings from our study suggest that certain subpopulations may be differentially affected by this class of drugs as well as other commonly used classes of anti-diabetic medication. However, we take cognizance of the fact that our approach should be viewed as a method for generating hypotheses about the specific subpopulations that are most likely to be impacted by the interventions. How such heterogeneity occurs or is realized is beyond the scope of the current approach and further investigations are warranted to confirm the generated hypotheses.

Current literature has primarily focused on studying the average treatment effect of interventions on binary outcomes[6], or on studying heterogeneous treatment effects of single interventions in clinical trials[3]. These approaches are, however, done separately. To the best of our knowledge, our study is the first to leverage and extend both causal inference and subset scanning techniques to study the effect of interventions on binary health outcomes. The techniques we have developed can provide researchers, care providers, and other stakeholders with the ability to identify positive or adverse clinical practices and subsequently institute better care delivery plans based on the identified insights. They can also be used as a basis for risk analysis and recommendation of follow-up interventions to improve care experiences and outcomes for individual patients, especially in differential service delivery and targeted intervention planning settings. Furthermore, they can enable payers to identify drivers of poor outcomes and unnecessary costs across patient subpopulations.

Whereas we took the necessary steps to ensure robustness in our study, several limitations can be observed. First, as would be expected, feature selection and engineering can affect propensity scoring modeling and the anomalous pattern detection, subsequently bias findings. We minimized this by testing our approach on a diabetes dataset validated by Strack et al.[11] while maintaining the features and feature values used in the aforementioned study. Second, the UPI score is currently unbounded and ranges from zero to infinity with higher scores implying better performance. However, a monotonic function that maps the UPI score to [0,1] can be applied without loss of generality. As part of our future work, we intend to refine, test, and compare different formulations of the UPI across different propensity score models trained on several publicly available datasets. Third, the subset scanning approach applied in this study can be misconstrued as conducting multiple hypotheses testings. However, we consider this and use parametric bootstrapped randomization tests to determine the statistical validity of anomalous subpopulations discovered by our algorithms. Fourth, the results of the subset scanning process can be complex and difficult to interpret even for persons with domain expertise. As part of our future work, we intend to incorporate into our pipeline a penalized version of Bias-Scan that uses a penalty function to minimizes the complexity and maximize the size of the identified subpopulations[15]. Lastly, our current approach can only analyze binary interventions and binary outcomes. We, however, acknowledge that there are different healthcare outcomes that can be binary (e.g. mortality), discrete (e.g. number of days spent in hospitals), continuous (e.g. blood glucose levels). At the same time, the intervention space in healthcare is often complex and can range from single interventions given only once (e.g. single dose medications or vaccinations), to multiple interventions used simultaneously (e.g. drug combinations in regimens), to sequential interventions (e.g. temporally dependent drug regimens, clinical pathways). As part of our future work, we plan to develop techniques for discovering anomalous patterns associated with these different types of interventions and outcomes in healthcare data. Additionally, we intend to compare our approach to state-of-the-art subgroup analysis techniques and to generate meaningful clinical insights, perspectives, and interpretations of the discovered anomalous subpopulation through

counterfactual analyses of modifiable risk factors that could serve as a baseline for targeted intervention planning.

Conclusion

We studied 30-day readmission in diabetes using methods that combine techniques from causal inference and anomalous pattern detection literature to study the heterogeneous effects of treatments. This study shows that a unified performance index can be used to select the best propensity score models among multiple models. It also shows that techniques such as propensity score matching and inverse probability of treatment weighting could be leveraged to study the impact of binary treatment on binary outcomes at the subpopulation-level and in a disciplined statistical approach. Furthermore, the study shows that for a given treatment, the studied algorithms result in the identification of subpopulations that are highly similar in terms of common covariate characteristics. Lastly, the study demonstrates that for certain subpopulations, the likelihood of 30-day hospital readmission among index diabetes encounter may be differentially impacted by the use of some anti-diabetic medication classes and that these differential effects may not be discernible at the overall populations. Future work includes delineating the merits and demerits of our algorithms, penalizing complexity while maximizing subset sizes for easier interpretability, and generalizing the approaches for application across disparate intervention and outcome data types.

References

[1] Daniel J Rubin. Correction to: hospital readmission of patients with diabetes. *Current diabetes reports*, 18(4):21, 2018.

[2] Daniel B Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):337–360, 2012.

[3] Edward McFowland III, Sriram Somanchi, and Daniel B Neill. Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection. *arXiv preprint arXiv:1803.09159*, 2018.

[4] Zhe Zhang and Daniel B Neill. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*, 2016.

[5] Edward Somanchi, Sriram McFowland III and Daniel B Neill. Detecting anomalous patterns of care using health insurance claims. *Presented at Conference on Information Systems and Technology*, 2017.

[6] Peter C Austin and Elizabeth A Stuart. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research*, 26(6):2505–2525, 2017.

[7] Peter C Austin and Elizabeth A Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

[8] Jasjeet S Sekhon. Alternative balance metrics for bias reduction in matching methods for causal inference. *Survey Research Center, University of California, Berkeley*, 2007.

[9] Massimiliano Pastore and Antonio Calcagnì. Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in psychology*, 10:1089, 2019.

[10] Daniel B Neill, Edward McFowland III, and Huanian Zheng. Fast subset scan for multivariate event detection. *Statistics in medicine*, 32(13):2185–2208, 2013.

[11] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[12] Frederick A Masoudi, Silvio E Inzucchi, Yongfei Wang, Edward P Havranek, JoAnne M Foody, and Harlan M Krumholz. Thiazolidinediones, metformin, and outcomes in older patients with diabetes and heart failure: an observational study. *Circulation*, 111(5):583–590, 2005.

[13] Fei-Yuan Hsiao, Yi-Wen Tsai, Yu-Wen Wen, Pei-Fen Chen, Hao-Yu Chou, Chen-Huan Chen, Ken N Kuo, and Weng-Foung Huang. Relationship between cumulative dose of thiazolidinediones and clinical outcomes in type 2 diabetic patients with history of heart failure: a population-based cohort study in taiwan. *Pharmacoepidemiology and drug safety*, 19(8):786–791, 2010.

[14] Silvio E Inzucchi, Frederick A Masoudi, Yongfei Wang, Mikhail Kosiborod, Joanne M Foody, John F Setaro, Edward P Havranek, and Harlan M Krumholz. Insulin-sensitizing antihyperglycemic drugs and mortality after acute myocardial infarction: insights from the national heart care project. *Diabetes Care*, 28(7):1680–1689, 2005.

[15] Skyler Speakman, Sriram Somanchi, Edward McFowland III, and Daniel B Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 25(2):382–404, 2016.

Comparison of Ease of Use and Comfort in Fitness Trackers for Participants Impaired by Parkinson's Disease: An exploratory study

Jay Patel, BDS, MS, PhD [1, 2], Patrick Lai, MPH [1], Doug Dormer, BGS [1], Rakesh Gullapelli, MHI, MPharm [1, 3], Huanmei Wu, PhD [1, 2], Josette J. Jones, RN, PhD [1]

1: Department of Bio-Health Informatics, School of Informatics and Computing, Indiana University Purdue University Indianapolis

2: Department of Health Services Administration and Policy, College of Public Health, Temple University

3: Center for Outcomes Research, Houston Methodist

Abstract

Parkinson's disease (PD) is an incurable, fatal neurodegenerative disease, and only available treatment is to minimize symptoms. Anecdotal evidence suggests whole body workout can help to reduce PD severity; however, it is challenging to quantify its effect on PD. The increased availability of fitness trackers can help in quantifying the effect of whole- body workout on PD. Before using any over the counter fitness tracker, we must study the ease of use of the fitness trackers in PD patients. We interviewed 32 PD patients with six over the counter fitness trackers and determined their perceptions and attitude towards the fitness trackers. Although none of the fitness trackers received perfect scores for ease of use or comfort due to the presence of tremors, two trackers performed significantly better than the others. Further study is warranted to understand the potential for fitness trackers to be used by PD patients.

Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disorder affecting 630,000 people in the US [1]. PD patients experience motor and non-motor symptoms such as stiff muscles, difficulty in standing, problems with coordination, loss of balance, slow bodily movements, and tremors, which lead to poor quality of life [2]. Currently, there is no approved therapy to prevent, delay, or reverse the progress of the disease [3]. The most common therapies today are medications to control the symptoms, but they do not prevent or slow the progression of the disease [4, 5]. In recent years, alternative therapies involving activity-based approaches such as whole-body workouts, forms of rigorous physical exercises such as dancing and non-contact boxing, have been shown to have a positive effect on patients with PD [6-9]. However, it is unknown how and to what extent whole-body workouts operate to slow or stop PD progression. One of the biggest challenges to conducting research in this domain could be the difficulty in recruiting patients (small sample size), the progress of the condition, and the effect of interventions [10-12]. Both clinicians and researchers typically use biophysical tests such as an electroencephalogram (EEG) to monitor disease severity and measure changes in motor and cognitive function [13-16]. These tests can only be administered by qualified technicians in a clinical facility, which is time-consuming, tedious, expensive, and labor-intensive [17]. The lack of a low cost, easy to use, continuously monitoring, standards-based means of measuring the motor function status, and progression of PD is a major hindrance to both improved management for PD patients and a constraint on research [18,19].

In the last decade, the use of fitness trackers/sensors such as Fitbit, Apple watch, and Jawbone by the general public in the US has become very common [20,21]. Each of these trackers/sensors contains an accelerometer that can record variations and intensity of movement. If this motion data can be interpreted to express the progression of PD and the effect of interventions, then these trackers/sensors may be useful and cost-effective for the large-scale monitoring of patients' exercise intensity and for quantifying the effect of exercise on PD progression. However, it's unclear the extent to which these trackers/sensors will be adopted and used continuously by PD patients due to the presence of motor symptoms such as tremors, which may make them uncomfortable or difficult to use [22,23]. From here on, we will refer "fitness trackers/sensors" as "trackers".

Based on our literature review, most of the existing studies focusing on the ease of use of over-the-counter fitness trackers are in people who do not have PD [24-28]. For instance, Ridgers (2018) performed a study on the usability and acceptability of fitness tracker use among Australian young adolescents and found the Fitbit Flex was the easiest to use activity tracker for tracking the activities [29]. Similarly, Chaparro et al. investigated the usability of six activity

trackers among 19 individuals and found that Fitbit and Garmin were the most popular trackers among the participants because of its lightweight and digital display [30]. While Rasche et al. studied the age-related usability of an activity tracker, authors found that the younger and older age groups used the activity tracker without difficulty independently without any specific training [31]. From these studies, we can conclude that there is some evidence showing that certain trackers work better than others for some users in specific circumstances; however, there is no study demonstrated the ease of use of over the counter trackers in PD patients.

The long-term goal of our research is to identify the best fitness trackers that will be highly adapted by the PD patients and to develop algorithms that allow researchers, clinicians, and patients to monitor activities, including whole body workout. Our short-term goal is to compare and evaluate the ease of wearability of several popular models of fitness trackers for PD patients in the presence of motor symptoms. Because of the limitations in dexterity associated with PD, limited our scope to basic functions such as putting on, taking off, and simply wearing these specific fitness trackers. We consider these attributes to be gating important factors before considering ease of use in the more complex device functions (industrial design of the device, and ergonomic aspects). The results of this study would determine the ease of use, attitude, and perception of six popular over the counter fitness trackers in patients impaired with PD and answer if over the counter fitness trackers could be used to quantify the effect of whole-body workouts in PD patients.

Methods

First, we identified six popular, commercially available trackers for which the manufacturer's marketing documentation indicated they had functions and features valuable to track activities in patients with PD. Second, we recruited 32 PD patients onsite (from here on referred as participants) with various levels of PD severity through Rock Steady Boxing gymnasium in Indianapolis, Indiana. Rock Steady Boxing is a nonprofit organization that offers non-contact boxing-based fitness activities (a form of whole-body workout) for patients with PD. We asked these participants to perform basic functions such as putting on and removing the fitness trackers and performing certain tasks while wearing the selected trackers. Next, we asked them to complete a survey of their perceptions and attitudes towards fitness trackers. Finally, we analyzed the results to determine which of these trackers the PD patients would be most likely to wear and use on a continuous basis.

Commercially available fitness trackers

We began our study by identifying the functions and features that were the most valuable to PD patients such as GPS, activity identification, touch, water resistance, sound, sensor, heart rate, exercise tagging and such as. We selected six trackers that had the most features in line with our feature list (feature list of 27 features). Included trackers in this study are: 1) Tracker A = Fitbit Blaze Large Black (2016 model); 2) Tracker B = Xiaomi Mi Band Black (2014 model; 3) Tracker C = Fitbit Flex Wireless Activity + Sleep Wristband Black (2013 model); 4) Tracker D = Jawbone UP2 Silver (2015 model); 5) Tracker E = Pebble SmartWatch Red (2016 model); 6) Tracker F = Pebble Steel Smartwatch Stainless (2014 model).

Survey design

Since this is a first study to assess the feasibility of wearing commercially available fitness trackers by PD patients, we developed our own survey questions without incorporating survey questions from the previous studies. Two of the team's members who are clinical informaticists developed these survey questions. Detailed information about the survey questions is described in Table 1. To make sure the survey questions are easy to understand by respondents who are unfamiliar with healthcare domain, we recruited two external reviewers to validate the survey. We asked them to evaluate the survey for question flow, sentence organization, grammar, and ease of understanding. The resulting survey scored very high (0.86) on Cohen's Kappa statistical test for agreement among the reviewers [32].

Data collection

We recruited participants from PD patients who are regular members in therapeutic exercise programs at Rock Steady Boxing. We used flyers to recruit participants. The 32 participants recruited for our study suffered from a wide range of disease intensities. None of the study participants had previously used fitness trackers. Each of the 32 participants was invited to work with all six of the fitness trackers. The order of the trackers was changed between participants to reduce the bias that comes with an increased general familiarity with using this type of devices. We verbally asked them to do the tasks like putting the tracker on and off and asked them questions regarding their experience with the trackers. We collected information about the initial impression of trackers, ease of use (the ability of the patient to put on and take off the device without hindrance or difficulty), comfort (see questions 4 to 6),

and overall likelihood (see question 7). We also collected information on how long and how often they have been attending the Rocksteady boxing facility (see questions 9, 10). We asked the questions 9 and 10 for our future study to find out how rigorous physical activity can help reduce PD progression. Due to the presence of tremors, we (all authors) noted down each response on paper and then transferred the information in a computer for further data analysis.

Statistical Analysis

To measure if there was an overall difference in opinion within the six different trackers for Questions 4-7, a nonparametric repeated measures omnibus analysis called the Friedman's test was conducted for each question. Each fitness tracker was compared to the other fitness trackers to determine individual preferences for Questions 4- 7, using a pairwise Wilcoxon Signed-Rank test. A Bonferroni adjustment was applied to the overall statistical significance to prevent the accumulation of Type I error and control for the familywise alpha level when conducting individual pairwise tests for p-values. The formula given to provide the Bonferroni adjusted p-value was $\alpha' = 1-(1-\alpha)1/k$ with α' representing the adjusted critical p-value from the adjustment, α representing the standard Type I error of 0.05, and k representing the number of multiple comparisons made which was a total of 15 comparisons for this study. Any p-value that was below the adjusted critical p-value was considered statistically significant. Summary and descriptive statistics on the time to wear and remove the trackers were performed for Questions 2 and 3, and comparison of these time measurements was analyzed using the Kruskal-Wallis nonparametric ANOVA test. The default level of significance was compared at $\alpha = 0.05$ for all analyses with the adjusted $\alpha = 0.00341$ for Questions 4-7 due to the multiple comparison tests and Bonferroni correction. All statistical analyses were performed using SAS® software version 9.4.

Table 1: Questionnaire to assess the feasibility of wearing fitness trackers in PD patients.

Survey Questions	Survey Response Options
1: What is your age group?	< 35 years, 36-45, 46-55, 56-65, 66-75, 76-85, > 85 years
2: How long did it take for you to put on and wear your device? (seconds)	/Open-ended
3: How long did it take for you to remove your device? (seconds)	Open-ended
4: How easy was it to wear your fitness device?	Very easy, Easy, Moderate, Hard, Very Hard, Other
5: How comfortable did the material feel against your arm?	Comfortable, Moderate, Rogh/Uncomfortable, Other
6: How comfortable is it to carry the device while working out?	Comfortable, Moderate, Rough/Uncomfortable, Other
7: How likely is it that would start to use/continue using this device?	Very likely, Maybe, Never, Other
8: How did Rock Steady Boxing help you?	Open-ended question
9: How long have been attending Rock Steady Boxing?	Open-ended question
10: How often do you come to Rock Steady Boxing?	Open-ended question

Results

Age of the participants

All of our participants were older than 45 and younger than 86 years of age. Three participants (9.4%) were between 46-55 years old, nine participants (28.1%) were between 55-65 years, and two participants (6.3%) were between 75-85 years. The 65-75 age group was the largest cohort in the study, with eighteen participants representing 56.25% of our sample.

Initial impression

Although all 32 participants were given the opportunity to use all six fitness trackers, none of the participants chose to test all six devices. The reasons why participants declined to try a fitness device included: the fitness tracker did not fit *("the band is too short"*); the participants didn't like either the attachment mechanism (such as a buckle), or the look or feel *("this one is 'too heavy' or 'clunky"*). In calculating our scores for each device, those that were selected more often received higher scores than those which the participants declined to try at all to use. Of the six

fitness trackers, 28 people tested tracker E and 29 tested tracker F, compared to only 11 people who chose tracker C and 12 who chose tracker D. Trackers A & B were tested by 22 participants.

Ease of use

We defined ease of use as the ability to put on and take off the device without hindrance or difficulty. We evaluated ease of use by measuring the amount of time in seconds that it took for a subject to both wear and take off a fitness tracker. For each fitness tracker, we calculated the mean and median times to put on or take off the fitness tracker along with the standard deviation, the lowest (shortest) time, and the highest (longest) time. We also noted the number of participants that declined to test each fitness tracker. Table 2 shows the results of the two ease of use questions.

Using the Kruskal-Wallis test, there was an overall statistically significant difference in the distribution of time to put on the fitness trackers ($\chi2(5) = 49.1619$, $p < 0.0001$) and to take them off ($\chi2(5) = 15.8928$, $p < 0.0072$). Tracker F had the lowest (quickest) mean and median times, followed by Tracker E for question 2, putting on the tracker, and Tracker C for question 3, removal. Tracker D had the highest meantime, and Tracker C had the highest median time putting on the fitness tracker. With respect to removal time, Tracker F had the lowest (quickest) mean and median time, while Tracker D had the highest (slowest) mean and median time.

Table 2: Summary statistics of time to put on and take off the fitness tracker.

Question 2: How long did it take for you to put on and wear your device? (seconds)								
Tracker	Samples	Mean	Median	Std Error	Std Dev	Lowest	Highest	Other
A	22	27.2	22.5	2.717	12.7	4	58	10
B	22	31.5	26.0	3.924	18.4	9	74	10
C	11	41.1	51.0	8.940	29.6	9	100	21
D	12	52.3	48.0	7.431	25.7	14	100	20
E	28	21.3	20.0	1.751	9.2	8	45	4
F	29	11.6	8.0	1.496	8.0	3	30	3
Question 3: How long did it take for you to remove your device? (seconds)								
Tracker	Samples	Mean	Median	Std Error	Std Dev	Lowest	Highest	Other
A	22	23.1	15.0	4.572	21.4	5	100	10
B	19	20.7	12.0	5.277	23.0	2	80	13
C	7	15.0	9.0	7.432	19.6	2	58	25
D	9	27.6	25.0	7.106	21.3	5	64	23
E	27	16.0	10.0	2.746	14.2	3	48	5
F	28	9.4	6.5	1.583	8.3	2	38	4

Comfort

Although not every participant was able to put on a fitness tracker by him/herself, some participants were able to wear the fitness trackers with assistance putting them on and taking them off. Hence, the number of people who responded to questions about comfort may be greater for some fitness trackers than those who could put on or remove the fitness trackers independently. Table 3 displays the number and percentage of participants responding to an opinion level for each tracker for Questions 4-7, while Figure 1 represents the same information in bar charts.

For Question 4, “How easy was it to wear your fitness device?” 75% of the participants reported that Tracker F and 56% of Tracker E was easy or very easy to wear, compared to less than 22% for each of the other fitness trackers (n=4). There was an overall statistically significant difference in opinions using Friedman’s test ($\chi2(5) = 41.7008$, $p < 0.0001$). As depicted in Table 4, the Wilcoxon Signed-Rank tests showed that Tracker E ($p < 0.0001$) and Tracker F ($p < 0.0001$) was significantly different from both Tracker C and Tracker D in terms of easiness of wearing the device. Tracker B was different than Tracker C ($p = 0.0008$) and Tracker D ($p = 0.0002$). There was no statistical difference in opinion among Tracker A users compared with other trackers. Overall, favorability was higher for Tracker F followed by Tracker E for this question.

Results from Question 5 “How comfortable did the material feel against your arm?” show that Tracker E had the most respondents stating that the tracker was comfortable (65.63%) while Tracker A (31.25%) and Tracker F

(34.38%) showed some favorability but at a lesser percentage. However, Tracker F had a slightly higher percentage of participants saying that the tracker was rough against the arm (37.50%). Furthermore, Tracker C (65.63%) and Tracker D (53.13%) had the highest participants saying "Other" as their response. Using the omnibus Friedman's test, the overall opinions of comfortability against the arm among the six trackers were statistically significant ($\chi2(5) = 31.4100$, $p < 0.0001$). Tracker C showed a statistically significant difference compared to Tracker A ($p=0.0013$), Tracker E ($p < 0.0001$), and Tracker F ($p < 0.0001$) while Tracker D was different than Tracker E ($p = 0.0009$). Overall, Tracker E scored highly in comfortability when worn against the arm over all the other devices.

For Question 6 "How comfortable is it to carry the device while working out?" results show that Tracker E had the most responses for comfortability (53.13%) while Tracker A (37.50%) and Tracker F (34.38%) showed some favorability but at a lesser percentage. Similar to Question 5, Tracker F had a moderate response rate for roughness (28.13%). Roughness was also found for a quarter of the respondents for Tracker C (25.00%). Among the "Other" opinion category, Tracker C (53.13%) and Tracker D (53.13%) had the same percentage of participants answering this response. The overall difference in comfortability when carrying the device using Friedman's test was statistically significant ($\chi2(5) = 26.2207$, $p < 0.0001$). Similar to results for Question 5, Tracker C showed a significant difference compared to Tracker E ($p < 0.0001$) and Tracker F ($p < 0.0001$) while Tracker D showed a difference compared to Tracker E ($p = 0.0017$). In summary, when comfortability in carrying the device was taken to account, Tracker E outperformed all other trackers.

Table 3: Frequency tables and proportions of responses for Questions 4 through 7 (asks about the comfort and the likelihood of continuing to use the tracker over the time).

Question 4: How easy is it to wear your fitness device?						
Opinion Level	Tracker A (%)	Tracker B (%)	Tracker C (%)	Tracker D (%)	Tracker E (%)	Tracker F (%)
Very Easy	3 (9.3)	5 (15.6)	1 (3.1)	1 (3.1)	11 (34.3)	16 (50.0)
Easy	4 (12.5)	2 (6.2)	1 (3.1)	2 (6.2)	7 (21.8)	8 (25.0)
Moderate	7 (21.8)	4 (12.5)	4 (12.5)	1 (3.1)	4 (12.5)	0 (0.0)
Hard	3 (9.3)	7 (21.8)	1 (3.1)	4 (12.5)	2 (6.2)	0 (0.0)
Very Hard	4 (12.5)	5 (15.6)	9 (28.1)	7 (21.8)	3 (9.3)	0 (0.0)
Other	11 (34.3)	9 (28.1)	16 (50.0)	17 (53.1)	5 (15.6)	8 (25.0)
Question 5: How comfortable did the material feel against your arm?						
Opinion Level	Tracker A	Tracker B	Tracker C	Tracker D	Tracker E	Tracker F
Comfortable	10 (31.2)	9 (28.1)	3 (9.3)	5 (15.6)	21 (65.6)	11 (34.3)
Moderate	7 (21.8)	5 (15.6)	6 (18.7)	7 (21.8)	4 (12.5)	6 (18.7)
Rough	4 (12.5)	4 (12.5)	2 (6.2)	3 (9.3)	1 (3.1)	12 (37.5)
Other	11 (34.3)	14 (43.7)	21 (65.6)	17 (53.1)	6 (18.7)	3 (9.3)
Question 6: How comfortable is it to carry the device while working out?						
Opinion Level	Tracker A	Tracker B	Tracker C	Tracker D	Tracker E	Tracker F
Comfortable	12 (37.5)	9 (28.1)	2 (6.2)	4 (12.5)	17 (53.1)	11 (34.3)
Moderate	1 (3.1)	8 (25.0)	5 (15.6)	6 (18.7)	7 (21.8)	7 (21.8)
Rough	7 (21.8)	3 (9.3)	8 (25.0)	5 (15.6)	2 (6.2)	9 (28.1)
Other	12 (37.5)	12 (37.5)	17 (53.1)	17 (53.1)	6 (18.7)	5 (15.6)
Question 7: How likely is it that would start to use/continue using this device?						
Opinion Level	Tracker A	Tracker B	Tracker C	Tracker D	Tracker E	Tracker F
Very Likely	2 (6.2)	5 (15.6)	0 (0.0)	1 (3.1)	17 (53.1)	10 (31.2)
Maybe	1 (3.1)	2 (6.2)	0 (0.0)	0 (0.0)	1 (3.1)	4 (12.5)
Never	2 (6.2)	3 (9.3)	2 (6.2)	7 (21.8)	0 (0.0)	1 (3.1)
Other	27 (84.3)	22 (68.7)	30 (93.7)	24 (75.0)	14 (43.7)	17 (53.1)

Overall likelihood

Question 7 "How likely is it that you would start to use/continue this device?" provided results that showed Tracker E having the highest percentage of respondents saying they were very likely to use the tracker among all our trackers studied (53.13%). However, there was a moderate percentage of participants who stated "Other" for the same tracker (43.75%). Overall, the percentage of participants saying "Other" for this question was very high for all trackers. No

participants preferred Tracker C while a sizable percentage stated they are not willing to wear Tracker D (21.88%). The overall Friedman's test was found to be statistically significant ($\chi 2(5) = 34.3140$, $p < 0.0001$). Tracker E ($p < 0.0001$) and Tracker F ($p < 0.0001$) was significantly different than both Tracker C and Tracker D. Additionally, Tracker A was found to be significantly different than Tracker E ($p < 0.0002$). From this analysis, there is a higher preference in using Tracker E and to a lesser extent Tracker F over the other tracking devices.

Figure 1. Graphical representation of participants' responses to Questions 4 through 7 (asks about the comfort and the likelihood of continuing to use the tracker over the time).

Discussion

We sought to identify trackers that demonstrate the best ease of use, comfort, and favorability (likelihood of continued use) among PD patients. Overall, Trackers E and F received the highest favorability while Trackers C and D received much lower favorability. The results suggest that each fitness tracker is different in terms of ease of use, comfort, and favorability among PD patients but that none stood above the others on all measures. Tracker E is the most comfortable to wear and is the model most likely to be worn over an extended period of time. Tracker F was not as comfortable against the arm while working out compared to Tracker E. Trackers A and B scored the lowest on ease of use but were only marginally more comfortable. Additionally, if a tracker was found to be highly comfortable against the arm, the tracker was also likely to be comfortable when worn during exercise activities. This is shown in Figure 1 where the distribution of responses among the six trackers in Question 5 was similar to Question 6. This may indicate that different measures of comfortability of the tracker might not all be that different from each other.

Table 4: Wilcoxon multiple comparison p-value denotations between trackers*

Tracker Comparison	Question 4	Question 5	Question 6	Question 7
A vs B	0.9	0.4	0.8	0.2
A vs C	0.0	0.0	0.0	0.1
A vs D	0.0	0.0	0.0	1.0
A vs E	0.0	0.0	0.0	0.0
A vs F	0.0	0.2	0.2	0.0
B vs C	0.0	0.0	0.0	0.0
B vs D	0.0	0.2	0.0	0.1
B vs E	0.0	0.0	0.0	0.0
B vs F	0.0	0.1	0.3	0.2
C vs D	0.8	0.2	0.3	0.0
C vs E	<.0001	<.0001	<.0001	<.0001
C vs F	<.0001	<.0001	0.0004	<.0001
D vs E	<.0001	0.0009	0.0	<.0001
D vs F	<.0001	0.0	0.0	0.0021
E vs F	0.4	0.0	0.2	0.3

*Shaded values represent statistically significant p-values according to the Bonferroni adjustment critical p-value at $\alpha = 0.00341$

For the quickest time necessary to wear the device, Tracker F performed the best overall and had the most participants successfully putting on and taking off the device. This finding was highly reflected in the results from Question 4 on the survey on the ease of use. Conversely, Trackers C and D took longer for the participants to put on the device and had the lowest number of participants who could successfully put on the device. Trackers C and D also had the smallest number of participants say that the tracker was easy to use. However, despite Tracker C having the longest time to put on the device, the removal time was the second quickest among the other trackers. The reason behind this was the design of the tracker's wristband which had a "lock within the hole" mechanism that made it harder to put on but easy to remove. Patients had to put the device on their arm and then push the device button located on the band into the hole to completely attach and lock it. However, for participants with more tremor, it was extremely difficult for them to put it on. In fact, some patients gave up trying to put the fitness tracker on because it was too difficult. However, removing the device was quick because it took only one pull to disengage the lock.
Further investigation may posit that the form or method of the tracker being worn may prove significant as some trackers may have replaceable straps, Velcro, or capable wristbands. Some wristbands may be difficult to wear or put on the wrist but are easily removed and vice versa. Recommendations on the best fitness tracker for PD patients should consider how the device can be worn appropriately and easily without limiting the physical activity of the wearer.

To minimize the effect of sponsor bias, we masked the brand names of all six trackers used in the study and referred to each tracker with alphabetical names. Our participants represented PD patients at all stages of the disease and symptom progression. Therefore, our results are not biased towards patients at any particular stage in the disease or symptom progression. Most of the patients have not used any fitness tracker devices as all the trackers were new to them. Some of our questions were open-ended so that participants could freely express their opinions without having

strictly defined options for answer choices.

Existing studies on wearable devices mainly focused on product design specifications and user-interface applications. Several reviews on wearable devices appearing on internet web pages have different opinions which are often subjective and lack experimental results to evaluate the type of participants and the accuracy of information on the devices. For example, a study was performed on "comparison of wearable fitness devices" where the authors focused on both subjective and objective data irrespective of manufacturer's claims to compare the user's satisfaction, user-friendliness, and accuracy of data collected and managed among four commercially popular fitness trackers, namely Fitbit Flex, Withings Pulse, Misfit Shine, and Jawbone [33]. This study reported that Misfit Shine scored the highest for its design and hardware features in contrast to Withings Pulse which was recommended by the highest number of participants for its user-friendliness, accuracy in activity tracking, and satisfaction levels among all the devices they tried. Furthermore, the study acknowledges that the design and the technology together are on the same levels to evaluate the quality of tracking devices both objectively and subjectively. In another study, the acceptance of wearable trackers (Fitbit Zip, Jawbone Up 24, Misfit Shine, Withings Pulse) is tested in participants with chronic illness. The study concluded that Fitbit Zip secured the highest mean acceptance score of 68 although having negative feedback that it is too short to put on and a mean acceptance score of 65 for Jawbone despite having positive comments on its design and ease of wearing [34]. Most of these studies compared the wearable trackers based on their design characteristics or user-interface applications. But there is no such study which can tell us about the ease of use and comfort while wearing these devices in patients with PD because patients are suffering from symptoms such as muscle rigidity, tremors, problems with muscle coordination and balance pose difficulty in wearing the trackers by themselves. This prevents them to choose a wearable tracker for long-term usage that could possibly affect the intervention to track the disease progression. Users are often compromising these behavioral characteristics for other reasons such as the cost of the device and features or applications the device provides for tracking the activities.

Limitations and future work

Our study was limited to 32 participants from a single organization, which may not be a large enough sample to achieve statistically valid estimates of patterns to answer our questions and generalize the entire US population. In the future, we will increase our sample size by recruiting more patients from the different locations of Rock Steady Boxing and different organizations to conduct a study which may be able to provide more generalizable results.
Second, we did not stratify our subject population by stage of progression or severity. It may be that people at different stages or levels of severity may react differently with respect to ease of use and comfort. In the future, we will categorize the patient population by the PD severity and examine their perception and attitude towards these fitness trackers. Third, our study looked only at the ease of use and comfort of the trackers without considering the potential value of new information that might benefit the subject in his or her management of their PD condition. It could be that those reactions that are perceived as barriers to ease of use or comfort initially may be less important once the subject sees the value of the information. Finally, patients were exposed to the fitness for a short duration of time. Therefore, the exposure to the trackers was relatively short. In the future, we will expand our study to examine the use of these trackers in daily life and identify the most popular tracker among patients impaired with PD. For the next step, we propose to use fitness tracker E and F to collect raw data of patients impaired with PD while they perform the exercise. We will use this raw data to develop algorithms that can identify the type of exercise (dancing, boxing, etc.), and its intensity. Using this information, we will evaluate its impact on PD progression. We will also consider other factors towards the user adoption such as battery life, Bluetooth syncing, and mobile applications. Future work should also include the involvement of PD patients in the long-term design acceptance and functional capabilities of exercise management programs to enhance adoption rate.

Conclusions

Owing to the recent advancements in devices designed specifically to measure motor function, commercially available fitness trackers play a key role in tracking the physical activities that could determine PD progression. Our study demonstrated the ease of use and comfort of various commercially available wearable trackers in PD patients. We conclude that commercially available fitness trackers should be considered as an alternative to dedicated devices or other devices that could be adapted for this purpose. Moreover, while none of the fitness trackers received high scores in all areas, at least two fitness trackers, E and F, received significantly higher positive responses than others while

two fitness trackers, C and D, performed significantly worse. Therefore, we conclude that further study is warranted to look more closely at the design and use of fitness trackers for PD patients.

Internal Review Board

This study was approved by the Internal Review Board at Indiana University Purdue University Indianapolis (Protocol approval # 1512203933).

Acknowledgments

The authors wish to thank the Rock Steady Boxing Team, Indianapolis, for their cooperation and Parvati Menon, a graduate student at the School of Informatics and Computing, Indiana University Purdue University Indianapolis, for her support.

Author Contributions

All the authors contributed equally to the development of this work.

Conflict of Interest

The authors report no conflicts of interest relating to this work.

References

1. Andoskin, P., Emelyanov, A., Nikolaev, M., Senkevich, K., Shilin, V., Yakimovskiy, A. Pchelina, S. J. U. Z. S.-P. G. M. U. i. A. I. P. (2015). Parkinson's disease (PD) is the most common neurodegenerative disease. 22(2), 14-17.
2. Karlsen, K. H., Tandberg, E., Årsland, D., Larsen, J. P. J. J. o. N., Neurosurgery, & Psychiatry. (2000). Health related quality of life in Parkinson's disease: a prospective longitudinal study. 69(5), 584-589.
3. Olanow, C. W., Kieburtz, K., Schapira, A. H. J. A. o. N. O. J. o. t. A. N. A., & Society, t. C. N. (2008). Why have we failed to achieve neuroprotection in Parkinson's disease? , 64(S2), S101-S110.
4. Horstink, M., Tolosa, E., Bonuccelli, U., Deuschl, G., Friedman, A., Kanovsky, P. Poewe, W. J. E. j. o. n. (2006). Review of the therapeutic management of Parkinson's disease. Report of a joint task force of the European Federation of Neurological Societies and the Movement Disorder Society–European Section. Part I: early (uncomplicated) Parkinson's disease. 13(11), 1170-1185.
5. Olanow, C. W., Kieburtz, K., Odin, P., Espay, A. J., Standaert, D. G., Fernandez, H. H. Robieson, W. Z. J. T. L. N. (2014). Continuous intrajejunal infusion of levodopa-carbidopa intestinal gel for patients with advanced Parkinson's disease: a randomised, controlled, double-blind, double-dummy study. 13(2), 141-149.
6. Hubble, R. P., Naughton, G., Silburn, P. A., Cole, M. H. J. A. j. o. p. m., & rehabilitation. (2018). Trunk Exercises Improve Gait Symmetry in Parkinson Disease: A Blind Phase II Randomized Controlled Trial. 97(3), 151-159.
7. Flach, A., Jaegers, L., Krieger, M., Bixler, E., Kelly, P., Weiss, E. P., & Ahmad, S. O. J. N. l. (2017). Endurance exercise improves function in individuals with Parkinson's disease: A meta-analysis. 659, 115-119.
8. Wu, P.-L., Lee, M., & Huang, T.-T. J. P. o. (2017). Effectiveness of physical activity on patients with depression and Parkinson's disease: a systematic review. 12(7), e0181515.
9. de Carvalho Oliveira, A., Murillo-Rodriguez, E., Rocha, N. B., Carta, M. G., Machado, S. J. C. p., CP, e. i. m. h., & EMH. (2018). Physical Exercise For Parkinson's Disease: Clinical And Experimental Evidence. 14, 89-98.
10. Lang, A. E., & Obeso, J. A. J. T. L. N. (2004). Challenges in Parkinson's disease: restoration of the nigrostriatal dopamine system is not enough. 3(5), 309-316.
11. Goodwin, V. A., Richards, S. H., Taylor, R. S., Taylor, A. H., & Campbell, J. L. J. M. d. (2008). The effectiveness of exercise interventions for people with Parkinson's disease: A systematic review and meta-analysis. 23(5), 631-640.
12. Pires, A. O., Teixeira, F. G., Mendes-Pinheiro, B., Serra, S. C., Sousa, N., & Salgado, A. J. J. P. i. n. (2017). Old and new challenges in Parkinson's disease therapeutics. 156, 69-89.
13. Geraedts, V. J., Marinus, J., Gouw, A. A., Mosch, A., Stam, C. J., van Hilten, J. J., . . . Tannemaat, M. R. J. C. N. (2018). Quantitative EEG reflects non-dopaminergic disease severity in Parkinson's disease.
14. Klassen, B., Hentz, J., Shill, H., Driver-Dunckley, E., Evidente, V., Sabbagh, M., . . . Caviness, J. J. N. (2011). Quantitative EEG as a predictive biomarker for Parkinson disease dementia. 77(2), 118-124.

15. Dahdal, P., Meyer, A., Chaturvedi, M., Nowak, K., Roesch, A. D., Fuhr, P., . . . disorders, g. c. (2016). Fine motor function skills in patients with Parkinson disease with and without mild cognitive impairment. 42(3-4), 127-134.
16. Meyer, A., Bogaarts, J., Cozac, V., Chaturvedi, M., Handabaka, I., Hatz, F., . . . Fuhr, P. J. C. N. (2018). P77. Prognosis of cognitive decline in Parkinsons disease: a combined marker of quantitative EEG and clinical variables improves prediction. 129(8), e98-e99.
17. Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE transactions on Biomedical Engineering, 57(4), 884-893.
18. Kooiman, T. J., Dontje, M. L., Sprenger, S. R., Krijnen, W. P., van der Schans, C. P., de Groot, M. J. B. s. s., medicine, & rehabilitation. (2015). Reliability and validity of ten consumer activity trackers. 7(1), 24.
19. Patel, S., Park, H., Bonato, P., Chan, L., Rodgers, M. J. J. o. n., & rehabilitation. (2012). A review of wearable sensors and systems with application in rehabilitation. 9(1), 21.
20. El-Amrawy, F., & Nounou, M. I. J. H. i. r. (2015). Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial? , 21(4), 315-320.
21. Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. J. P. M. (2016). The rise of consumer health wearables: promises and barriers. 13(2), e1001953.
22. van der Kolk, N. M., & King, L. A. J. M. D. (2013). Effects of exercise on mobility in people with Parkinson's disease. 28(11), 1587-1596.
23. Cancela, J., Pastorino, M., Tzallas, A. T., Tsipouras, M. G., Rigas, G., Arredondo, M. T., & Fotiadis, D. I. J. S. (2014). Wearability assessment of a wearable system for Parkinson's disease remote monitoring based on a body area network of sensors. 14(9), 17235-17255.
24. Michaelis, J. R., Rupp, M. A., Kozachuk, J., Ho, B., Zapata-Ocampo, D., McConnell, D. S., & Smither, J. A. (2016). Describing the user experience of wearable fitness technology through online product reviews. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
25. Preusse, K. C., Mitzner, T. L., Fausset, C. B., & Rogers, W. A. J. J. o. A. G. (2017). Older adults' acceptance of activity trackers. 36(2), 127-155.
26. Vooijs, M., Alpay, L. L., Snoeck-Stroband, J. B., Beerthuizen, T., Siemonsma, P. C., Abbink, J. J., . . . Rövekamp, T. A. J. I. j. o. m. r. (2014). Validity and usability of low-cost accelerometers for internet-based self-monitoring of physical activity in patients with chronic obstructive pulmonary disease. 3(4).
27. Steinert, A., Haesner, M., & Steinhagen-Thiessen, E. J. U. A. i. t. I. S. (2018). Activity-tracking devices for older adults: comparison and preferences. 17(2), 411-419.
28. Hickey, A. M., & Freedson, P. S. J. P. i. c. d. (2016). Utility of consumer physical activity trackers as an intervention tool in cardiovascular disease prevention and treatment. 58(6), 613-619.
29. Ridgers, N. D., Timperio, A., Brown, H., Ball, K., Macfarlane, S., Lai, S. K., . . . uHealth. (2018). Wearable Activity Tracker Use Among Australian Adolescents: Usability and Acceptability Study. 6(4).
30. Pfannenstiel, A., & Chaparro, B. S. (2015). An investigation of the usability and desirability of health and fitness-tracking devices. Paper presented at the International Conference on Human-Computer Interaction.
31. Rasche, P., Schäfer, K., Theis, S., Bröhl, C., Wille, M., Mertens, A. J. I. J. o. H. F., & Ergonomics. (2016). Age-related usability investigation of an activity tracker. 4(3-4), 187-212.
32. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3), 276-282.
33. Kaewkannate, K., & Kim, S. J. B. P. H. (2016). A comparison of wearable fitness devices. 16(1), 433.
34. Mercer, K., Giangregorio, L., Schneider, E., Chilana, P., Li, M., Grindrod, K. J. J. m., & uHealth. (2016). Acceptance of commercially available wearable activity trackers among adults aged over 50 and with chronic illness: a mixed-methods evaluation. 4(1).

An Examination of the Statistical Laws of Semantic Change in Clinical Notes

Kevin J. Peterson, MS[1,3], Hongfang Liu, PhD[2]
[1] Department of Information Technology, Mayo Clinic, Rochester, MN
[2] Department of Health Sciences Research, Mayo Clinic, Rochester, MN
[3] Bioinformatics and Computational Biology Program, University of Minnesota, Minneapolis, MN

ABSTRACT

Natural language is continually changing. Given the prevalence of unstructured, free-text clinical notes in the healthcare domain, understanding the aspects of this change is of critical importance to clinical Natural Language Processing (NLP) systems. In this study, we examine two previously described semantic change laws based on word frequency and polysemy, and analyze how they apply to the clinical domain. We also explore a new facet of change: whether domain-specific clinical terms exhibit different change patterns compared to general-purpose English. Using a corpus spanning eighteen years of clinical notes, we find that the previously described laws of semantic change hold for our data set. We also find that domain-specific biomedical terms change faster compared to general English words.

INTRODUCTION

Even with the increasing digitization of the medical chart by modern Electronic Health Records (EHRs), a large amount of patient information is still encoded using unstructured, free-text representations.[1,2] This text-based clinical narrative is the mechanism through which the patient "story" is conveyed, providing background and context for patients' pertinent health issues.[3] The linguistic characteristics of this narrative are varied and diverse,[4] and the effective use of this language in the clinical setting can be a contributing factor to clinical outcomes.[5] As such, understanding the linguistic characteristics of clinical language is an essential part of understanding how information is communicated in the clinical domain as a whole.[6]

The language used to describe clinical care is not static, however. Natural language is continually evolving,[7] and these observable linguistic differences over time are known as the "diachronic change" of language. Given the prominence of free-text in the biomedical domain, diachronic change is important to consider when applying Natural Language Processing (NLP) techniques. Robust NLP systems must be able to detect changes in grammar, syntax, and semantics in order to adapt to changing medical practices.[8]

To better understand and quantify this change, Hamilton et al. formalized several methods to statistically analyze shifts in word meaning, resulting in two laws describing semantic drift over time, or the *Laws of Semantic Change*.[9] These two laws provide a quantitative framework for linguistic analysis that can be used to examine diachronic semantic change:

- **Law of Conformity:** Frequently used words in a corpus will, on average, change meaning more slowly compared to infrequently used words.
- **Law of Innovation:** Polysemous words (or words with many different senses) change faster than words with a single or limited set of meanings.

The goal of this study is to test these laws of semantic change in the context of clinical notes. We apply the methods described by Hamilton et al.[9] to test if the *Law of Conformity* and the *Law of Innovation* hold for our clinical data set. We also extend beyond these two laws and examine one further aspect of change: whether domain-specific clinical terms change at a different rate compared to general-purpose words. Because clinical notes contain a mix of domain-specific medical terms and general English words, we are interested in quantifying any differences in the rate of change between the subsets.

BACKGROUND & SIGNIFICANCE

A key aspect of diachronic change analysis is a focus on word-level semantics, or how individual words change meaning over time. To facilitate this comparison, words are often transformed from discrete tokens to more comparable *word embeddings*, or vectorized representations of words. These real-valued vectors can then be compared, analyzed, and used to quantify diachronic change.[10] Advances in embedding techniques, specifically the introduction of models based on artificial neural networks, have helped improve word embeddings' effectiveness in detecting semantic shift.[11]

An important consideration for diachronic studies is choosing the time granularity with which to detect change. Traditionally, diachronic analysis has focused on detecting change at the granularity of decades or longer. With the advent of neural language models, however, these time spans have shortened drastically. Recent work has shown that semantic change in certain domains can be detected in a corpus spanning a total of only five years,[12] and several studies have suggested that change can be effectively measured at the granularity of a single year.[10,13] Detecting change in smaller time increments has significant pragmatic advantages – specifically, if detected quickly enough, diachronic change analysis may be used to improve and adjust running NLP systems. This can be especially important for systems where failing to account for change may pose significant performance or safety risks, such as clinical NLP. In this study, we aim to demonstrate that semantic shift with respect to the laws of semantic change can be detected in year-by-year time increments on a corpus of clinical notes.

Irrespective of how diachronic change is ultimately quantified, failure to take this change into account when processing text spanning multiple time periods can lead to incorrect or invalid conclusions.[14] Likewise, language change must be actively accounted for in deployed information systems in order to avoid performance degradation of fundamental downstream NLP tasks.[15] It is also anticipated that the increased pace of digitization of older data will exacerbate the problem, forcing existing NLP systems to adopt data-driven approaches to solving the problem of normalizing outdated language, as opposed to manual or one-off intervention by subject matter experts.[16,17]

Despite the potential impact to clinical NLP systems, there have been few reported studies where diachronic change analysis has been applied to the domain of clinical notes. One recent study proposed methods to track disease change via Wikipedia articles,[18] attempting to quantify the change in disease meaning using edit activity on public Wikipedia pages. Because of the importance of controlled terminologies and ontologies to the clinical domain, other studies have explored change through the evolution of publicly available clinical terminologies.[19,20] As semantic mismatches between standard terminologies and real-world clinical text can be subtle and difficult to detect,[21,22] we devote a portion of this study to examining how controlled terms change relative to general-purpose English words.

For this study, the motivation for quantifying language change in clinical notes is pragmatic: understanding the characteristics of diachronic drift will allow us to plan appropriately for changes to running clinical NLP systems. Sufficient monitoring may even allow NLP systems to become more robust to this drift, as there is evidence that even a small-to-moderate effort to account for language change has a positive impact on downstream NLP tasks.[23] Ultimately, a robust quantification of change can lead not only to improved detection and monitoring, but to normalization efforts such as reconciling differences in spelling[24] or vocabulary[25] over time.

METHODS

We structured our methods to align with the techniques used by Hamilton et al. in their examination of semantic change,[9] replicating their analysis of the *Law of Conformity* and *Law of Innovation* as closely as possible using a clinical notes corpus as input. We further extend their study methods and propose a new axis of semantic change important to our domain: the relative rate of change of domain-specific clinical terms.

Dataset

Our analysis was performed on a large clinical corpus of free-text notes from over 50,000 patients spanning eighteen years. Clinical notes were drawn from the years 2000 to 2017 for consistency, as in 2018 a new EHR implementation significantly changed note structure. Roughly seven million total clinical notes in total were included in the corpus. We focused specifically on clinical diagnoses and their descriptions, extracting text from the *Impression, Report and*

Plan (IRP) section, which is the narrative section of the clinical note that further describes a specific diagnosis or clinical problem. This section provides supporting context and detail regarding the diagnosis, along with current or proposed treatment plans. The text underwent minimal preprocessing to remove stop words and punctuation, and all text was converted to lower case. The corpus was then segmented by years to allow for temporal analysis. Note that this reflects a difference from the approach of Hamilton et al., where the text was grouped into decades.[9] This change in granularity was necessary as our clinical notes corpus has a relatively limited temporal span, but also deliberate as the ability to detect change in terms of smaller time increments is an emphasis of our approach.

Our clinical notes corpus was quite heterogeneous in regards to note type composition and included over one hundred different categories of notes, including progress notes, evaluations, discharge summaries, and so on. Although there were many possible types, the bulk of our corpus was made up of a limited set. Figure 1 shows the total distribution of note types, limited to the top ten types by count.

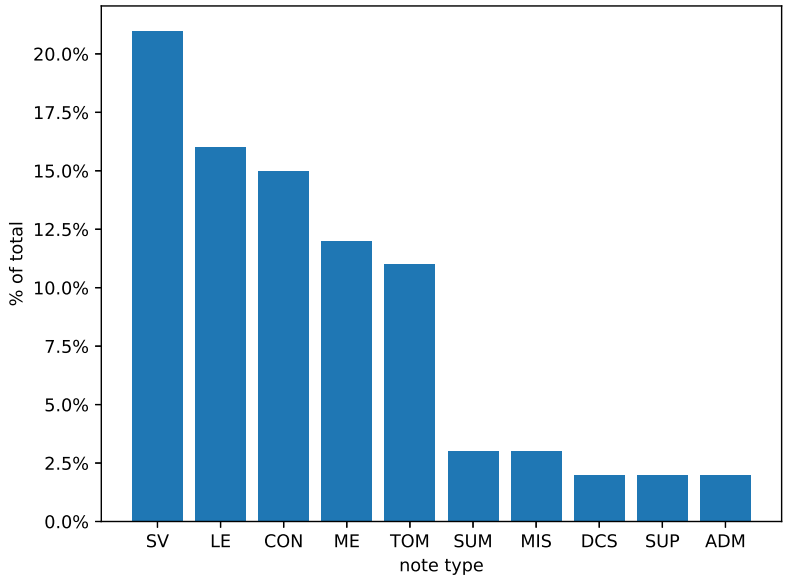


Figure 1: The distribution of the top ten most frequent note types drawn from our clinical corpus. Note type abbreviations are defined as follows:

SV	Subsequent Visit	SUM	Dismissal Summary
LE	Limited Exam	MIS	Miscellaneous
CON	Consult	DCS	Discharge Summary
ME	Multi-system Evaluation	SUP	Supervisory
TOM	Test-Oriented Miscellaneous	ADM	Hospital Admission Note

Another source of variation in our corpus was the number of words per year. We found that after grouping by year, word counts steadily increased year-by-year, reflecting the increasing ability to capture and store aspects of the digital patient record over time. To account for these differences, stratified sampling was used to produce a final corpus homogeneous in terms of both words-per-year and note type composition. Sampling was conducted such that (1) the overall note type distribution shown in Figure 1 was consistent for each year, and (2) the total number of notes sampled for each year was the same. Because clinical notes have some variation in word count, our final stratified corpus contained an average of 2,614,792 words per year with a standard deviation of 48,671 words.

Word Embedding

Word embeddings are transformations of discrete words into a continuous vector space. Embeddings allow words to be compared mathematically via cosine similarity or other measures. Many word embedding techniques are rooted in the theories of *distributional semantics* – the hypothesis that semantically similar words will be found in similar contexts.[26] In terms of word embeddings, distributional semantics can be leveraged to ensure that semantically similar words will be closer together in the vector space.

Word2vec,[27] a prominent unsupervised machine learning method for word embedding, uses neural networks to map words to vectors, and is the embedding model used for this study. Word2vec represents each word as a single, high-dimensional vector, regardless of the number of senses the word may carry. As a consequence, polysemous words are represented as a single vector that is either skewed toward one sense used predominately in the corpus, or some aggregation of the multiple sense vectors.[28] This is usually seen as a disadvantage and detrimental to downstream tasks – as such, several techniques have been introduced to allow for sense-specific embeddings.[29,30,31] For our purposes, however, having one vector per word (as opposed to one per word sense) is desirable, as changes to the vector of a polysemous word can show average movement to or from certain senses over time.

For a given word, if no change in meaning happens over time, we assume the word2vec embedding vector will also remain constant. This also applies to polysemous words – the vector representing the composition of the various senses of a word should remain fixed. We would, however, expect the vector to change in two circumstances: first, if the meaning of a word changes to something new over time. This would cause the vector to shift to a new portion of the vector space. Second, if a polysemous word undergoes a shift to or from one or more of the senses. This would force the vector to move away from the less-used senses and toward the preferred usage. In both instances, leveraging the fact that word2vec only allows for one vector per word, we will be able to detect both of these scenarios.

To facilitate year-by-year change analysis, we created a separate word2vec model for each year of our corpus. Word embedding processing was conducted using the Gensim[32] word2vec implementation using default settings: 100-dimensional word vectors, a window size of 5 words, and using the Continuous Bag of Words (CBOW) model.

Embedding Alignment

To examine vector changes over time, we compared the separately trained word2vec models from every time period under study – in our case, for every year. There is an issue regarding this technique, however. Two word2vec models cannot be directly compared, as vectors in any two word2vec models are subject to randomizations within the training algorithm and will not meaningfully align, even if the training corpus and hyperparameters are held constant.[33] We assume, however, that even though absolute vector positioning between models cannot be compared, vectors do maintain their relative position as compared to other vectors.[11] This means that a linear transformation applied to one vector space could be used to align it to another space. As such, a common solution to the problem is to learn a linear transformation that maps each vector in a source vector space to a target space.[34,9]

This transformation can be accomplished through an orthogonal Procrustes transformation.[35] Using this method, an orthogonal matrix is learned such that the sum of squares of word vector distances from one vector space to another is minimized. The following equation represents this transformation:

$$R = \underset{\Omega:\, \Omega^T\Omega = I}{\arg\min} \sum_{w \in V} ||\Omega \mathbf{v}_w^{(i)} - \mathbf{v}_w^{(j)}||^2$$

where i and j represent two different vector spaces (i.e., different word2vec models). The alignment is done based on word2vec vectors $\mathbf{v}$ for a given word w from a given year i or j. The vocabulary used in the Procrustes alignment, denoted as V, is the intersection of the vocabularies of both years of interest.

We calculate the diachronic embedding similarity of a word's meaning across two years i and j by first taking the word vector from year i and applying the Procrustes transformation R to transform it into the target vector space. We then compute the cosine distance of the resulting vector with the corresponding word vector from year j:

$$\Delta_w^{(i,j)} = cosine\_distance(R\mathbf{v}_w^{(i)}, \mathbf{v}_w^{(j)})$$

Figure 2 shows how an example Procrustes transform can align two different word2vec vector spaces. In this example, we learn a transform that moves the year 2000 vectors (shown as the solid lines) as close as possible to the year 2010 vectors (the dashed lines), with the small arrows representing the transformation. A single linear transform is able to align most vectors, meaning that most words will be represented by similar vectors (and thus, have similar meanings) once the vector spaces are aligned. Some word vectors, however, will not fully align even after the transformation is applied, signifying a vector change beyond that which could be accounted for by the alignment of the vector spaces. The word "portal" in this example is one of those cases. As seen, the alignment transformation places the predicted 2010 vector far away from the actual 2010 vector. We can infer from this that the meaning of the word "portal" did in fact change over this time span.

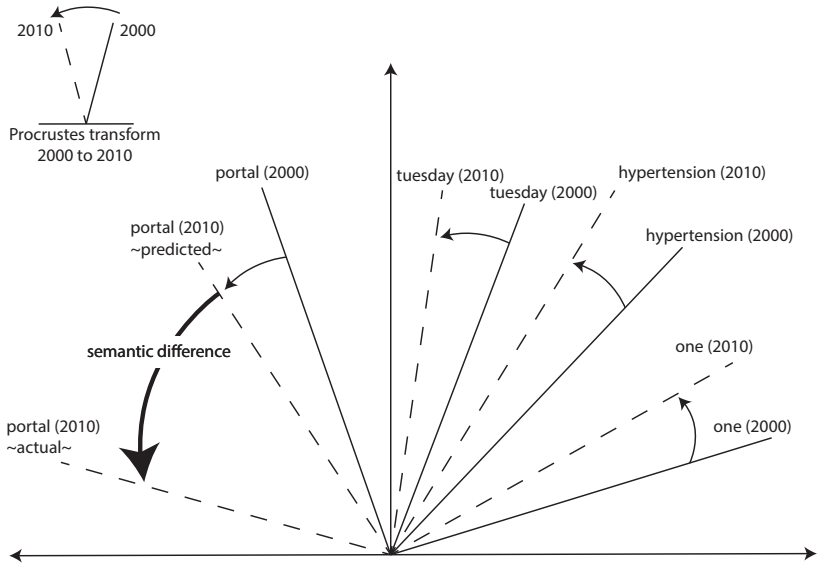


Figure 2: An example alignment of two word2vec vector spaces using a Procrustes linear transform. The "semantic difference" arrow here illustrates a temporal vector difference that is not accounted for via vector space alignment. We interpret this discrepancy as an indication of semantic drift.

Quantifying Polysemy

In order to analyze the *Law of Innovation*, a technique for quantifying polysemy was needed. In this study, polysemy was measured using the same technique as Hamilton et al.,[9] i.e., the negative of the local clustering coefficient.[36] In this context, the clustering coefficient of a given word is the degree to which a set of words that co-occur with it also co-occur with each other. Words with high clustering coefficients have predictable and highly interconnected co-occurrence graphs (and thus less polysemy, as this indicates that they occur in a limited set of contexts). Conversely, a low clustering coefficient indicates that the word co-occurs with a diverse set of words that are less interconnected, indicating usage in a broader set of contexts, or more polysemy.

Detecting Biomedical Terms

In this study we also aim to quantify any differences in the rates of change between biomedical terms and general-purpose English words. We determine whether a word in our corpus is a domain-specific biomedical term using the Unified Medical Language System (UMLS),[37] a large collection of controlled biomedical terminologies. We used UMLS version 2018AB (Level 0 + SNOMED CT) as a reference for biomedical terms, as 2018 represents the end of our corpus. QuickUMLS,[38] a concept extraction tool built on the UMLS, was used to detect whether a word was a biomedical term. A word was considered a biomedical term if QuickUMLS returned one or more matches to a UMLS concept for the word using the default QuickUMLS match settings.

Analysis of Known Semantic Shifts

To validate our techniques, words with known shifts in semantics were analyzed to determine if our word embedding approach could detect these expected changes. Two words were chosen, 'portal' and 'guideline,' as they correspond to known changes to clinical documentation and can be pinpointed to specific times. First, 'portal' was chosen due to the launch of the Mayo Clinic Patient Portal, an online tool to assist with telehealth and provider-to-patient communication. This online tool was first introduced in April 2010, with a broader release in August 2011.[39] Next, the word 'guideline' was examined. We selected this word due to the introduction in 2009 of AskMayoExpert (AME), a centralized system to disseminate clinical knowledge in the form of standardized care guidelines.[40] We expect that, due to their impact on clinical documentation, the changing semantics of these two words can be traced to the development timelines of these tools.

To conduct this analysis, for each of the two chosen words we measured the cosine distance between word vectors from a starting (or baseline) year in our corpus to every subsequent year. Although the Procrustes transformation was used to align word vectors between years, words that are changing meaning will remain misaligned even after transformation. The degree of this word vector misalignment for each year reflects relative semantic change compared to the baseline year.

Analysis of the Laws of Semantic Change

Semantic change is modeled via a linear mixed-effects regression model corresponding closely to the model used by Hamilton et al.[9] The model is scoped to determine word vector change for one time step increment, and in our case, a time step is defined as one year (from $year$ to $year + 1$). Note that this differs from the Hamilton et al. approach where the time unit was one decade.[9] The full model is specified as follows:

$$\Delta_w^{(year,year+1)} = \beta_{freq}\, log(frequency(w)) + \beta_{poly}\, log(polysemy(w)) + \beta_{year} + \beta_{type} + z_w + \varepsilon$$

where β_{freq} represents the log-transformed relative frequency of the word, β_{poly} the log-transformed polysemy value, and β_{year} the year in which we are examining word change. We include one random effect z_w, a categorical variable reflecting the individual word, to account for variation in word-specific change rate and allow our model to use repeated measures of the same word over different time periods. The resulting semantic distance changes were standardized and transformed via a Box-Cox transformation.[41] Our addition to this model – the comparison of change rates between biomedical terms vs. general English words – is accounted for via a binary fixed effect variable β_{type} corresponding to whether or not the word is a biomedical term.

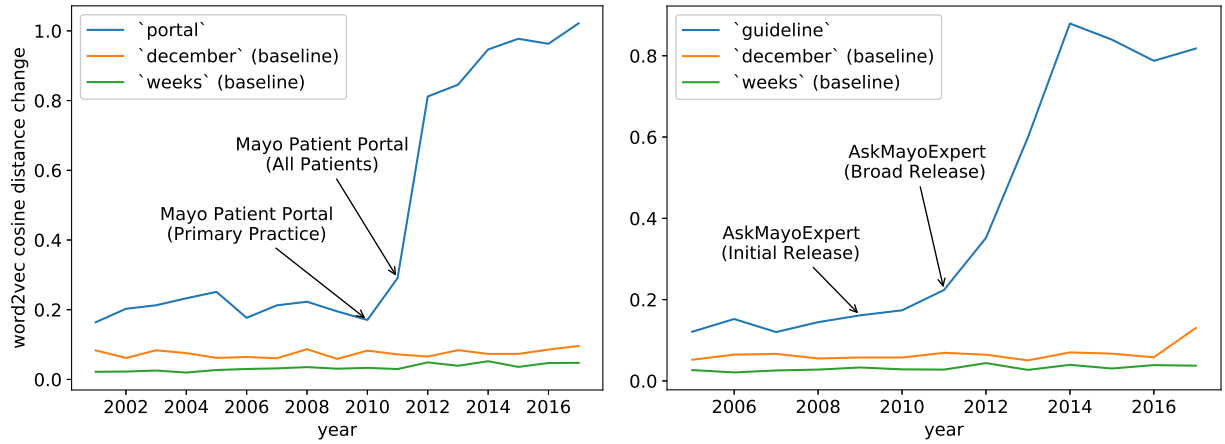
Model training was initiated by first grouping word embeddings into ($year$, $year + 1$) pairs for all years from 2000 to 2017. For each pair, all common words with 100 or more occurrences in both years were selected for analysis. The word embeddings were then aligned using the Procrustes alignment described above, and semantic distance was calculated for each word individually. All features were then processed as described and input into the model.

RESULTS

Analysis of Known Semantic Shifts

As shown in Figure 3, the two scenarios of known semantic change are detectable via different rates of change within the word2vec vector space. Baseline words like "december" and "weeks" change very little over time, while the words "guideline" and "portal" show considerable amounts of change. In these figures, significant events hypothesized to be associated with the semantic change are plotted for reference.

The word clouds for the selected words "guideline" and "portal" are shown in Figure 4, indicating the set of words most closely related via word2vec vector similarity for a given year. In this figure, each word cloud is visualized using t-SNE dimensionality reduction of the word vectors.[42] Figure 4a illustrates how the word "portal" has changed in meaning over time. In this example, it moved from having a mostly anatomical meaning to representing various emerging forms of telehealth – specifically the Mayo Clinic Patient Portal and its increasing integration with facets of clinical care documentation. For the word "guideline" in Figure 4b, differences in the two word clouds can be

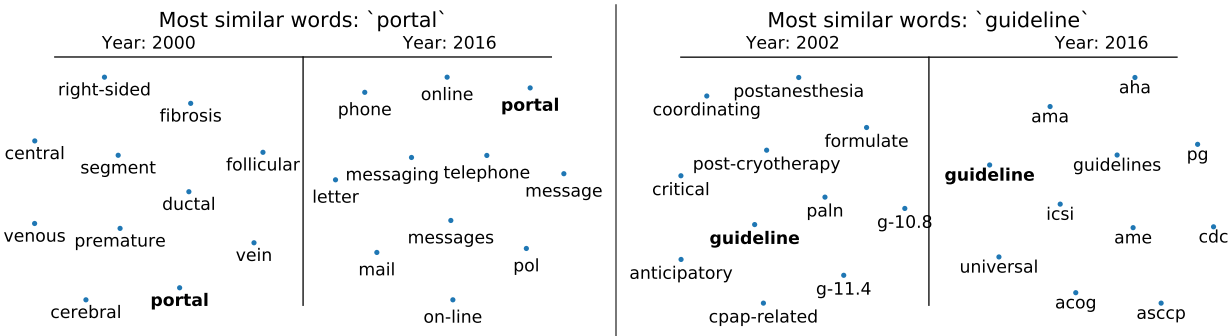


(a) Semantic change of the word word "portal."

(b) Semantic change of the word word "guideline."

Figure 3: Associating semantic word change with known past events. In these figures, plot lines represent word2vec cosine distance for each year compared to the initial year. As shown, the words "portal" and "guideline" show significant change corresponding to the introduction of two new clinical tools.

interpreted as differences over time in patient care guideline usage. As shown, as care guideline usage evolves, so does the neighborhood of similar word2vec vectors for the word "guideline."



(a) The neighborhood of similar words for the word "portal" from year 2000 (left) vs. 2016 (right).

(b) The neighborhood of similar words for the word "guideline" from year 2002 (left) vs. 2016 (right).

Figure 4: Contrasting similarity word clouds for two words exhibiting substantial change in meaning over time.

Analysis of the Laws of Semantic Change

Table 1 shows the regression coefficients of interest from our trained model. A negative coefficient in this case indicates that the particular effect corresponds on average to slower rates of change. In this table, β_{type} reflects the positive case, meaning it shows the impact to rate of change if a word was recognized as a biomedical term. Overall, we found that 42% of words analyzed could be linked to a controlled term in the UMLS.

Discussion

Table 1 generally corresponds to the findings of Hamilton et al.[9] in that higher frequency corresponds to slower rates on average of semantic change, and higher polysemy is associated with faster rates of change. These two findings are evidence that the *Law of Conformity* and the *Law of Innovation* both hold for our corpus. We also found a positive

Table 1: Results of the Laws of Semantic Change analysis in terms of frequency (*Law of Conformity*), polysemy (*Law of Innovation*), and whether the word is a domain-specific biomedical term.

Effect	Coefficient	95% CI	p-value
β_{freq} (Frequency)	-0.584	(-0.604, -0.564)	<1e-04
β_{poly} (Polysemy)	0.119	(0.087, 0.150)	<1e-04
β_{type} (Biomedical Term)	0.169	(0.111, 0.227)	<1e-04

relationship between frequency and polysemy (0.42 Pearson correlation), which is in line with correlations reported by Hamilton et al.[9] In addition, we also find that words in our corpus corresponding to biomedical terms in the UMLS showed on average higher rates of change when compared to general-purpose words. This finding suggests that these domain-specific terms may be subject to more volatility, possibly driven by the pace of change in the healthcare domain itself. More exploration into this finding is necessary, however, due to the inherent difficulty of precisely defining and classifying the notion of a "biomedical term" (see the Limitations & Future Work section below for further information).

Figure 3 demonstrates that known shifts in word meaning do in fact correspond to changes in word embeddings. More importantly, we were able to see a correspondence with semantic change and potential drivers of change – in this case, the implementation of two new clinical tools. This not only adds evidence to support that word embeddings can be effectively used to detect change, but suggests that it is possible to pinpoint change at the granularity of year increments.

To further explore this change, Figure 4 examines the change in word clouds for the two words in our known change scenarios. Figure 4a shows the shift of the word "portal" to an entirely different sense, moving from a meaning focused on anatomy to one centered on messaging and patient interaction via technology. We believe this change is due to general trends of technology change including the introduction of the Mayo Clinic Patient Portal and the expanding role of online patient services at the Mayo Clinic over that time. For the word "guideline" in Figure 4b, we see what we believe to be the influence of AskMayoExpert. The impact of this tool is visible via the new guideline types in the word cloud, as well as the appearance of the indicative "ame" acronym.

CONCLUSION

In this study we have conducted an examination of the statistical laws of semantic change in clinical notes. We find that the two laws of change described by Hamilton et al., the *Law of Conformity* and the *Law of Innovation*,[9] both hold for our corpus of clinical text. In addition, we find some evidence in our corpus that words corresponding to biomedical terms in the UMLS change more frequently than general language. More work is needed to precisely define which types of biomedical terms are changing and how, but we believe this finding may be an important consideration when conducting further analysis of the evolving semantics of clinical notes.

LIMITATIONS & FUTURE WORK

A limitation of this study is the narrow time span of the corpus, driven mostly by the relatively recent large-scale digitization of the patient record. Also, using stratified sampling such that each year contained the same amount of words greatly reduced the fraction of our corpus that was usable, as recent years are significantly larger in terms of size. While the reduction in usable size of our corpus was not ideal, we felt that controlling for variations in size and clinical note type distribution over the years was important.

Another large limitation is the ambiguity in defining and classifying biomedical terms. The UMLS covers an expansive array of terms and acronyms, often overlapping with general English terms, making our definition of a "biomedical term" quite broad. That, paired with inaccuracies of the QuickUMLS concept extraction, makes it difficult to accurately subset biomedical terms. Moreover, further analysis is needed to analyze the change characteristics of different categories of biomedical terms at a fine-grained level, such as by semantic type. We hypothesize that different types of biomedical terms will exhibit different change characteristics, but such an analysis is beyond the scope of this work.

Future directions may include moving from word2vec to more contextualized embeddings, as recent work suggests that they may be applicable to this task.[13] Furthermore, more work is needed to integrate all aspects of clinical language change over time (such as shifts in grammar, morphology, or vocabulary). We believe a holistic quantification of change in clinical language necessitates a multi-faceted approach.

Acknowledgment. This study was funded by NCATS U01TR002062.

References

[1] Roberts A. Language, structure, and reuse in the electronic health record. AMA Journal of Ethics. 2017;19(3):281–288.

[2] Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351–1352.

[3] Helman CG. Doctors and Patients-An Anthology. CRC Press; 2018.

[4] Zeng QT, Redd D, Divita G, Jarad S, Brandt C, Nebeker J. Characterizing clinical text and sublanguage: A case study of the VA clinical notes. J Health Med Informat S. 2011;3:2.

[5] Ferguson G. 13. In: English for Medical Purposes. John Wiley & Sons, Ltd; 2012. p. 243–261.

[6] Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of Biomedical Informatics. 2002;35(4):222–235.

[7] Fitch WT. Empirical approaches to the study of language evolution. Psychonomic Bulletin & Review. 2017;24(1):3–33.

[8] Vashisth G, Voigt-Antons JN, Mikhailov M, Roller R. Exploring diachronic changes of biomedical knowledge using distributed concept representations. In: Proc. 18th BioNLP Workshop and Shared Task; 2019. p. 348–358.

[9] Hamilton WL, Leskovec J, Jurafsky D. Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics; 2016. p. 1489–1501.

[10] Kim Y, Chiu YI, Hanaki K, Hegde D, Petrov S. Temporal analysis of language through neural language models. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. Baltimore, MD, USA: Association for Computational Linguistics; 2014. p. 61–65.

[11] Kulkarni V, Al-Rfou R, Perozzi B, Skiena S. Statistically significant detection of linguistic change. In: Proceedings of the 24th International Conference on World Wide Web; 2015. p. 625–635.

[12] Jawahar G, Seddah D. Contextualized diachronic word representations. In: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. Florence, Italy: Association for Computational Linguistics; 2019. p. 35–47.

[13] Martinc M, Kralj Novak P, Pollak S. Leveraging contextual embeddings for detecting diachronic semantic shift. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2020. p. 4811–4819.

[14] Koplenig A. Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. Digital Scholarship in the Humanities. 2017;32(1):159–168.

[15] Ehrmann M, Colavizza G, Rochat Y, Kaplan F. Diachronic evaluation of NER systems on old newspapers. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). CONF. Bochumer Linguistische Arbeitsberichte; 2016. p. 97–107.

[16] Jatowt A, Duh K. A framework for analyzing semantic change of words across time. In: IEEE/ACM Joint Conference on Digital Libraries. IEEE; 2014. p. 229–238.

[17] Jassem K, Skórzewski P. Processing historical texts with contemporary NLP tools. Proceedings of the 8th Language and Technology Conference. 2017;p. 152–157.

[18] Lagunes-García G, Rodríguez-González A, Prieto-Santamaría L, del Valle EPG, Zanin M, Menasalvas-Ruiz E. How Wikipedia disease information evolve over time? an analysis of disease-based articles changes. Information Processing & Management. 2020;57(3):102225.

[19] Grigonyte G, Rinaldi F, Volk M. Change of biomedical domain terminology over time. In: Baltic HLT; 2012. p. 74–81.

[20] Oliver DE, Shahar Y, Shortliffe EH, Musen MA. Representation of change in controlled medical terminologies. Artificial Intelligence in Medicine. 1999;15(1):53–76.

[21] Cimino J. High-quality, standard, controlled healthcare terminologies come of age. Methods of Information in Medicine. 2011;50(02):101–104.
[22] Wang AY, Barrett JW, Bentley T, et al. Mapping between SNOMED RT and Clinical Terms version 3: a key component of the SNOMED CT development process. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001. p. 741.
[23] Pettersson E, Megyesi B, Nivre J. Parsing the past: identification of verb constructions in historical text. In: Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Association for Computational Linguistics; 2012. p. 65–74.
[24] Bollmann M, Søgaard A. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee; 2016. p. 131–139.
[25] Allauzen A, Gauvain JL. Diachronic vocabulary adaptation for broadcast news transcription. In: 9th European Conference on Speech Communication and Technology; 2005. p. 1305–1308.
[26] Harris ZS. Distributional structure. Word. 1954;10(2-3):146–162.
[27] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems; 2013. p. 3111–3119.
[28] Li J, Jurafsky D. Do multi-sense embeddings improve natural language understanding? In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics; 2015. p. 1722–1732.
[29] Iacobacci I, Pilehvar MT, Navigli R. SensEmbed: learning sense embeddings for word and relational similarity. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); 2015. p. 95–105.
[30] Mancini M, Camacho-Collados J, Iacobacci I, Navigli R. Embedding words and senses together via joint knowledge-enhanced training. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: Association for Computational Linguistics; 2017. p. 100–111.
[31] Huang EH, Socher R, Manning CD, Ng AY. Improving word representations via global context and multiple word prototypes. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics; 2012. p. 873–882.
[32] Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA; 2010. p. 45–50.
[33] Kutuzov A, Øvrelid L, Szymanski T, Velldal E. Diachronic word embeddings and semantic shifts: a survey. In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018. p. 1384–1397.
[34] Smith SL, Turban DH, Hamblin S, Hammerla NY. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv preprint arXiv:170203859. 2017;.
[35] Schönemann PH. A generalized solution of the orthogonal Procrustes problem. Psychometrika. 1966;31(1):1–10.
[36] Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. Nature. 1998;393(6684):440–442.
[37] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004;32(Database issue):D267–D270.
[38] Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, SIGIR; 2016. p. 1–4.
[39] North F, Crane SJ, Chaudhry R, et al. Impact of patient portal secure messages and electronic visits on adult primary care office visits. Telemedicine and e-Health. 2014;20(3):192–198.
[40] Shellum JL, Nishimura RA, Milliner DS, Harper Jr CM, Noseworthy JH. Knowledge management in the era of digital medicine: a programmatic approach to optimize patient care in an academic medical center. Learning Health Systems. 2017;1(2):e10022.
[41] Box GE, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological). 1964;26(2):211–243.
[42] Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9(Nov):2579–2605.

Preferential Mixture-of-Experts: Interpretable Models that Rely on Human Expertise As Much As Possible

Melanie F. Pradier, PhD[1,2,*], Javier Zazo, PhD[1,2,*], Sonali Parbhoo, PhD[1,*],

Roy H. Perlis, MD MSc[3,4], Maurizio Zazzi, MD[5], Finale Doshi-Velez, PhD[1]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA;
[2]Health Intelligence, Microsoft Research, Cambridge, Cambridgeshire, UK;
[3]Center for Quantitative Health, Massachusetts General Hospital, Boston, MA, USA;
[4]Harvard Medical School, Boston, MA, USA [5]University of Sienna, Italy.

Abstract *We propose Preferential MoE, a novel human-ML mixture-of-experts model that augments human expertise in decision making with a data-based classifier only when necessary for predictive performance. Our model exhibits an interpretable gating function that provides information on when human rules should be followed or avoided. The gating function is maximized for using human-based rules, and classification errors are minimized. We propose solving a coupled multi-objective problem with convex subproblems. We develop approximate algorithms and study their performance and convergence. Finally, we demonstrate the utility of Preferential MoE on two clinical applications for the treatment of Human Immunodeficiency Virus (HIV) and management of Major Depressive Disorder (MDD).*

1 Introduction

In the last few years, there has been a growth in the use of machine learning (ML) methods for decision-making in complex domains such as loan approvals, medical diagnosis and criminal justice. In particular, ML currently plays a key role in the healthcare sector for several tasks such as developing medical procedures [1, 2], handling patient data and records [3] and treating chronic diseases [4]. However, these algorithms typically require large amounts of data to make reasonable predictions. Additionally in the health sector, variability in practice between clinicians, patient heterogeneity, different disease prevalences, and confidentiality issues all result in final training cohorts being relatively small. Moreover, a clinician is often faced with rare events or outlier cases, where classic ML approaches suffer from insufficient training samples. In each of these scenarios, it is crucial to be able to incorporate clinical experience and domain knowledge.

Specifically, in practice, clinicians often rely on relatively simple human-based rules that reflect reasonable approaches to handle a situation. These rules can be seen as an additional source of knowledge that can be leveraged when building ML systems for clinical decision-support. For instance, clinicians treating patients with HIV tend to adhere to a list of guidelines for administering first and second-line therapies specified by several organizations [5, 6]; other well-known guidelines exist for prescribing antidepressants to address Major Depressive Disorder (MDD) [7]. Often, these rules provide benefits that are not easily formalized into a machine learning objective, for example, in terms of safety [8], or tolerability [9] (e.g., not giving excitatory drugs to a patient that has insomnia). Thus, one might prefer an ML system that agrees with these human-based rules as much as possible.

Several ML methods have been proposed that combine human expertise in conjunction with training data to perform a prediction task [10, 11]. Some of these methods such as [12] explicitly focus on modeling the *interaction* between an automated ML model and an external decision-maker; the decision-maker determines whether to reject a particular decision made by the model based on the model's confidence and the expertise of the decision-maker. An extension to this procedure in [10] describes when to defer decisions to a downstream decision-maker based solely on samples of the expert's decisions. In contrast to these approaches, we propose a ML system that complements human expertise only when needed, that is, it gives preference to human-based rules as much as possible, subject to explicit performance constraints in the optimization problem.

In this work, we develop a novel mixture-of-experts (MoE) approach, called *Preferential MoE*, that explicitly incorporates human expertise in learning to provide predictions that align with human-based rules as frequently as possible

*Equal contribution.

without losing performance. The MoE framework allows for an intuitive way to combine ML with clinical expertise. Importantly, Preferential MoE provides a means of enforcing preference for the human decision rules, as well as an interpretable gating function that allows us to understand when data-driven or clinical expertise should be used. Specifically, we identify when a human decision rule should be followed, and when it makes more sense to provide an alternative data-driven prediction. Overall, by explicitly incorporating and optimizing for human expertise in our predictions, we obtain models that aligns better with human knowledge, making them easier to inspect, audit and trust.

2 Related Work

Human-ML decision making systems. There is a long history of approaches to incorporate human expertise in the architecture of ML systems. In particular, [13] and [14] propose methods that map rules to elements of a neural network. [15] incorporates human-based knowledge gates into Recurrent Neural Networks for question answering or text matching. Closer in spirit, [16] constrained a ML model to be more credible by relying as much as possible on input predictors that are intuitive for human experts. All these approaches include human expertise as input or intermediate features, whereas we assume that the expert information is available in the form of output decision rules, on which we want to rely as much as possible. Recently, [17] learns a ML system complementary to humans by modeling the residual of humans in the context of timeseries. Here we focus on classification, and additionally provide an interpretable explanation about when to rely on human-based rules. Finally, [18] proposes a knowledge distillation approach, where human decisions are used as a teacher, and a student network is trained to mimic the human decisions while performing well on test data. Unlike implicitly assuming human expertise as additional ground truth labels (teacher), this work has the capacity of ignoring human rules if those are found unreliable.

Mixture of Experts. In the ML community, mixture-of-expert (MoE) models [19, 20] are frequently used to leverage different types of expertise in decision-making. The model works by explicitly learning a partition of the input space such that different regions of the domain may be assigned to different specialized sub-models or experts. MoEs have also been applied to several healthcare domains such as HIV [21, 22]. The proposed approach Preferential MoE is different in three regards: first, we explicitly incorporate human knowledge in the form of therapy standards and guidelines for medical decision-making. Second, our framework expresses an explicit preference for a specific expert (human-based), and trains an ML-based expert to complement the human expert; third, we learn a gating function, which makes the model easy-to-interpret and give us information on when human-based rules should be followed.

Learning to defer approaches. [12, 10] propose MoE classification models to be used as triage tools, where only the most critical decisions are deferred to a medical expert, whilst relying on data-driven approaches the majority of the time. Specifically, these classifiers are trained based solely on the samples of an expert's decisions. Other approaches for integrating human expertise in decision-making such as [4, 23] train a standard classifier on the data and subsequently obtain uncertainty estimates based on this classifier and the human expert. The decision is ultimately deferred to the expert with the lowest uncertainty. Unlike triage methods, we view human expertise as *complementary* to data-driven approaches and explicitly leverage these sources of knowledge to inform better predictions. That is, we optimize to rely on human expertise as much as possible, except for those regions for which human-based rules are inadequate. Our training samples consist of generic (potentially partial) rules that have been specified a priori.

3 Methodology

In this section, we present Preferential MoE. The proposed approach fulfills two desiderata. First, Preferential MoE relies on the human rules as much as possible while preserving predictive performance. When the human-based rules are damaging w.r.t the prediction task, the proposed approach is able to overrule them (that is, we recover the same solution as the unconstrained standard MoE formulation). Second, the gating function is interpretable, providing information on when each human guideline is applicable.

An overview of how Preferential MoE operates is illustrated in Figure 1. Colors in columns b)-d) represent predictive decision boundaries. In the proposed example, the human-based rule predicts red everywhere in the input space (sketch b). The third column (sketch c) shows the final predictions (colors) for each region of the input space. Each prediction either comes from the human decision rule, or from a data-based ML classifier. We learn a gating function (highlighted in purple) to select which classifier to rely on, as well as a complementary ML classifier to make predictions in regions

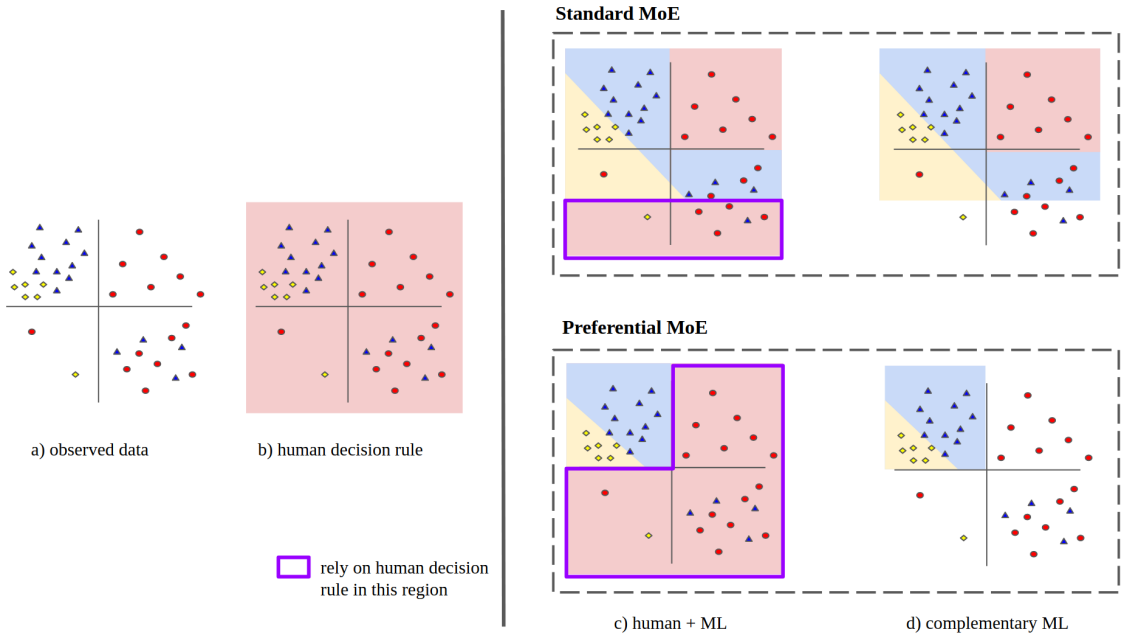


Figure 1: Preferential MoE: a mixture-of-experts (MoE) approach that relies on human decision rules subject to performance guarantees, and only disagrees with humans when a data-driven model can do better. A standard MoE exhibits similar predictive performance, but relies less on human rules. The proposed approach also provides insights on when/why human rules should be followed or not via an interpretable gating function (region highlighted in purple).

outside of the purple region. In this diagram, both the standard MoE and the preferential MoE exhibit same predictive performance; however, the preferential MoE relies on humans much more often.

More formally, let $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ be a dataset of observations where $\mathbf{x}_n \in \mathbb{R}^d$ are the covariates and $y_n \in \{1, ..., K\}$ is a categorical outcome for a specific prediction task. Let the *guideline* function $g : \mathbb{R}^d \to \{1, 0, -1\}$ be an aggregated function encoding all available human decision rules, whose input are the covariates, and the output might be any output category, or a flag (-1) indicating that the rule is not applicable (human does not know). This human guideline function g is fixed a priori by domain knowledge or well-established medical practice. In the case of not having access to an explicit human function g, but samples of past human decisions instead, we can pre-train a classifier to mimic those human decisions beforehand, and use such classifier as our human-based rules function g.

Given dataset $\mathcal{D}$, there might exist several functions that exhibit similar predictive performance, but are qualitatively different. We want to use expert knowledge (via human-based rules) to guide the optimization such that we are able to find models that have high predictive performance and agree with the human-based rules as much as possible. In order to accomplish that, we will include both objectives in the proposed optimization.

Modeling. Our goal is to make predictions that prioritize human-based rules when the data supports (or does not contradict) such knowledge, and learn to defer to another trainable ML expert when the human rules counters empirical evidence. For that, we propose a new classification model based on a mixture of experts formulation. Let $f_\theta : \mathbb{R}^d \to \Delta^K$ be a trainable ML expert parameterized by θ, where Δ^K denotes the $(K-1)$-simplex (outcome vectors of f_θ should sum to one). Our approach combines the predictions of the ML expert f_θ and the human expert g via the gating function $\rho_w : \mathbb{R}^d \to \{0, 1\}$ parametrized by w; ρ_w is another classifier that selects which expert to rely on given the covariates $\mathbf{x}$. The prediction model of Preferential MoE is formalized as

$$\hat{y}_{\theta,w}(\mathbf{x}) = \begin{cases} (1 - \rho_w(\mathbf{x}))f_\theta(\mathbf{x}) + \rho_w(\mathbf{x})g(\mathbf{x}) & \text{if } g(\mathbf{x}) \neq -1 \\ f_\theta(\mathbf{x}) & \text{if } g(\mathbf{x}) = -1 \end{cases} \tag{1}$$

where $\hat{y}_{\theta,w}(\mathbf{x}) \in \Delta^K$, and the likelihood function is given by

$$y|\mathbf{x} \sim \text{Categorical}\left(\hat{y}_{\theta,w}(\mathbf{x})\right) \tag{2}$$

Note that the ML expert f_θ might make predictions and specialize in input regions where the human rule g is not applicable or is inaccurate. In summary, $y|\mathbf{x}$ assumes that every data point $\mathbf{x}$ can be discriminated by f_θ or g, and ρ_w makes a deterministic decision on which expert to rely on. When learning ρ_w, we will prioritize human-based rules g during inference. Notice that (1) produces a non-convex prediction model which may be difficult to optimize.

The gating function selects when a decision should rely on human-based rule or a trained expert. By making ρ_w an interpretable function, e.g., a linear classifier or a decision tree, the model learns which features are important for human-based decision making, and identifies the regions of other expert classifiers. We note that, even if the gating function ρ_w is chosen to be interpretable, our approach does not provide theoretical guarantees on identifying all the regions suitable for human decision rules. More generally, ρ_w can also be a non-interpretable function, e.g., a neural network. In such case, the gating function still identifies regions appropriate for human-based decisions, although it may miss the interpretability of the parameters w. Overall, our framework allows model constructions that balance flexibility and interpretability suitable to different applications.

Problem formalization. Our formulation as an optimization problem needs to reflect the following criteria: (i) we want to minimize the predictive error, (ii) we want to follow human-based rules as frequently as possible without hurting performance.

We optimize for predictive performance by minimizing the cross-entropy $L^\gamma_{\theta,w}(\mathcal{D})$ with respect to the predictions from Equation (1); this corresponds to a standard maximum log-likelihood estimator for the probabilistic model $y|\mathbf{x}$ with an additional regularizer. For example, if the outcome is binary we write

$$L^\gamma_{\theta,w}(\mathcal{D}) = \sum_{n=1}^{N} -\left[y_n \ln(\hat{y}_{\theta,w}(\mathbf{x}_n)) + (1-y_n)\ln(1-\hat{y}_{\theta,w}(\mathbf{x}_n))\right] + \gamma||w||_1, \tag{3}$$

where $\gamma \geq 0$ is a regularization weight that controls the trade-off between predictive performance and sparsity of w. A sparse w can help identify important features for the gating function $\rho_w(x)$.

We bound the cross-entropy loss $L^\gamma_{\theta,w}(\mathcal{D})$ with a prefixed optimized value for performance guarantees. Denote $L^\gamma_{\theta^*,w^*}(\mathcal{D})$ an attainable loss where θ^* and w^* are solutions of minimizing $L^\gamma_{\theta,w}(\mathcal{D})$ for the stated MoE in Equation (2). Consider a margin $\varepsilon \geq 0$ measuring an acceptable performance decrease, and consider the constraint:

$$L^\gamma_{\theta,w}(\mathcal{D}) \leq (1+\varepsilon) L^\gamma_{\theta^*,w^*}(\mathcal{D}). \tag{4}$$

Equation (4) guarantees that the performance loss will not increase more than specified, and will maintain predictive error results. We introduce sets $\Theta \subset \mathbb{R}^q$ and $W \subset \mathbb{R}^p$ such that $\theta \in \Theta$ and $w \in W$. Variables θ and w do not need to have same dimensions and can be constructed with different model classifiers.

We present next the problem formulation for Preferential MoE:

$$\mathcal{G}: \quad \begin{array}{lll} \text{(player 1)} & \min\limits_{w\in W} & -\Sigma_{n=1}^{N} \ln(\rho_w(x)) \\ & \text{s.t.} & L^\gamma_{\theta,w}(\mathcal{D}) \leq (1+\varepsilon) L^\gamma_{\theta^*,w^*}(\mathcal{D}) \end{array} \qquad \begin{array}{lll} \text{(player 2)} & \min\limits_{\theta\in\Theta} & L^\gamma_{\theta,w}(\mathcal{D}). \end{array} \tag{5}$$

We refer to (5) as game $\mathcal{G}$. Using game theory terminology, there are 2 players and each player optimizes their own objective, variables and constraints, while taking into account the other player's decisions. Notice that $\mathcal{G}$ explicitly models our discussed goals: player 1, which is optimizing the gating function, maximizes the number of human-based decisions; player 2, which optimizes the classifier f_θ, minimizes prediction error according to the loss function (3). Note that the negative logarithm is a monotone transformation that helps obtain a convex objective for player 1. Player 1 also imposes the performance constraint and limits the classification loss.

4 Inference Algorithms

Inference for determining θ and w from $\mathcal{G}$ proceeds in two steps:

1. **Unconstrained optimization**: we train a standard MoE model from Equation (2) by minimizing the performance loss $L^{\gamma}_{\theta,w}(\mathcal{D})$ described in Equation (3). This step yields a performance reference value of $L^{\gamma}_{\theta^*,w^*}(\mathcal{D})$ which we will aim to maintain up to a certain margin ε. We use the optimal parameters θ^* and w^* from the unconstrained problem as warm initialization for the next step.
2. **Constrained optimization**: we solve game $\mathcal{G}$ initializing from previous solution.

We discuss two algorithms for solving $\mathcal{G}$. The first proposal combines both objectives and uses a log-barrier method to approximate a solution. The second proposal takes gradient steps that minimize each objectives alternatively and projects to the feasible region. Both methods have convergence guarantees.

Log-Barrier Method. We want to approximate a solution of $\mathcal{G}$ by simplifying its formulation. We move player 1's constraint to the objective using a log-barrier penalty used in interior point methods [24, Chapter 11], and we get

$$\min_{\theta\in\Theta, w\in W} \quad -t\sum_{n=1}^{N}\ln\left(\rho_w(x_n)\right) - \ln\left((1+\varepsilon)L^{\gamma}_{\theta^*,w^*}(\mathcal{D}) - L^{\gamma}_{\theta,w}(\mathcal{D})\right). \tag{6}$$

The first term of equation (6) corresponds to player 1's objective, and the second term to the log-barrier function $\widehat{I}(u) = -1/t\ln(-u)$ transforming its constraint, which also aligns with player 2's objective.

The log-barrier argument is susceptible of becoming negative inside the logarithm and be a source of numerical instability, so care needs to be taken with step sizes and correct initialization (warm-start). Parameter t is a hyperparameter that weights the satisfiability of the constraint, and the approximation improves as t grows. Note that this approximated form encourages that the difference $L^{\gamma}_{\theta^*,w^*}(\mathcal{D}) - L^{\gamma}_{\theta,w}(\mathcal{D})$ becomes large, regardless of the constraint already being satisfied. This has the desirable effect of continuously minimizing $L^{\gamma}_{\theta,w}(\mathcal{D})$. Finally, because of the non-convex nature of the problem, gradient descent methods only guarantee convergence to stationary solutions.

Projected Gradient Method. Player 2's decisions affect player 1's constraint, and player 1's affect player's 2 objective in game $\mathcal{G}$. A simple algorithm would be to alternate solving subproblems and repeat until convergence. Such schemes are only guaranteed to converge under very stringent conditions of monotonicity of the game. Monotonocity is a desirable property of multivariate mappings, informally stating that a small change in the input guarantees a bounded change in the output, therefore permitting dynamics of control towards stable solutions. We refer the reader to [25] for definitions, properties and algorithms for solving monotone games.

We present Algorithm 1 for solving $\mathcal{G}$. The algorithm makes a gradient update on each objective, and projects the result onto the feasibility region. We denote estimates on iteration k with θ^k and w^k. The feasibility region is denoted with $K_\varepsilon(\theta^{k+1})$, and is formally introduced in the appendix.* The operation $\Pi_{K_\varepsilon(\theta^{k+1})}$ denotes projection of w onto the set $K_\varepsilon(\theta^{k+1})$. The projection operation solves the following optimization problem $\Pi_{K_\varepsilon(\theta^{k+1})}(z) = \arg\min_{w\in K_\varepsilon}\frac{1}{2}\|w-z\|^2$, whose solution can be efficiently computed via a bisection search, described in Algorithm 2. The optimization inside the while loop in Algorithm 2 can be solved via L-BFGS [26].

Algorithm 1: Projected Gradient Descent

Input: $\mathcal{D}$, ε, $L^{\gamma}_{\theta^*,w^*}$, $\{\alpha^k\}$
Output: θ and w.
Initialization: $\theta^0 \leftarrow \theta^*$, $w^0 \leftarrow w^*$, $k \leftarrow 0$;
while *stopping criteria not satisfied* **do**
  $\theta^{k+1} \leftarrow \theta^k - \alpha^k \frac{\partial}{\partial\theta} L^{\gamma}_{\theta^k,w^k}$;
  $w^{k+1} \leftarrow \Pi_{K_\varepsilon(\theta^{k+1})}\left(w^k - \alpha^k \frac{\partial}{\partial\theta} L^{\gamma}_{\theta^{k+1},w^k}\right)$;
  $k \leftarrow k+1$
end

Algorithm 2: $\Pi_{K_\varepsilon(\theta^{k+1})}$ (bisection search)

Input: θ^{k+1}, ε, $z \in \mathbb{R}^p$, $\overline{\lambda}$; Initialization: $\underline{\lambda} \leftarrow 0$
Output: w^{k+1}.
while $(\overline{\lambda} - \underline{\lambda}) \geq$ *tolerance* **do**
  $\lambda \leftarrow (\overline{\lambda} + \underline{\lambda})/2$;
  $w \leftarrow \arg\min_{w\in W} \quad \frac{1}{2}\|w-z\|^2 + \lambda L^{\gamma}_{\theta,w}(\mathcal{D})$;
  if $L^{\gamma}_{\theta,w}(\mathcal{D}) - (1+\varepsilon)L^{\gamma}_{\theta^*,w^*}(\mathcal{D}) > 0$ **then** $\underline{\lambda} \leftarrow \lambda$;
  else $\overline{\lambda} \leftarrow \lambda$;
end

*The appendix is available at: https://arxiv.org/abs/2101.05360

5 Results

We compare the performance of Preferential MoE against several baselines for two medical tasks for the treatment of Human Immunodeficiency Virus (HIV), or pharmacological management of Major Depressive Disorder (MDD). Our baselines include using predictions a) based on a human expert alone; b) a logistic regression ML expert alone; c) a standard mixture-of-experts model (standard MoE); d) the learn-to-defer model in [12]; and e) a learn-to-defer model from [10]. For the standard MoE and Preferential MoE, we train models either assuming discrete $\rho(x)$ values to begin with, or assuming continuous $\rho(x)$ values and the discretizing at the end, exploring all operating points for the threshold of the gating function. Here we report the latter, which seems to work better in practice.

Hyperparameter selection. For both prediction tasks, we explore different learning rates for both, the unconstrained and constrained optimization steps, in the range of $\{10^{-4}, 10^{-3}, 0.01, 0.1\}$. We also explore a range of regularization parameters $\gamma \in \{0.0, 0.001, 0.01, 0.05, 0.1, 1.0\}$ for the gating function, and select those that maximize predictive performance in a validation set. For the psychiatry dataset, we additionally regularize the ML classifier with an L1 penalty to avoid overfitting due to the high-dimensionality of the input space. We fix the margin $\varepsilon = 0.1$, and the trade-off parameter $t = 5.0$ for the log-barrier penalty in Equation (6). Our results were stable to perturbations of these parameters. Intuitively, t can be matched to existing interior-point algorithms and is quite robust with appropriate gradient step sizes. The margin ϵ affects model´s accuracy, but even if there is no direct mapping from its value to a desired performance level, its impact was similar in the range $\epsilon \in [1e^{-2}, 2e^{-1}]$. Setting ϵ too small can make the model not move from the initialization point, and its solution stay similar to the standard MoE's.

Evaluation metrics. To evaluate Preferential MoE and other baselines, we measure performance as Area-Under-the-operating-ROC-Curve (AUC), as well as predictive accuracy (percentage of correct predictions) for a fine-grid of threshold values, both for the gating function and final predictions. Note that all thresholds are chosen by cross-validation, we thus guarantee that the right thresholds (w.r.t the most adequate metric for each downstream task) are selected, in a data-driven manner. We report *coverage* as a measure of how frequently (in percentage) each model relies on the human-based guideline function g. More specifically, we define *soft-coverage* as $100.0 \times \mathbb{E}[\rho(x)]$ and *hard-coverage*(t) as $100.0 \times \mathbb{E}\big[\mathbb{1}[\rho(x) \geq t]\big]$ for a given gating function threshold t.

5.1 Human Immunodeficiency Virus (HIV) Therapy Outcome Prediction

HIV currently affects more than 36 million people worldwide. The life-long use of combinations of antiretrovirals has largely helped combat the virus in most parts of the world and has transformed the virus from a life-threatening condition to a chronic illness. However, administering therapies is tricky as patients frequently suffer from drug resistance, viral relapses or spikes, as well as adherence issues and several other side-effects from use of antiretrovirals.

We identified individuals between 18-72 years of age from the EuResist database comprising of genotype, phenotype and clinical information of over 65 000 individuals in response to antiretroviral therapy administered between the years 1983 and 2018. We focus on a subset of 36 780 of these patients who received at least 3 prior treatments and base our predictions on the genotype, phenotype, clinical and demographic information of these individuals. The curated dataset contains a total of 384 such features. Our goal is to predict short-term therapy success where viral suppression is maintained for at least 40 days after a therapy is administered.

Table 1 shows predictive performance and coverage results for the proposed approach and competing baselines. Compared to other approaches, Preferential MoE exhibits highest soft coverage while either retaining or improving predictive performance. Figure 3 compares the accuracy relative to hard thresholding of the coverage for each of the MoE models. In the HIV setting, both variants of the Preferential MoE outperform the standard MoE approach at various coverage values. At 60% coverage, the methods all seem to perform relatively similarly in terms of accuracy.

Importantly, Preferential MoE allows us to incorporate human expertise into the prediction task and provides us with insights of when it makes sense to follow the rules based on the gating function. Table 2 provides a sparse list of predictors and corresponding weights averaged over 10 random seeds for the gating function. These predictors are associated with regions where it makes sense to follow human intuition. While Standard MoE identifies blood count

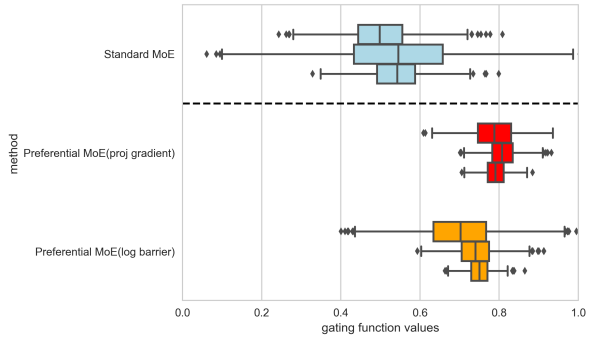


Figure 2: $\rho(x)$ values in the test set for HIV. Preferential MoE pushes up the values for the gating functions, favoring human decision rules more frequently in the input space. Each box plot corresponds to a different random seed (we report 3 different initializations per method).

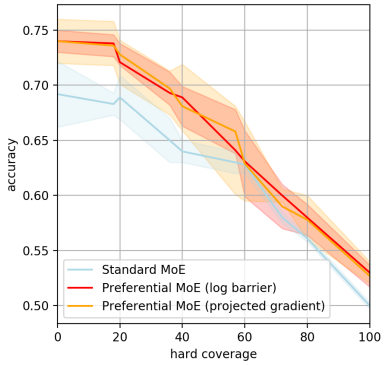


Figure 3: Accuracy-coverage trade-off. Preferential MoE (trained by the log barrier or projected gradient method) for HIV either relies more on human rules for the same predictive accuracy, or gets higher accuracy for the same coverage with human rules.

data, certain mutations and a patient's risk group as meaningful factors, Preferential MoE identifies a significantly different set of predictors. Notably, many of the predictors identified in the latter correspond to cases where patients have additional conditions such as lipodystrophy or side effects to medication where it is preferable to rely on human judgement to determine how to treat these individuals. Figure 2 compares the gating function values $\rho(x)$ in the test set for HIV. Unsurprisingly, Preferential MoE shows a higher preference for relying on human rules.

5.2 Prediction of Antipsychotic for Major Depressive Disorder (MDD)

Antidepressant prescription for MDD often involves trial and error. Roughly 2/3 of individuals diagnosed with MDD do not yield remission with their initial treatment, and 1/4 of patients is expected to dropout against clinical advice before finishing their treatment [27, 28]. The list of potential side-effects translates in tolerability and safety concerns that need to be taken into account while prescribing antidepressants. Here we focus on predicting prescription of antipsychotics, which is a class of medication primarily used to manage psychosis, but often used as an adjunctive treatment in the pharmacological management of MDD. The guideline function g for this prediction task is as follows: if the patient has anxiety or insomnia, promote antipsychotic (predict positive label), if the patient has overweight, avoid antipsychotic (predict negative label).

We identified individuals age 18-80 years drawn from the outpatient clinical networks of two academic medical centers in New England, Massachusetts General Hospital and Brigham and Women's Hospital. These patients had received

	AUC		soft coverage (%)	
Baselines	mean	CI	mean	CI
ML only	0.64	[0.63-0.65]	0.00	[0.00-0.00]
Learn-to-defer[12]	0.71	[0.68-0.72]	54.07	[48.18 - 55.63]
Consistent Learn-to-defer [10]	0.66	[0.62-0.69]	56.81	[50.02 - 57.62]
Standard MoE (unconstrained)	0.69	[0.69-0.70]	52.87	[51.19-54.55]
Preferential MoE (log barrier)	**0.74**	[0.72-0.76]	62.06	[60.8-63.32]
Preferential MoE (projected gradient)	**0.74**	[0.73-0.75]	**63.18**	[61.7-64.66]

Table 1: Performance vs Coverage (HIV): Preferential MoE relies much more often on human expertise while preserving predictive performance. Predictive performance measured by Area-Under-the-operating-ROC-Curve (AUC); Reliance on human decision rules based on soft coverage.

Weight w	Covariate Description
+0.1612 $\pm$ 0.014	CD8+ cell count (cells/ml)
-0.1161 $\pm$ 0.002	Reverse Transcriptase Mutation 67N
-0.0310 $\pm$ 0.025	Protease Mutation 20M
0.0280 $\pm$ 0.001	Blood count; complete (CBC)
-0.0195 $\pm$ 0.005	Co-infection of Hepatitis C
-0.0156 $\pm$ 0.001	Stavudine
-0.0124 $\pm$ 0.011	Reverse Transcriptase Mutation 215YF
+0.0121 $\pm$ 0.020	Nevirapine
-0.0068 $\pm$ 0.031	Risk group MSM
-0.0055 $\pm$ 0.005	Age

Weight w	Covariate Description
+0.0359 $\pm$ 0.022	CD4 + cell count (cells/ml)
-0.0236 $\pm$ 0.027	Baseline Viral Load
+0.0151 $\pm$ 0.030	High Adherence
+0.0150 $\pm$ 0.001	Number of Prior Treatment Lines
+0.076 $\pm$ 0.007	Pregnancy
-0.0055 $\pm$0.016	Reverse Transcriptase Mutation 184V
-0.0035 $\pm$0.002	Race black
-0.0026 $\pm$0.001	Lamivudine
+0.0025 $\pm$ 0.003	Anaemia
+0.0012 $\pm$ 0.007	Lipodystrophy

Table 2: Interpretation of gating function (HIV). Sparse list of predictors describing the regions where human decision rules are followed. We report weight parameters averaged across 10 different random seeds, and for regularization γ=0.1). (left) Standard MoE (predictors after step 1 in training); (right) preferential MoE (predictors after step 2 in training). Highlighted in red/green are those predictors that disappear/pop-up after step 2 in training.

at least one electronically-prescribed antidepressant between March 2008 and December 2017 with a diagnosis of MDD or depressive disorder at the nearest visit to that prescription. The goal is to predict prescription of antipsychotic based on demographic information (gender, race) as well as diagnostic and procedure codes. Race and gender were self-identified features and were included as a proxy for socio-economic variables. The curated dataset consists of 3,865 individuals and 1,680 features.

Table 3 shows predictive performance and soft coverage results for the proposed approach and competing baselines, averaged across 5 random initializations. We encountered issues training the Learn-to-defer approaches to this data (probably due to its high-dimensionality), so we only include the other baselines. Preferential MoE exhibits highest soft coverage (reliance on human rules) while maintaining (or even slightly improving) predictive performance.

	AUC		soft coverage (%)	
Baselines	mean	CI	mean	CI
ML only	0.70	[0.69-0.71]	0.00	[0.00-0.00]
Standard MoE (unconstrained)	0.71	[0.70-0.71]	31.41	[28.48-34.56]
Preferential MoE (log barrier)	**0.72**	[0.71-0.73]	**48.34**	[46.24-51.74]
Preferential MoE (projected gradient)	0.72	[0.71-0.72]	45.06	[42.85-46.70]

Table 3: Performance vs Coverage (Psychiatry): Preferential MoE relies much more often on human expertise while preserving predictive performance. Predictive performance measured by Area-Under-the-operating-ROC-Curve (AUC); Reliance on human decision rules based on soft coverage.

Preferential MoE gives us additional information on when to follow such human rules by inspecting the gating function. Table 4 presents the sparse list of predictors for the gating function, associated to regions where human decision rules are followed. By regularizing the gating function classifier with an L1-penalty, we get concise list of predictors to describe those regions. The list on the left correspond to Standard MoE (unconstrained optimization), and the list on the right correspond to Preferential MoE (constrained optimization maximizing reliance on humans). In both lists, most predictors corresponding to general patient care (examination, hospital care, etc) are negatively-correlated: this can be interpreted as higher reliance on humans in the absence of patient care related codes. In the case of Preferential MoE, additional covariates coding for cardiovascular risk factors (highlighted in green) are positively-correlated with reliance on human rules. Such information can be used to explore refinements of the human-based rules.

Figure 4 compares the histogram of the gating function values $\rho(x)$ in the test set. As expected, Preferential MoE pushes those values up, reflecting a preference for relying on human rules when possible. Although these values are continuous, we can discretize them using a specific threshold v calibrated on the validation set. Each threshold v

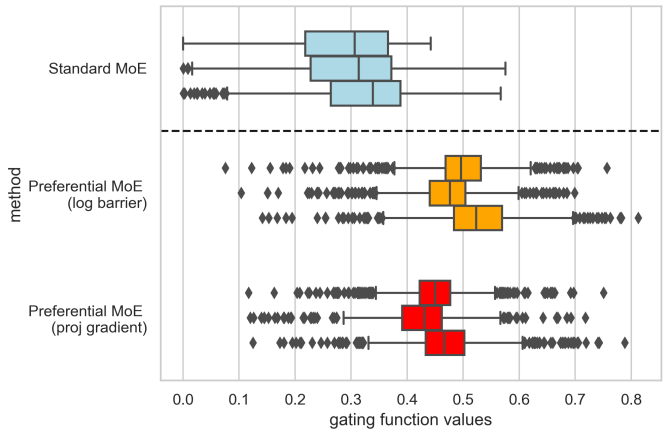


Figure 4: Histograms for $\rho(x)$ in the test set. Preferential MoE pushes up the values for the gating function, favoring relying on human decision rules more frequently. Each box plot corresponds to a different random seed (3 per method).

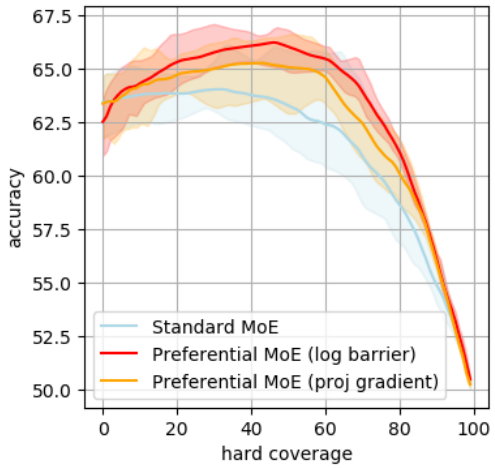


Figure 5: Accuracy-coverage trade-off. Preferential MoE either relies more on human rules for the same predictive accuracy, or gets higher accuracy for the same coverage with human rules.

Weight w	Covariate Description
-0.0303 $\pm$ 0.0143	Subsequent hospital care
-0.0242 $\pm$ 0.0152	MDD, recurrent episode
-0.0235 $\pm$ 0.0126	Psychiatric examination
-0.0208 $\pm$ 0.0071	Depressive disorder
-0.0153 $\pm$ 0.0106	Anxiety state
-0.0117 $\pm$ 0.0089	Office or outpatient visit
-0.0083 $\pm$ 0.0033	Radiologic examination
-0.0073 $\pm$ 0.0062	Trazodone
-0.0068 $\pm$ 0.0049	Emergency department visit
-0.0031 $\pm$ 0.0062	race white

Weight w	Covariate Description
-0.0202 $\pm$ 0.0043	Subsequent hospital care
-0.0115 $\pm$ 0.0075	MDD, recurrent episode
-0.0110 $\pm$ 0.0107	Psychiatric examination
0.0103 $\pm$ 0.0041	Office or outpatient visit
-0.0070 $\pm$ 0.0112	Depressive disorder
0.0068 $\pm$ 0.0022	General medical examination
0.0061 $\pm$ 0.0022	Type II diabetes
0.0037 $\pm$ 0.0016	Hypertension
-0.0035 $\pm$ 0.0046	Anxiety state
-0.0034 $\pm$ 0.0083	Trazodone

Table 4: Interpretation of gating function. Sparse list of predictors describing the regions where human decision rules are followed. We report weights averaged across 10 different random seeds, and for a regularization parameter γ=0.1). (left) Standard MoE (predictors after unconstrained step 1 in training); (right) Preferential MoE (predictors after step 2 in training). Highlighted in red/green are those predictors that disappear/pop-up after step 2 in training.

yields a different trade-off between accuracy and coverage. Figure 5 shows the trade-off between accuracy and hard coverage reachable by these models. As a reference point, the human decision rules have an accuracy of 49.87% for this prediction task. The curves are averaged over 10 different random seeds, each curve is obtained by changing the thresholds for the gating function and final decision. Overall, Preferential MoE is able to reach better trade-offs, either better accuracy for a given fixed hard coverage, or more hard coverage for a given accuracy level.

6 Conclusion

We presented *Preferential MoE*, a mixture of experts that learns and combines a ML general classifier with a human expert, prioritizing the human-based rules. We presented a game formulation of two objectives, which we solve by a log-barrier method or alternating projected gradient descent. We evaluate both approaches in the prediction of HIV therapy success, and prescription of antipsychotic for MDD. Both algorithms preserve performance and maximize coverage of human-based decisions compared to other baselines, assuming soft and hard decision assignments of the gating function. Future work will further explore other MoE formulations balancing performance and global optimality of the MoE formulation.

References

[1] K. Hamid et al. "Machine learning with abstention for automated liver disease diagnosis". In: *2017 International Conference on Frontiers of Information Technology (FIT)*. IEEE. 2017, pp. 356–361.

[2] A. Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), pp. 115–118.

[3] O. S. Pianykh et al. "Improving healthcare operations management with machine learning". In: *Nature Machine Intelligence* 2.5 (2020), pp. 266–273.

[4] M. Raghu et al. "The algorithmic automation problem: Prediction, triage, and human effort". In: (2019).

[5] W. H. Organization. *Tackling HIV drug resistance: trends, guidelines and global action*. Tech. rep. 2017.

[6] OARAC. "Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV". In: *Panel on Antiretroviral Guidelines for Adults and Adolescents* (2017).

[7] I. Lage et al. "Do clinicians follow heuristics in prescribing antidepressants?" In: *submitted* (2020).

[8] M. Stone et al. "Risk of suicidality in clinical trials of antidepressants in adults: analysis of proprietary data submitted to US Food and Drug Administration". In: *BMJ (Clinical research ed.)* 339 (2009).

[9] S. R. Blumenthal et al. "An electronic health records study of long-term weight gain following antidepressant use". In: *JAMA psychiatry* 71.8 (Aug. 2014). ISSN: 2168-6238.

[10] H. Mozannar and D. Sontag. "Consistent Estimators for Learning to Defer to an Expert". In: (2020).

[11] E. D. Gennatas et al. "Expert-augmented machine learning". In: *Proceedings of the National Academy of Sciences* 117.9 (2020), pp. 4571–4577.

[12] D. Madras, T. Pitassi, and R. Zemel. "Predict responsibly: improving fairness and accuracy by learning to defer". In: *Advances in Neural Information Processing Systems*. 2018, pp. 6147–6157.

[13] G. G. Towell and J. W. Shavlik. "Knowledge-based artificial neural networks". In: *Artificial Intelligence* (1994).

[14] S. N. Tran and A. S. d'Avila Garcez. "Deep Logic Networks: Inserting and Extracting Knowledge From Deep Belief Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 29.2 (Feb. 2018).

[15] Y. Wu et al. "Knowledge enhanced hybrid neural network for text matching". In: *AAAI Conference*. 2018.

[16] J. Wang et al. "Learning credible Models". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (July 2018). arXiv: 1711.03190.

[17] M. A. Chattha et al. "KINN: Incorporating Expert Knowledge in Neural Networks". In: (2019).

[18] Z. Hu et al. "Harnessing deep neural networks with logic rules". In: *arXiv:1603.06318* (2016).

[19] R. A. Jacobs et al. "Adaptive mixtures of local experts". In: *Neural computation* 3.1 (1991), pp. 79–87.

[20] M. I. Jordan and R. A. Jacobs. "Hierarchical mixtures of experts and the EM algorithm". In: *Neural computation* 6.2 (1994), pp. 181–214.

[21] S. Parbhoo et al. "Combining kernel and model based learning for hiv therapy selection". In: *AMIA Summits on Translational Science Proceedings* 2017 (2017), p. 239.

[22] S. Parbhoo et al. "Improving counterfactual reasoning with kernelised dynamic mixing models". In: *PloS one* 13.11 (2018), e0205839.

[23] B. Wilder, E. Horvitz, and E. Kamar. "Learning to Complement Humans". In: *arXiv:2005.00582* (2020).

[24] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Mar. 2004.

[25] G. Scutari et al. "Monotone games for cognitive radio systems". In: Springer, 2012.

[26] D. C. Liu and J. Nocedal. "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1-3 (1989), pp. 503–528.

[27] M. C. Hughes et al. "Assessment of a Prediction Model for Antidepressant Treatment Stability Using Supervised Topic Models". In: *JAMA Network Open* 3 (May 2020).

[28] M. F. Pradier et al. "Predicting treatment dropout after antidepressant initiation". In: *Tr. Psychiatry* (2020).

SAEgnal: A Predictive Assessment Framework for Optimizing Safety Profiles in Immuno-Oncology Combination Trials

Andrzej Prokop, PhD[1,2], Youyi Zhang, PhD[2], Pralay Mukhopadhyay, PhD[3, 4],

Faisal Khan, PhD[2], Khader Shameer, PhD[2*]

[1] Biometrics, Oncology R&D, AstraZeneca, Warsaw, Poland; [2] Data Science and Artificial intelligence, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD, USA; [3] Late ImmunoOncology Statistics, GMD, AstraZeneca, Gaithersburg, US; [4] Current Affiliation: Otsuka Biopharmaceuticals, Rockville, MD, USA

**Corresponding Author Email: shameer.khader@astrazeneca.com*

Abstract:

Combination therapies are an emerging drug development strategy in cancer, particularly in the immuno-oncology (IO) space. Many combination studies do not meet their safety objectives due to serious adverse events (SAEs). Prediction of SAEs based on evidence from single and combination studies would be highly beneficial. To address the emerging challenge of optimizing the safety and efficacy of combination studies, we have assembled a novel oncology clinical trial data set with 329 trials, 685 arms (279 unique treatment arms), including 200 combinations, 79 mono arms, and 59 curated adverse event categories in the setting of non-small cell lung cancer (NSCLC). We integrated the database with an analytical framework: SAEgnal. Using SAEgnal, we have investigated the difference in the risk of 39 adverse event types between combination and monotherapy arms across a subset of 34 combination trials. We observed different risk profiles between combination and monotherapies; interestingly, while the risk of elevated AST/ALT is lower in combination arms (in 1/8 trials, p-value < 0.05), it is higher for bleeding (7/8 trials, p-value < 0.05). We envisage that the SAEgnal framework would enable rapid predictive analytics of SAEs in oncology and accelerate drug development in oncology.

Introduction

Combination therapy is a dosing regimen where a single treatment arm contains more than one drug to achieve a higher therapeutic effect. With a rapidly changing landscape of immuno-oncology agents (IO, such as checkpoint inhibitors), more clinical trials attempt to assess the safety and efficacy of combination therapies. Such clinical trials are designed in such a way that IO therapies are either dosed together with a different IO or with chemotherapy. Such treatments have been proven to be effective in treating various cancers as higher efficacy can be achieved with a lower dosage of each drug and hence – reduce the risk of adverse events. A higher efficacy is typically achieved due to both drugs inhibiting cancer development or boosting immune responses by targeting different genomic pathways[1,2]. For example, a CTLA4 inhibitor dosed together with PDL1/PD1 inhibitor renders cancer cells to immune cells' cytotoxic activities by enhancing their functions and targeting them to cancer environmen[3]. Combination therapies have already been frequently approved across different indications, including breast cancer. However, the application of combination therapies is still a relatively new branch in Oncology, where questions concerning safety and efficacy are not always clear to answer. Despite the number of successful cases, many new therapies fail and do not meet their efficacy or safety milestones[4,5]. Drug-drug interactions may also affect the safety parameters, including pharmacokinetics and pharmacodynamics profiles of a therapeutic agent. A simple example could include a situation where both drugs target the same receptor, therefore, competing

over a single resource and bringing limited treatment benefit. Moreover, most studies fail to achieve their safety objectives due to unexpected toxicity effects among the investigated cohorts. Thus, at the phase of study design, a high degree of attention must be put while attempting to determine the target population, inclusion criteria, drugs used in a combination, and their dosing schedule.

The study-design level decision-making process typically relies on collective insights gained from historical data on similar studies. Similar successful combinations, doses, populations, and indications introduce a controlled environment where a single factor can be changed on a study basis. Such an approach eventually leads to the closing of a gap in knowledge by creating a landscape of possibilities[6]. Our work attempts to answer whether the construction and the analyses of publicly available clinical trial metadata can benefit decision making in drug development. Given the limited availability of the efficacy data, we primarily focused on studying safety profiles of chemo-IO combinations. We generated a dataset comparing different treatment arms across non-small cell lung cancer (NSCLC) to answer a question: *Can we use the historical data on IO and chemotherapy mono-arms to predict the risk of the adverse events in the combination treatment arms?* To answer these questions, we designed, developed, and evaluated a new analytical framework called SAEgnal.

Methods

In the next section we explain our strategy on data selection and collection. The following section introduces *SAEgnal,* our method of transforming the data and preparing it for the analysis. Finally, we summarize our analytical approach and highlight few interesting results.

Exploratory Data Analyses

Given the limited availability of dataset containing a summary of risks of adverse events across combination IO studies, the Data Resource Exploration was initiated to form an adverse events table generated based on publicly available data resources that provide sufficient information. Our approach was to use the TrialTrove database to search for all NSCLC studies where an IO drug was used (either in combination with chemotherapy or alone). TrialTrove contains a set of filters that enable a quick search of clinical databases to pre-select studies of interest. This search engine significantly decreases the amount of work needed to identify studies of particular indication or design. We also selected studies where a chemo monotherapy was used, and matched chemotherapeutic agents used in the chosen combo studies. Once the list of study decodes comprised of IO+chemo combos, IO mono arm studies, chemo mono arm studies, we then downloaded the Aggregate Analysis of ClincalTrials.gov (AACT) database from March 27, 2016, as a SAS CPORT export. AACT contains an extensive information on clinical studies, including inclusion criteria, demography, reported adverse events, endpoints data etc. We used the TrialTrove list of study identifiers matching ClinicalTrials.gov Identifiers (NCT ID) to search the AACT database for published numbers of cases of adverse events and cohorts' size. For the speed of processing, we have used a SAS-Grid server to extract the information on treatment arms of our interest including drug names, the number of subjects in a cohort, adverse event name, frequency of an adverse event in population, demography information, inclusion criteria information – entry age; squamous/non-squamous type of disease; the line of chemotherapy; CNS metastasis; disease stage. After the data export, in total, we have found 329 studies, 279 unique pairs of drugs, and 11 monotherapy IO studies serving as one of the control arms.

Data Curation

Adverse events published within AACT are typically not coded under standardized terms and allow the use of synonyms of adverse events (shivers, chills, as one of the examples). Given the low data volume, we targeted to avoid a situation where too many adverse events were considered distinct because it would inevitably result in the loss of information: discrete adverse events with a low coverage across the population. With the support of Medical Dictionary for Regulatory Activities Terminology (MedDRA), we have simplified our data set and grouped similar adverse events into distinct categories to enable interpretability. The criteria for groupings preserved medical meaning of a condition in the data set at the general level, e.g. limb or joint pain were grouped regardless of which limb was affected. In many cases, groupings were done by switching the order of words in the AE term; example "pain of a limb" became "limb pain'. In some cases, terms were grouped by their similar meaning: chills and shivers were grouped across all studies into a single word called "chills." Groupings were created with extensive use of regular expressions in Python. Similarly, the names of treatment arms were standardized across the dataset for further grouping. The names of drugs within treatment arms are often displayed under various synonyms within the AACT database, and a uniform convention was chosen for all drugs. Combo treatment arm names were constructed in such a way, that compounds were alphabetically listed and separated by a "+" sign. Therapies were finally mapped to treatments within combo arms and their corresponding monotherapies. Figure 1 portrays the data flow from Study IDs from the TrialTrove merged with AACT data set by NCT ID. Both treatment arm names and similar adverse events were standardized across the data set. In case of IO components reported under trade names or abbreviations we have chosen names with "mabs" suffix. In each combo treatment arm, we have used an alphabetical order for drugs, e.g. bevacizumab+cisplatin but not cisplatin+bevacizumab.

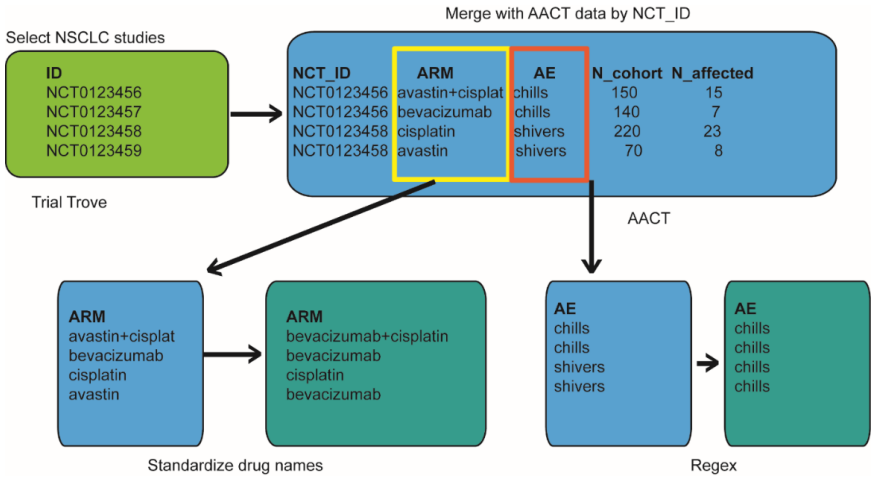


Figure 1. SAEgnal data exploration and curation strategy.

Data Analytics

The data curation described in the previous section enabled us to apply analytical methods further to explore the data. The primary goal of our analysis is to determine whether the risk of adverse events in combination treatment arms was different from an estimated combined risk within chemo and IO arms. To achieve this goal, we first summarized the distribution of adverse events across treatment arms (Figure 2A) independently across individual clinical studies. We also investigated the total counts of adverse events grouped by their types of irrespective of

the population size within different studies, to understand the overall prevalence of adverse events (Figure 2B).

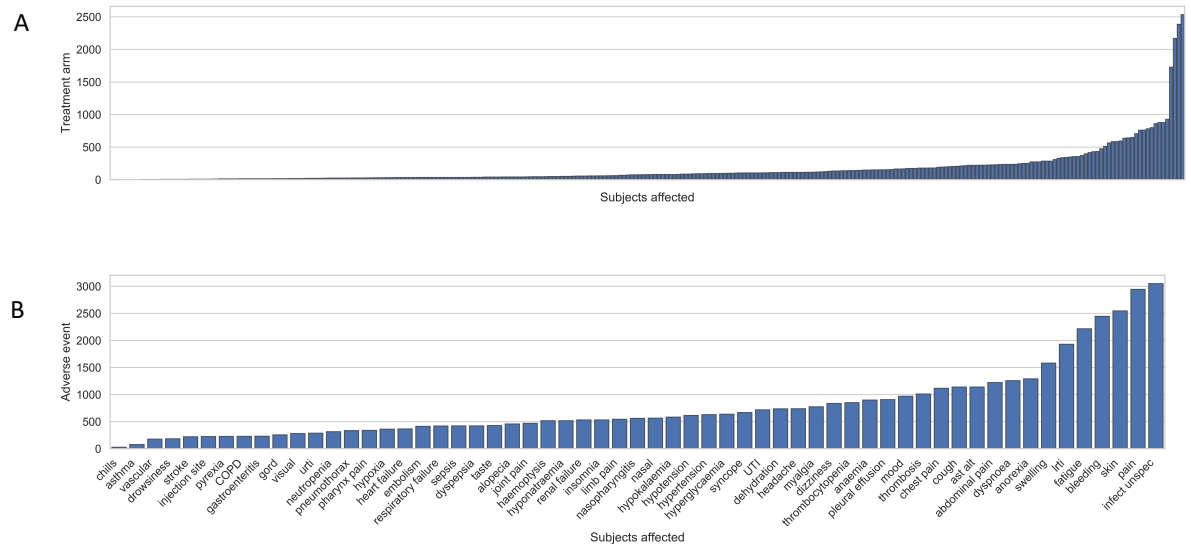


Figure 2. (A) Distribution of adverse events across 279 treatment arms in the data set. (B) Number of subjects affected by each AE across the data set.

Each adverse event in each treatment arm from a data set is represented by the number of subjects affected per total population enrolled into that arm. A significant analytical challenge lies in comparing the risk of an adverse event in a combination arm to single mono arms. The ideal scenario would be to have the same set of multiple studies where three treatment arms were designed: IO+Chemo arm, IO control arm, Chemo control arm. However, the situation where an IO drug is studied as a control within a mono arm is quite rare. Most clinical combination studies apply chemo as a control arm. To overcome this challenge, we have decided to leverage the data set containing available IO mono-arms across all available studies by calculating the risks of individual AEs in the pooled data irrespective of the type of IOs. Based on the evidence, IO agents, like proteins, mostly cause a similar range of immune-related adverse events[8]. The pooled risk of AEs in chemo monotherapies was calculated for each AE and each chemotherapeutic drug. The combined risk of an AE for IO and Chemo was estimated as $1 - (1 - \text{prob\_ae\_IO}) \times (1 - \text{prob\_ae\_Chemo})$. We could then answer a question of whether there is a statistically significant difference in the observed risk of AEs in combination treatment arms compared to an expected risk in mono arms. To assess this, we have used Python-based Scipy statsmodel API to build a 2x2 contingency table containing: an observed number of patients who suffered/not suffered a specific AE in a given combination of agents on a study; an expected number of AE/non-AE cases in that combination calculated by multiplying the cohort size by an estimated probability calculated before. We then performed the Fisher's exact test for each individual AE in each study for each combination treatment arm using $\alpha = 0.05$. Multiple testing correction was achieved with an α threshold of Bonferroni Correction p-value= 0.00084. In further analysis, we only present data for odds ratios significant after multiple testing correction (39 AEs across 34 studies and 29 treatment arms). The resulting table contained risk ratios together with their p-values and an upper and lower confidence interval. We selected AEs with the risk ratios p-value < 0.05 and the confidence intervals, not covering one for further analysis. The analytics strategy used in our analyses is described in Figure 3. Here, the null hypothesis assumed no difference between an observed number of cases ae/non-ae event in a combination regimen and an estimated number of cases of ae/non-ae. The estimated numbers were obtained by multiplying an estimated probability of

AE by the size of cohort in the combination study. The estimated probability of AE in combination was in turn calculated from chemo and IO mono arms. Probability of an AE in IO was calculated once for each AE irrespective of the type of IO. Probability of AE in chemo mono arm was, however, calculated for each separate chemotherapeutic drug. A resulting 2x2 contingency table allowed testing of H_0 by providing risk ratios, p-values and upper and lower confidence intervals. Risk ratios for AEs in combinations together with all the auxiliary variables used for their computations were saved in a data set. The data set also contains the information about inclusion/exclusion criteria for each study; line of chemotherapy; minimum recruitment age; disease stage; disease origin (squamous/non-squamous). The data set is available in a format which allows for an application of machine learning methods (decision trees, neural networks etc.) to predict the risk of particular AEs based on treatment labels and other variables. All data mapping and analytics were performed using Python 2.7 using Jupyter Notebook.

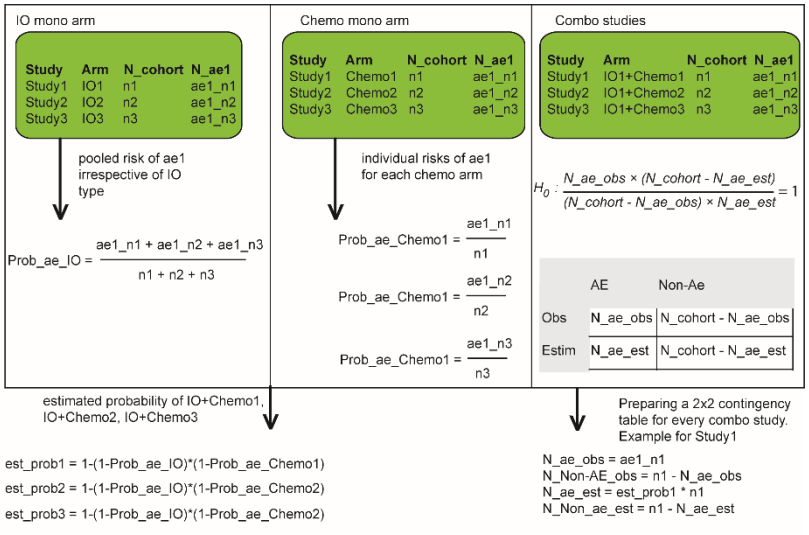


Figure 3. SAEgnal analytics strategy.

Results

The data integrated as part of the development of the initial version of SAEgnal framework discussed in this paper is available at the URL: https://doi.org/10.6084/m9.figshare.13483521. We visualized our result in depicting profiles of AEs across combination therapies with each point representing a risk ratio for a single-combination study on a logarithmic scale (Figure 4, Figure 5, Figure 6(A)). Collectively, these plots provide information on AEs' directionality, where the points grouped at the bottom side of the axis represent the clinical combination studies whose risk is lower than estimated and vice versa. We have investigated it further as a binary problem by counting the number of cohorts (treatment groups) per AE, where the observed risk was lower or higher than estimated (Figure 6 (B)). We have found that out of 39 significant AEs under study 28 placed themselves at the bottom side of the axis (meaning the risk was reduced compared to mono arms), 4 AEs were equally distributed between "high" and "low" risk and 7 AEs had an elevated risk in combination arms compared to mono arms. Interestingly, we have found that the risk of bleeding was increased in 7 studies cohorts/treatment combination arms, which is also the highest number in this data set. On the contrary, the risk of cough was reduced in 7 treatment arms under study. An interesting trend can also be observed concerning pain: risks of general pain, chest pain, abdominal pain, joint and limb pain is reduced in combination arms. In Figure 4, we are presenting data on two SAE types (bleeding and elevated AST and ALT).

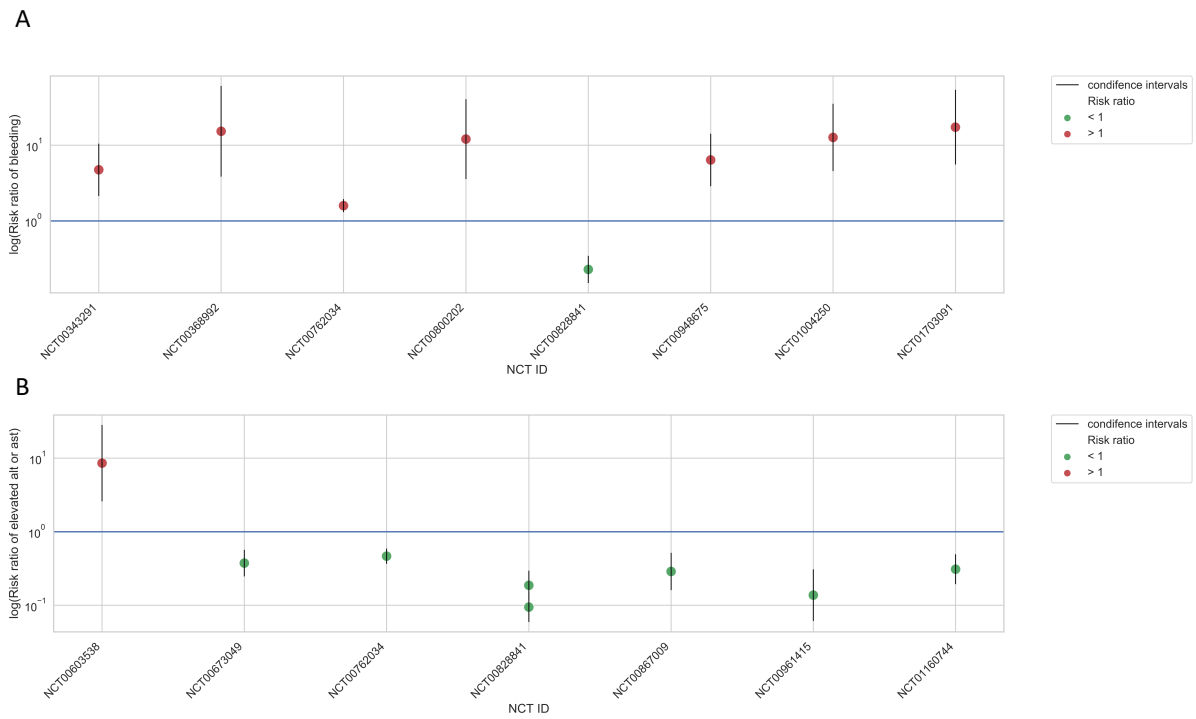


Figure 4. Example of statistical analysis performed for two AE types: bleeding (A) and elevated AST ALT (B). Risk ratios are represented on the y-axis. More than a single point per study indicates more than one treatment arm per study.

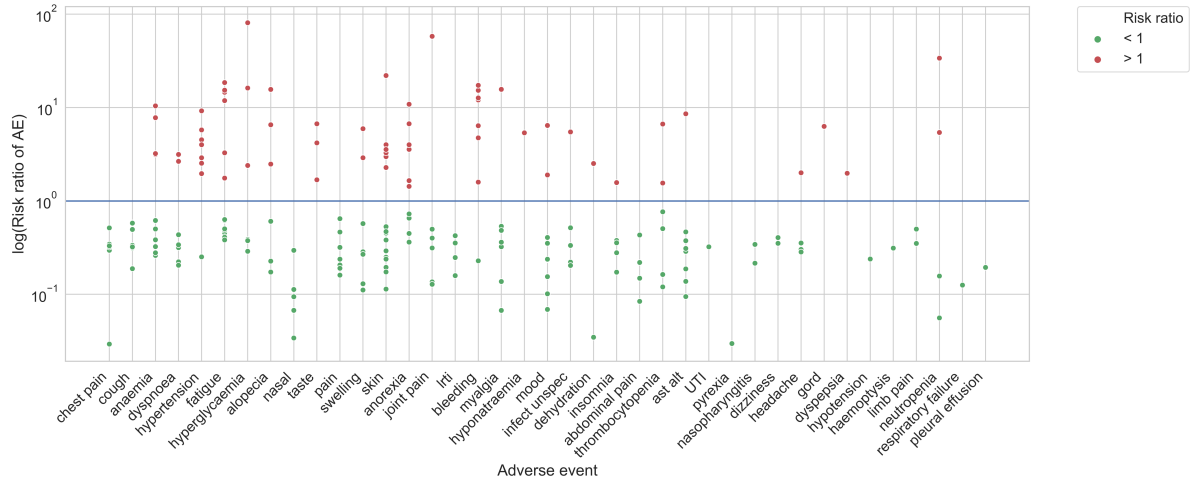


Figure 5. Risk ratios for 39 adverse events in the SAEgnal framework. Each data point represents a risk ratio for each combination treatment arms in studies where a given AE was reported. Each color is specific to an adverse event.

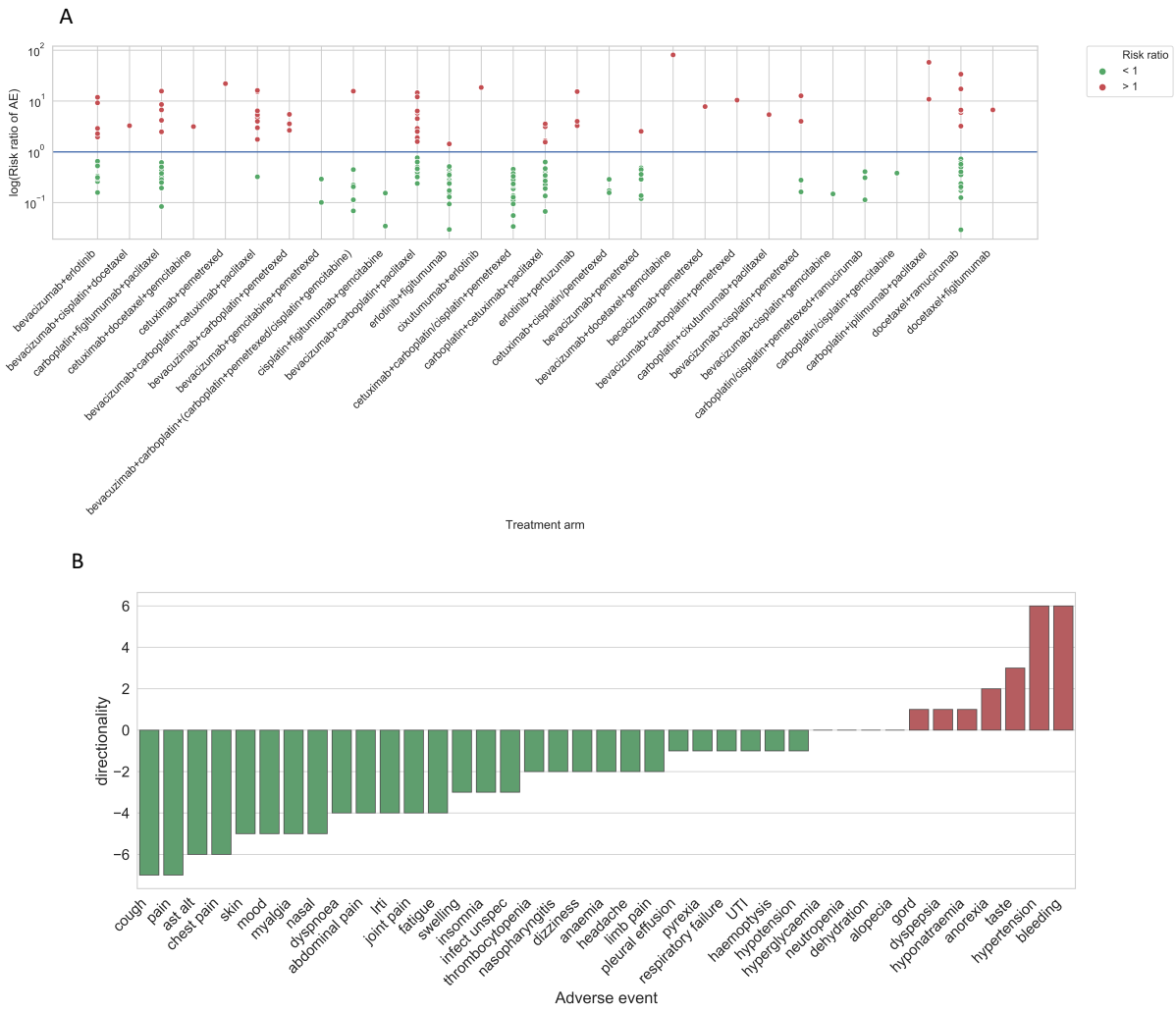


Figure 6. (A) Risk ratios for adverse events across treatment arm types. (B) Directionality of AEs in combinations. The y-axis contains a sum of combination treatment arms in each study where an observed risk of adverse event was lower than estimated from mono arms (green) and higher, than estimated (red).

Discussion

As combination therapeutics development increases, we think it will be essential to establish a data science pipeline to gather trial-level data, analyze and interpret, with interpretable results that eventually lead to the accelerating decision making to enable drug development. Using SAEgnal, we have proposed a framework to identify suitable studies, curate the data, transform and prepare in a format where advanced quantitative analyses such as machine learning methods can be applied. Besides, we have statistically analyzed the data to compare the risks of adverse events in combination with arms with an estimated risk of AEs in mono arms. Finally, we assessed the risks of AEs in combinations as lower or higher than the estimated values in mono arms. To obtain the risks of AEs in IO and chemo treatment arms, we have pooled all IO studies together, irrespective of a drug used. Chemotherapy arms were pooled as well, but it was performed for each kind of chemotherapeutic agent separately. This step may be controversial and questionable as it is based on a simplified assumption that the profiles of IO's adverse events are similar. IO drugs are of protein origin and are typically linked with immune-related AEs, irAEs (30% of patients treated with immunotherapies suffer from skin-related adverse events such as rash and mucositis[7]), since as proteins; they have different potential to activate immune system[8]. However, we are aware that some adverse

events are more specific to the type of IO drugs. For example, CTLA4 antibodies were reported to cause an increased onset of gastrointestinal disorders, such as colitis and diarrhea (30%-40% of dosed patients[9]). Our goal of developing SAEgnal as a data compendium and analytics framework was to answer the question: *is a risk of AE in combination different from an estimated risk in separate mono arms?* Many methods can answer this question, including logistic regression, bootstrapping, etc. We proposed a contingency 2x2 table as a valid method for this task. Contingency or cross-tabulation tables have been well documented as a method chosen for clinical trial data analysis[10,11,12]. We have taken the observed numbers of AEs and non-AEs from each combination study in the first step. We have then estimated pooled risks of AE across IO and chemo arms (all IO grouped, and chemotherapeutic agents grouped respectively to their types). Each combination study population was then multiplied by an estimated risk to obtain an estimated number of AE and non-AE events. We have then created a 2x2 table where we obtained risk ratios, p-values, upper and lower confidence intervals for observed AEs/non-AEs vs. estimated AEs/non-AEs in combination therapy. These results were saved in a separate data set and subjected to further analysis. Interestingly, we have found that the observed risk of the majority of AEs across different studies is lower than expected. It may be a result of a decreased dosing of individual components within combination therapies. Depending on a study type and drugs combined, the cumulative dosing of two drugs may be either reduced or not, and both safety and efficacy are considered. Such decisions are made individually for each type of combination treatment; for example, bevacizumab and telatinib or bevacizumab and vatalanib combination can be applied at 60% of the additive dose since both drugs overlap in their target angiogenesis[13,14,15]. Elevated ALT and AST indicating liver damage have also been decreased in the IO-chemo combinations compared to mono-arms. This observation could also be due to the result of reduced dosing of individual drugs in combinations. Published information about liver damage mechanisms indicates that chemotherapy may directly induce liver damage by putting stress on the filtering functions, causing a toxic effect[16]. Immunotherapies, however, may result in liver damage as a result of an immune system-mediated response. Such risk has been reported to increase when two IO drugs are used in the combination[17]. We cannot exclude a possibility that IO and chemo drugs in combination, especially when a possibility of lower dosing is considered, may not activate any of the liver-damage mechanisms as much as individual therapies. It is possible that in the case of studying IO-IO combination, the directionality of liver damage would be opposite to what we have found in our study. Interestingly, the risk of bleeding in combination treatment arms from our analysis seems to be elevated compared to individual mono arms. In mono-arms where chemotherapy is introduced, thrombocytopenia is typically connected with an increased risk of bleeding because a lower platelet count fosters the formation of clots that otherwise stop bleeding[18]. Immunotherapies were also linked to increased bleeding but through a different mechanism, also immune system-mediated[19]. While in our analysis, we did not find an increased risk for thrombocytopenia across multiple treatments (2 treatments had an increased than the estimated risk of thrombocytopenia and 4 studies, decreased) we cannot exclude a possibility of a synergistic effect, where bleeding occurs through different mechanisms of action of combination drugs. Although IO drugs are reported to cause skin problems, our IO-chemo combination analysis did not indicate an increased risk of such AEs. Risks of different types of pain were also decreased compared to mono arms in our data. The decrease in pain may both be related to an increased efficacy as pain is often associated with developing cancer disease and reduced toxicity of both drugs in combination.

Predictive Modeling

Predictive analytics of SAEs in combination therapies using signals from monotherapies would accelerate the development process of IO-combination therapies[20]. With the development of machine learning techniques, our data could be analyzed into a predictive framework. We have performed a quick test where we applied a random forest to predict the directionality of adverse events, using just treatment labels to train an algorithm. By splitting our data set into a training part and a validation part, we achieved ~89% of validation accuracy. The addition of inclusion criteria columns, together with population variables, makes this framework especially useful for the clinical teams working on the design of the new studies. New studies are typically designed based on historical data, subject matter expertise, and step-by-step approaches by further expanding current indications. We think that utilizing publicly available metadata will be a powerful tool to support the data-driven decision-making process to accelerate drug development. Predictive analytics of SAEs in combination therapies using signals from monotherapies would help in the data-driven development process of IO-combination therapies.

Limitations

The AACT data set that we have used in the SAEgnal framework was uploaded in 2016, and the combination therapies were relatively new in the industry. Hence, we face the challenge of a low volume of publicly available data. This severely limits the utility of the SAEgnal in its current iteration. Also, as a proof-of-concept study, we focused only on one tumor type. Furthermore, the mapping of SAEs was performed using a rational mapping strategy to simplify the terms' definition. We did not utilize any standard clinical ontologies other MedDRA in the current iteration of SAEgnal, but it could be explored in the next iteration. A regular update of SAEgnal that provides access to pan-cancer clinical trial results with ontology mapped datasets could become an indispensable tool for evaluating SAEs. With enough future data available, some IO drugs may result in different profiles of AEs, which might also depend on the treated population. However, the availability of IO data in mono arms is insufficient to allow a sensitivity analysis at the current state. Coverage of drugs and treatment arm for IO mono arms is so low that it does not justify utilizing them as control arms for respective IO drugs in combinations. Therefore, pooling is a reasonable option to overcome this challenge. A similar question arises with the pooling of chemo drugs across moon arms. However, in this case, drugs are pooled across studies depending on the chemotherapy regimen. Still, it is important to ask whether the pooling of different studies does not introduce a bias. One way to overcome it was to select studies from the same indication: NSCLC, with similar populations, a similar range of inclusion and exclusion criteria, and disease characteristics. Once the availability of data increases over time, pooling any data will be no more demanded.

Conclusions

To conclude, in this paper, we have proposed SAEgnal, a new immuno-oncology analytical framework to study SAEs in combination therapies by integrating a novel dataset from historical, publicly available clinical trial data on combination arms and mono arms. We have identified that although scarce, clinical data available so far can provide statistically significant information on the interaction between different treatments. We have presented several scenarios where results are in alignment with the published data on adverse events. The data set we have created can serve as an input for machine learning algorithms to predict the occurrence of adverse events and their directionality based on drugs planned in treatment arms and other information available at the study design and

strategic planning stage of clinical development. Further, by expanding the data backend and feature capabilities, the SAEgnal framework could be a useful tool for pre-emptive and predictive adverse event monitoring using clinical trial big data and machine intelligence.

Acknowledgements

The authors would like to acknowledge Drs. Hesham Abdullah, Dejan Pavlovic, Claire Morgan, and Bhaskar Dutta for discussions.

References

1. Yadav KK, Shameer K, Readhead B, Yadav SS, Li L, Kasarskis A, et al. Systems medicine approaches to improving understanding, treatment, and clinical management of neuroendocrine prostate cancer. Curr Pharm Des. 2016;22(34):5234-48.
2. Camidge DR, Doebele RC, Kerr KM. Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC. Nat Rev Clin Oncol. 2019;16(6):341-55.
3. Rotte A. Combination of CTLA-4 and PD-1 blockers for treatment of cancer. J Exp Clin Cancer Res. 2019;38(1):255.
4. Maitland ML, Hudoba C, Snider KL, Ratain MJ. Analysis of the yield of phase II combination therapy trials in medical oncology. Clin Cancer Res. 2010;16(21):5296-302.
5. Paller CJ, Huang EP, Luechtefeld T, Massett HA, Williams CC, Zhao J, et al. Factors affecting combination trial success (FACTS): investigator survey results on early-phase combination trials. Front Med (Lausanne). 2019;6:122.
6. Wu M, Sirota M, Butte AJ, Chen B. Characteristics of drug combination therapy in oncology by analyzing clinical trial data on ClinicalTrials.gov. Pac Symp Biocomput. 2015:68-79.
7. Kumar V, Chaudhary N, Garg M, Floudas CS, Soni P, Chandra AB. Current diagnosis and management of immune related adverse events (irAEs) induced by immune checkpoint inhibitor therapy. Front Pharmacol. 2017;8:49.
8. Liu YH, Zang XY, Wang JC, Huang SS, Xu J, Zhang P. Diagnosis and management of immune related adverse events (irAEs) in cancer immunotherapy. Biomed Pharmacother. 2019;120:109437.
9. Hodi FS, O'Day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, et al. Improved survival with ipilimumab in patients with metastatic melanoma. N Engl J Med. 2010;363(8):711-23.
10. Noguchi Y, Ueno A, Otsubo M, Katsuno H, Sugita I, Kanematsu Y, et al. A simple method for exploring adverse drug events in patients with different primary diseases using spontaneous reporting system. BMC Bioinformatics. 2018;19(1):124.
11. Hasselblad V LY. Tests for 2 x 2 tables in clinical trials. Journal of Modern Applied Statistical Methods. 2007;6(10).
12. Sauerbrei W, Blettner M. Interpreting results in 2 x 2 tables: part 9 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2009;106(48):795-800.
13. Langenberg MHG, Witteveen PO, Roodhart J, Lolkema MP, Verheul HMW, Mergui-Roelvink M, et al. Phase I evaluation of telatinib, a VEGF receptor tyrosine kinase inhibitor, in combination with bevacizumab in subjects with advanced solid tumors. Ann Oncol. 2011;22(11):2508-15.
14. Jones SF, Spigel DR, Yardley DA, Thompson DF, Burris HA, 3rd. A phase I trial of vatalanib (PTK/ZK) in combination with bevacizumab in patients with refractory and/or advanced malignancies. Clin Adv Hematol Oncol. 2011;9(11):845-52.
15. Liu S, Nikanjam M, Kurzrock R. Dosing de novo combinations of two targeted drugs: Towards a customized precision medicine approach to advanced cancers. Oncotarget. 2016;7(10):11310-20.
16. Sharma A, Houshyar R, Bhosale P, Choi JI, Gulati R, Lall C. Chemotherapy induced liver abnormalities: an imaging perspective. Clin Mol Hepatol. 2014;20(3):317-26.
17. Bose V SD, Penn J, Rustgi V, John T, Mishra A. Immunotherapy induced liver injury: a retrospective analysis. American Journal of Gastroenterology. 2018;113(p s45).
18. Goldberg GL, Gibbon DG, Smith HO, DeVictoria C, Runowicz CD, Burns ER. Clinical impact of chemotherapy-induced thrombocytopenia in patients with gynecologic cancer. J Clin Oncol. 1994;12(11):2317-20.
19. Eberst G, Lakhzoum W, Tomasini P, Andreotti N, Abcaya J, Mascaux C, et al. Autoimmune-related bleeding occurring during combined immunotherapy for lung cancer - Case report. Rev Mal Respir. 2018;35(9):974-7.
20. Maymani H, Hess K, Groisberg R, Hong DS, Naing A, Piha-Paul S, et al. Predicting outcomes in patients with advanced non-small cell lung cancer enrolled in early phase immunotherapy trials. Lung Cancer. 2018;120:137-41.

Association of Neighborhood-Level Factors and COVID-19 Infection Patterns in Philadelphia Using Spatial Regression

Mary Regina Boland, MA, MPhil, PhD, FAMIA[1], Jessica Liu[1], Cecilia Balocchi, PhD, Jessica Meeker, MPH[1], Ray Bai, PhD[2], Ian Mellis, PhD[3], Danielle L. Mowery, PhD, MS, MS[1], Daniel Herman, MD, PhD[3]

[1]Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [2]Department of Statistics, University of South Carolina, SC, USA; [3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Abstract

As of August 2020, there were ~6 million COVID-19 cases in the United States of America, resulting in ~200,000 deaths. Informatics approaches are needed to better understand the role of individual and community risk factors for COVID-19. We developed an informatics method to integrate SARS-CoV-2 data with multiple neighborhood-level factors from the American Community Survey and opendataphilly.org. We assessed the spatial association between neighborhood-level factors and the frequency of SARS-CoV-2 positivity, separately across all patients and across asymptomatic patients. We found that neighborhoods with higher proportions of individuals with a high-school degree and/or who were identified as Hispanic/Latinx were more likely to have higher SARS-CoV-2 positivity rates, after adjusting for other neighborhood covariates. Patients from neighborhoods with higher proportions of individuals receiving public assistance and/or identified as White were less likely to test positive for SARS-CoV-2. Our approach and its findings could inform future public health efforts.

1. Introduction

1.1 Background on COVID-19

A novel strain of coronavirus (COVID-19) was discovered in Wuhan, China in December 2019. Initially, COVID-19 began as an epidemic in that local region, resulting in strict lock-downs of approximately 500 million people in China. Since then COVID-19 has since spread across the globe, impacting every continent including Antarctica (as of December 2020). As of August 25, 2020, there had been over 23 million confirmed cases and 800,000 deaths worldwide[1].

1.2 Importance of Identifying Neighborhood-Level Factors in COVID-19

Since March 2020, a plethora of COVID-19 related research papers have correlated social determinants of health with COVID-19 spread. Factors such as poverty and income can affect an individual's ability to effectively distance physically from others (colloquially as 'social distancing'). In addition, socioeconomic status is closely associated with job type [2]. Occupations, such as grocery store workers and pharmacy store workers are at increased risk of COVID-19 [3], and a large proportion of these jobs are held by those of lower socioeconomic status and/or immigrant populations [4]. It follows that neighborhood-level characteristics (e.g., income, poverty) of certain census tracts (i.e., neighborhoods) within a large diverse city such as Philadelphia, could be associated with different SARS-CoV-2 positivity rates. However, it is unclear at this present time what neighborhood-level factors are the most closely linked with COVID-19 spread and their relative importance to each other (e.g., is education or poverty or income more informative?). What is lacking is sufficient understanding of neighborhood-level characteristics to tailor public health messaging and other interventions to reduce the SARS-CoV-2 transmission across communities that may be otherwise high-risk for spread (e.g., certain low-income communities).

1.3 Significant Variability Exists in Neighborhood Characteristics in Philadelphia

Neighborhoods are defined in this paper as census tracts. Great variability exists across Philadelphia neighborhoods and it is truly a 'tale of two cities'. One is rich and the other poor in material resources. These neighborhood (i.e., census-tract) level associations within Philadelphia are directly applicable to healthcare. One example of neighborhood-level or census tract-level differences within Philadelphia is

income with low-income neighborhoods having been linked with increases in fire-related injuries [5]. There also exists considerable disparity in terms of the number of primary care providers. The average ratio of adults per primary care provider was 1,073 across all census tracts within the city of Philadelphia [6]. However, some census tracts had only 105 while others had 10,321 [6]. This disparity affects physical availability access to care. In addition, we have shown that neighborhood-level factors in Philadelphia affect C-section rates among pregnant women [7]. Overall, significant variability exists within the city of Philadelphia in terms of the neighborhood-level characteristics (i.e., census tract-level characteristics). This is also true of many other urban centers (e.g., New York City, San Francisco) and therefore our approach should generalize to other urban centers.

1.4 Need for Informatics Approaches to Address these Gaps

Biomedical informatics methods can be developed to identify neighborhood-level factors that are associated with COVID-19 infection. If these factors were known, it could assist in effective containment of COVID-19 or other pandemics in future [8]. SARS-CoV-2 testing is being offered in a variety of settings. Some patients are received through an outpatient office or via drive through testing sites. Others are received via the in-patient care system (e.g., Emergency Departments). The testing orders, results, and associated clinical information such as clinical symptomology are stored and distributed in a complex web. For this study, we focused on testing being performed in one academic medical center on specimens collected from a variety of sites and patients with a variety of indications, all aggregated in one integrated electronic health record (EHR) system. Effective informatics solutions are needed to identify relevant patient cohorts (e.g., asymptomatic vs. symptomatic) and to understand associated neighborhood-level factors for these different cohorts.

1.5 Purpose of Study

The goals of this study are two-fold: 1.) develop an informatics approach to integrate multiple, potentially relevant neighborhood-level factors (e.g., income, poverty, education) from American Community Survey (ACS) and other Philadelphia-specific open datasets (e.g., violence, housing quality) from OpenDataPhilly.org with SARS-CoV-2 testing data; and 2.) determine what neighborhood-level factors are associated spatially with COVID-19 spread within the city of Philadelphia. These neighborhood-level factors are available at the census-tract level. We envision that this information could be used for future city and institutional planning to more readily identify COVID-19 'hotspots' within the city and the risk factors associated with these hotspots (e.g., poverty). Moreover, for healthcare institutions, these information should be useful for institutional planning purposes to understand what types of communities are most at risk to enable adaptation of testing and contact tracing strategies for those communities.

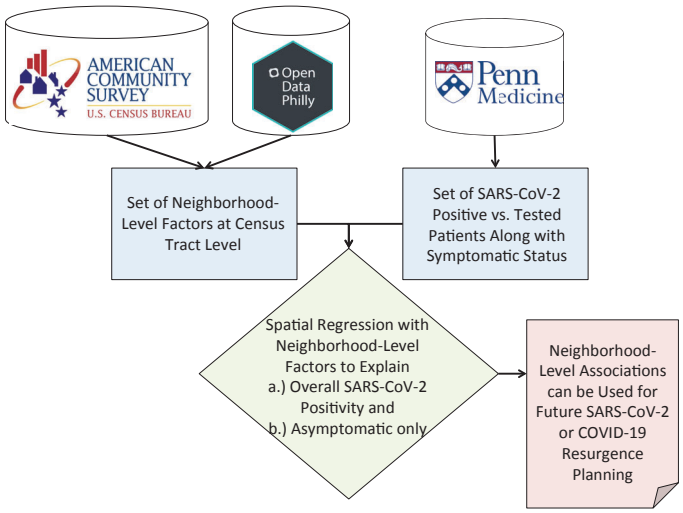


Figure 1. Overview of Informatics Approach to Aggregate a Set of Neighborhood-Level Covariates from Public Sources and Link Using Spatial Regression with SARS-CoV-2 Positivity Status for All Patients Tested and All Asymptomatic Patients Tested.

2. Materials and Methods

Our informatics approach involves assembling and integrating data SARS-CoV-2 testing data with various sources on neighborhood-level factors. We obtained clinical data of SARS-CoV-2 positivity rates broken down by symptomatic status (i.e., overall vs. asymptomatic patients only). We then built a spatial regression model to assess the role of each neighborhood-level factor on SARS-CoV-2 positivity rates within the city of Philadelphia. Our overall strategy is visualized in **Figure 1.** This study was performed as a quality

assurance and quality improvement project within the Department of Pathology and Laboratory Medicine in the University of Pennsylvania.

2.1 Obtaining Neighborhood-Level Descriptors

Table 1 includes the neighborhood-level factors included in our model along with each source. All of our 'neighborhoods' consist of census tracts. Therefore, for the purpose of this study, a neighborhood is a census tract. We chose census-tract characteristics to inform our models regarding the neighborhood and communities living situations within that census tract.

2.1.1 American Community Survey

We queried the United States Census Bureau website for data on neighborhood-level factors (at the census-tract level) for inclusion in our models. We then queried the US Census query interface called CEDSCI (Center for Enterprise Dissemination Services and Consumer Innovation) accessible at: https://data.census.gov/cedsci/. We downloaded all data for 2010-2017 and only included 2017 data in our modeling of the neighborhood-level associations with COVID-19, to focus on the most recent survey data. Specific data file names are given in **Table 1**. Although, these data are available for all US census tracts, we restricted our datasets to Philadelphia-only census tracts to study the relationship between neighborhood-level factors and SARS-CoV-2 positivity rates within this most immediate urban catchment area for Penn Medicine.

2.1.2 OpenDataPhilly

We incorporated OpenDataPhilly data accessible at https://www.opendataphilly.org/ for information on housing quality. This housing dataset contains information on buildings and units that have been cited by law enforcement and/or city regulations and inspections officials for violations to housing quality laws. We integrated general violations data from the Licenses and Inspections violations, obtainable at: https://www.opendataphilly.org/dataset/licenses-and-inspections-violations. We also only used the general violations category to ensure that we had data across as many Philadelphia census tracts as possible for the year 2017. For information on the violent and non-violent crime rates, we included data previously obtained by Dr. Balocchi [9] that was obtained from the Philadelphia Police Department, available on https://www.opendataphilly.org/dataset/crime-incidents. To be consistent with the ACS data, we only included data from 2017.

Table 1. Sources and Data Files for Neighborhood-Level Factors Included in Our Spatial Regression Model

Neighborhood-Level Factors	Source	Data File
Prop. of women aged 15-50 years in each census tract below 100 percent poverty level	ACS	S1301
Prop. of women aged 15-50 years in each census tract that graduated high school (including equivalency)	ACS	S1301
Prop. of women aged 16-50 years in each census tract that are in the labor force	ACS	S1301
Prop. of women aged 15-50 years in each census tract that received public assistance income in the past 12 months	ACS	S1301
Prop. of occupied housing units in each census tract that are owner-occupied	ACS	S2502
Prop. of occupied housing units in each census tract that are renter-occupied	ACS	S2502
Median family income (dollars)	ACS	S1903
Prop. of each census tract that identifies as Asian Alone	ACS	B01001D
Prop. of each census tract that identifies as Black or African-American	ACS	B01001B
Prop. of each census tract that identifies as Hispanic or Latinx	ACS	B01001I
Prop. of each census tract that identifies as White Alone	ACS	B01001A
Housing Violations	OpenDataPhilly	
Violent Crime Rate	OpenDataPhilly	
Non-Violent Crime Rate	OpenDataPhilly	

2.2 SARS-CoV-2 Positive Patients Broken Down by Symptomatic Status

We identified whether patients being tested for SARS-CoV-2 were symptomatic based on documentation in the test's ordering questions, as stored in EHR. This allowed us to investigate all patients grouped together regardless of symptomatic status (comparison 1) and also only asymptomatic patients (comparison 2). Asymptomatic patients receiving a SARS-CoV-2 test did not report any specific signs or symptoms of COVID-19 (e.g., fever, cough, loss of taste or smell) and were either tested due to presumed or confirmed

exposure (e.g., living with a COVID-19 positive individual), screening prior to clinical care (e.g. surgical procedure, hospital admission, pregnancy labor & delivery), the result of contact tracing efforts. We were able to perform this stratification because we had access to documentation of symptomatic status.

2.3 Statistical Analysis: Spatial Regression Model of Effect of Neighborhood-Level Factors on SARS-CoV-2 Positivity Status

First, we performed a log-transformation of housing violation data, income, and both violent and non-violent crime to normalize them. Next, we evaluated two models: one to predict SARS-CoV-2 positivity rate at the neighborhood-level for all patients and the second model to predict SARS-CoV-2 positivity rate among asymptomatic patients only. All analyses were performed using R statistical software (version 3.6.1). We used the R spautolm function from spdep package to perform spatial autoregression, fit using Maximum Likelihood estimation. This method is optimized for sparse matrices, as observed in our dataset. Our outcome was the % SARS-CoV-2 positive out of total conclusive test results (i.e., Number of Positive / (Number of Positive + Number of Negative)) for each census tract. After univariable results were obtained, we then explored multi-predictor models and included all nominally significant ($p <= 0.05$) results into a single spatial regression model. This allowed us to assess the importance of each nominally significant neighborhood-level predictor while accounting for the other factors in an adjusted model.

3. Results

3.1 Penn Medicine Cohort of Individuals Tested for SARS-CoV-2

Our population consists of 46,001 unique individuals tested for SARS-CoV-2 between March 3, 2020 and June 9, 2020. We geocoded patient addresses using ArcGIS and mapped the corresponding latitude and longitude coordinates to census tracts within the city of Philadelphia. We observed 19,281 distinct patients lived within the city of Philadelphia and 26,824 distinct test results that could be linked to census tract information for analysis. Several patients were tested multiple times.

Most of our patients had a documented symptomatic status (i.e., asymptomatic or symptomatic) allowing us to stratify our population and perform subanalyses among asymptomatic individuals only. **Table 2** shows the positivity rates by symptomatic status, revealing a rate of 10.1% SARS-CoV-2 positivity among asymptomatic individuals as compared to 20.7% positivity rate for all patients during this time frame. For simplicity, we excluded rare inconclusive and non-discrete results from our statistical analysis.

Table 2. SARS-CoV-2 Positivity Rates by Symptomatic Status, Penn Medicine March-June 2020

SARS-CoV-2 Test Result	All Patients	Known Asymptomatic
Positive	4781 (20.73%)	680 (10.14%)
Negative	18278 (79.27%)	6023 (89.86%)
Total Positive or Negative Result	23059	6703
Total Completed Results	26824	7395

3.2 Neighborhood-Level Associations with SARS-CoV-2 Positivity Status: Overall

3.2.1 Single Predictor Spatial Regression Model

We performed single predictor or univariable analysis to obtain results for each variable and its relationship with SARS-CoV-2 positivity (**Table 3).** All neighborhood-level predictors were associated ($p < 0.001$) with the exception of four, the proportion of those identifying as Asian living in a census-tract, the non-violent crime rate, and the proportion of occupied housing units that was either renter occupied or owner occupied. Interestingly, three neighborhood-level factors were associated with *lower* than expected SARS-CoV-2 positivity: median family income, proportion of those identifying as White living in a census-tract, proportion of women aged 16-50 that participated in the labor force (i.e., women who were working). In other words as these three factors (i.e., income, White, labor force participation) increased in a given neighborhood the rate of SARS-CoV-2 positivity went down.

3.2.2 Multi-Predictor Spatial Regression Model

We next constructed a multi-predictor spatial regression model using all 10 statistically significant neighborhood-level factors from the single predictor models (**Table 4).** We report the goodness-of-fit statistics for our multi-predictor model, including lambda=-0.0707, Hessian Standard Error of lambda=0.0842, log-likelihood=372.12 and log likelihood ratio of 0.6665 with a corresponding p-value=0.4143. Because the p-value of the log-likelihood ratio >=0.05, we can conclude that spatial

clustering of SARS-CoV-2 positivity rates that may appear in our data has been accounted for by the factors included in this multi-predictor model with results shown in **Table 4**[10]. We visualized the SARS-CoV-2 positivity rates in Philadelphia by census tract (**Figure 2)** and compared this with one of the nominally significant variables that is potentially actionable, the proportion of women 15-50 receiving public assistance (**Figure 3).**

Table 3. Results of Single Predictor Spatial Regression Analyses of Neighborhood-Level Factors on Neighborhood-Level SARS-CoV-2 Positivity Rate for All Patients

Short Name	Neighborhood-Level Factors	Odds Ratio	P-value
Income	Median family **income** (dollars) (log-transformed variable)	0.9144	<0.001
Education	Prop. of women aged 15-50 years in each census tract that **graduated high school** (including equivalency)	1.4471	<0.001
White	Prop. of each census tract that identifies as **White** Alone	0.8601	<0.001
Violent Crime	**Violent Crime Rate** (log-transformed variable)	1.0463	<0.001
Black	Prop. of each census tract that identifies as **Black or African-American** Alone	1.1117	<0.001
Poverty	Prop. of women aged 15-50 years in each census tract below 100 percent **poverty** level	1.2243	<0.001
Public Assistance	Prop. of women aged 15-50 years in each census tract that **received public assistance** income in the past 12 months	1.6357	<0.001
Housing Violations	**Housing Violations** (log-transformed variable)	1.0302	<0.001
Hispanic	Prop. of each census tract that identifies as **Hispanic or Latinx**	1.1853	<0.001
Labor Force	Prop. of women aged 16-50 years in each census tract that are **in the labor force**	0.8162	<0.001
Asian	Prop. of each census tract that identifies as **Asian** Alone	0.96	0.5582
Renter	Prop. of occupied housing units in each census tract that are **renter**-occupied	0.9841	0.5879
Owner	Prop. of occupied housing units in each census tract that are **owner**-occupied	1.0161	0.5879
Non-Violent Crime	**Non-Violent Crime** Rate (log-transformed variable)	1.0017	0.8439

Table 4. Results of Multi-Predictor Spatial Regression Model of Neighborhood-Level Factors on Neighborhood-Level SARS-CoV-2 Positivity Rate for All Patients

Short Name	Neighborhood-Level Factors	Adj. Odds Ratio	P-value
Education	Prop. of women aged 15-50 years in each census tract that **graduated high school** (including equivalency)	1.2114	<0.001
White	Prop. of each census tract that identifies as **White** Alone	0.8283	0.0011
Hispanic	Prop. of each census tract that identifies as **Hispanic or Latinx**	1.1302	0.0027
Public Assistance	Prop. of women aged 15-50 years in each census tract that **received public assistance** income in the past 12 months	0.8136	0.0406
Labor Force	Prop. of women aged 16-50 years in each census tract that are **in the labor force**	1.0942	0.1045
Black	Prop. of each census tract that identifies as **Black or African-American** Alone	0.9421	0.2708
Income	Median family **income** (dollars) (log-transformed variable)	0.9891	0.5592
Housing Violations	**Housing Violations** (log-transformed variable)	1.0027	0.7263
Violent Crime	**Violent Crime Rate** (log-transformed variable)	0.9989	0.9116
Poverty	Prop. of women aged 15-50 years in each census tract below 100 percent **poverty** level	1.0012	0.9816

Four neighborhood-level factors appeared nominally associated ($p < 0.05$) in our multi-predictor model. Two variables were associated with *increased* SARS-CoV-2 positivity rates at the neighborhood-level: the proportion of women 15-50 having at least a high school education and the proportion of individuals identifying as Hispanic/Latinx in that neighborhood. Two neighborhood-level factors were associated with *decreased* SARS-CoV-2 positivity rates: the proportion of individuals identifying as White alone in a given census-tract and the proportion of women aged 15-50 years old receiving public assistance. Of note,

according to the Census Bureau public assistance includes both cash and non-cash benefits (e.g., Temporary Assistance for Needy Families or TANF, Supplemental Nutrition Assistance Program or SNAP) to low-income families or individuals [11]. Importantly, the proportion of a census tract that identifies, as Black or African American was no longer associated with SARS-CoV-2 positivity in the multi-predictor model, suggesting other correlated neighborhood-level factors were more informative in SARS-CoV-2 positivity rates.

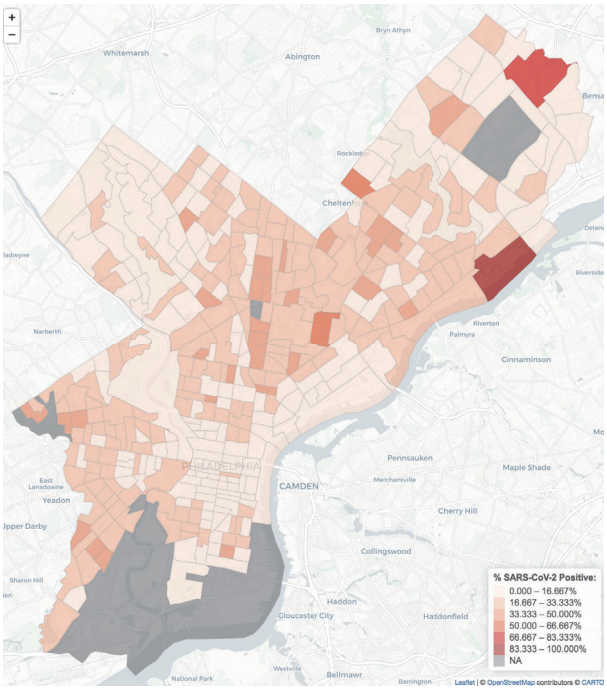


Figure 2. SARS-CoV-2 Positivity Rates (March-June 2020)

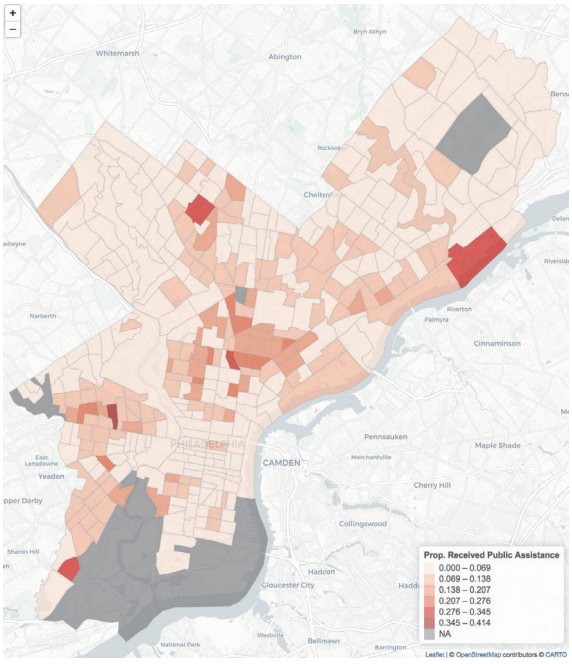


Figure 3. Proportion of Women 15-50 Receiving Public Assistance in 2017 in Philadelphia

3.3 Neighborhood-Level Associations with SARS-CoV-2 Positivity Status: Asymptomatic Only

We conducted a sub-analysis among only patients that were identified as being asymptomatic at the time of testing. We had 7,395 COVID-19 test results that were from asymptomatic patients who live within the city of Philadelphia (**Table 2**) and we focused the patients with confirmed positive or negative COVID-19 test results. The asymptomatic population has a lower overall positivity rate, as expected.

3.3.1 Single Predictor Spatial Regression Model

We performed single predictor (univariable) regression to characterize each variable's relationship with SARS-CoV-2 positivity among asymptomatic patients. We found that eight neighborhood-level factors were associated ($p < 0.05$) with positivity (**Table 5**). Two of these neighborhood-level factors *lower* SARS-CoV-2 positivity rates, including the proportion of the census tract that identified as being of White race and the median family income in a census tract. The remaining six associations were positively correlated, indicating that they were associated with *higher* SARS-CoV-2 positivity rates. Those neighborhood characteristics included, the proportion of individuals identifying as being of Black or African American race, the violent crime rate, the number of housing violations and the proportion of women 15-50 that graduated high school and the proportion of women 15-50 that received public assistance. As in the overall model (which included those that were symptomatic or asymptomatic), we observed that the proportion of housing units that were either renter occupied or owner occupied was not significantly correlated with SARS-CoV-2 positivity rates.

Table 5. Results of Single Predictor Spatial Regression Analysis of Neighborhood-Level Factors on Neighborhood-Level SARS-CoV-2 Positivity Rate: Asymptomatic Only

Short Name	Neighborhood-Level Factors	Odds Ratio	P-value
Black	Prop. of each census tract that identifies as **Black or African-American** Alone	1.0831	<0.001
White	Prop. of each census tract that identifies as **White** Alone	0.9177	<0.001
Income	Median family **income** (dollars) (log-transformed variable)	0.9574	<0.001
Violent Crime	**Violent Crime Rate** (log-transformed variable)	1.0261	<0.001
Housing Violations	**Housing Violations** (log-transformed variable)	1.0216	<0.001
Education	Prop. of women aged 15-50 years in each census tract that **graduated high school** (including equivalency)	1.1716	<0.001
Poverty	Prop. of women aged 15-50 years in each census tract below 100 percent **poverty** level	1.1000	0.0033
Public Assistance	Prop. of women aged 15-50 years in each census tract that **received public assistance** income in the past 12 months	1.2549	0.0108
Non-Violent Crime	**Non-Violent Crime** Rate (log-transformed variable)	1.0138	0.1176
Asian	Prop. of each census tract that identifies as **Asian** Alone	0.9007	0.1274
Labor Force	Prop. of women aged 16-50 years in each census tract that are **in the labor force**	0.9370	0.1371
Owner	Prop. of occupied housing units in each census tract that are **owner**-occupied	0.9554	0.1425
Renter	Prop. of occupied housing units in each census tract that are **renter**-occupied	1.0467	0.1425
Hispanic	Prop. of each census tract that identifies as **Hispanic or Latinx**	0.9928	0.8288

3.3.2 Multi-Predictor Spatial Regression Model

We included all 8 nominally statistically significant neighborhood-level factors from our asymptomatic only univariable models into a multi-predictor model (**Table 6).** Metrics denoting the goodness-of-fit for the model, include lambda=-0.0493, Hessian Standard Error of lambda=0.0860, log-likelihood=291.73 and log likelihood ratio of 0.3234 with a corresponding p-value=0.5695. Because the p-value of the log-likelihood ratio >=0.05, we can conclude that any spatial clustering of SARS-CoV-2 positivity rates that may appear in our data has been accounted for by the factors included in this multi-predictor model [10]. In this model, we observed no neighborhood-level factors that were significantly associated with SARS-CoV-2 positivity. The factor with the lowest p-value (p = 0.05) was the proportion of those identifying as being Black or African American in a census tract. Importantly, because the log likelihood ratio was 0.3234 with a corresponding p-value=0.5695, our model does remove any spatial variability in the SARS-CoV-2 positivity rates (if present) among asymptomatic individuals.

Table 6. Results of Multi-Predictor Spatial Regression Model of Neighborhood-Level Factors on Neighborhood-Level SARS-CoV-2 Positivity Rate: Asymptomatic only

Short Name	Neighborhood-Level Factors	Adj. Odds Ratio	P-value
Black	Prop. of each census tract that identifies as **Black or African-American** Alone	1.1032	0.0505
Poverty	Prop. of women aged 15-50 years in each census tract below 100 percent **poverty** level	1.0653	0.2209
Education	Prop. of women aged 15-50 years in each census tract that **graduated high school** (including equivalency)	1.0586	0.3684
White	Prop. of each census tract that identifies as **White** Alone	1.0499	0.4336
Public Assistance	Prop. of women aged 15-50 years in each census tract that **received public assistance** income in the past 12 months	0.9338	0.5776
Housing Violations	**Housing Violations** (log-transformed variable)	1.0052	0.5809
Violent Crime	**Violent Crime Rate** (log-transformed variable)	1.0040	0.7369
Income	Median family **income** (dollars) (log-transformed variable)	0.9993	0.9769

4. Discussion

4.1 Method Enables Rapid Identification of Neighborhood-Level Factors Linked to SARS-CoV-2 Positivity

Our informatics approach enabled us to identify four neighborhood-level factors associated with the frequency of SARS-CoV-2 test positivity, after adjusting for other correlated neighborhood-level covariates. Two neighborhood-level factors were positively correlated with increased SARS-CoV-2 positivity rates in our adjusted model; conversely, two other factors were negatively correlated with SARS-CoV-2 positivity rates. We discuss the negative or protective associations in section 4.2 and focus here on the neighborhood-level variables that are *positively* associated with SARS-CoV-2 test positivity. The two variables that were associated with higher SARS-CoV-2 positivity rates at the neighborhood-level were: the proportion of women 15-50 having at least a high school education, which is a proxy for education status, and the proportion of individuals identifying as Hispanic/Latinx in that neighborhood (**Table 4**). Patients from neighborhoods with increased proportions of individuals with at least a high school degree appeared more likely to be SARS-CoV-2 positive (adjusted odds ratio [aOR] = 1.2). Neighborhoods with increased proportions of individuals identifying as Hispanic or Latinx showed slightly higher SARS-CoV-2 positivity rates in the adjusted model (aOR=1.130).

It is extremely challenging to translate these apparent associations into causal relationships, because of the complex ascertainment schemas in these data, including the associations between neighborhood factors and symptom status. That said, a study conducted in the Washington DC area also found higher rates of SARS-CoV-2 positivity among Hispanic and Latinx populations [12]. Although generalizations tend to be overly simplistic [13], potential explanations include reduced rates of healthcare insurance and utilization among Hispanic populations [12]. Another factor worth considering is reduced ability to physically distance due to living in tightly crowded communities or inability to stop working due to economic constraints [14]. Finally, there is the added complexity of undocumented immigrants, who within the past year were disenrolled from public assistance programs (i.e., SNAP) [14].

We found that the proportion identifying as Black or African American in a given census tract did not appear independently associated with SARS-CoV-2 positivity in the multivariable model (**Table 4**). However, it was associated in the univariable model (**Table 3**), suggesting that some other neighborhood-level factors were correlated and potentially more informative for SARS-CoV-2 positivity rates. Of note, many other studies that observed associations between Black/African American communities and SARS-CoV-2 positivity [15] did not adjust for other correlated factors that may either increase risk (e.g., poverty) or decrease risk (e.g., public assistance). In addition, our adjusted model including other highly correlated factors, including the proportion of individuals identifying as White alone, which was observed to be associated with *lower* rates of SARS-CoV-2 positivity among neighborhoods (aOR = 0.83).

4.2 Our Method Can Identify Neighborhood-Level Factors Associated with Lower SARS-CoV-2 Positivity Rates in Large Metropolitan Areas

Our method assesses the role of a number of neighborhood-level factors that could impact SARS-CoV-2 positivity rates. We first examined each of these factors at the single variable level to identify those factors that are associated in some way with SARS-CoV-2 positivity. We then included all of the significant variables in a multivariable model to identify a set of variables that appear independently associated with test positivity. We found that two neighborhood-level factors were associated with *lower* SARS-CoV-2 positivity rates. One of these associations has been well established, namely that neighborhoods with a higher proportion of individuals identifying as White is associated with a decreased SARS-CoV-2 positivity rate. Cordes et Castro's also observed an association with reduced SARS-CoV-2 positivity rates among neighborhoods with higher proportions of White individuals analysis in their New York City population [15]. However, their analysis was only conducted at the univariable level and they did not conduct a multivariable model to adjust for other factors. Our multivariable model suggests that race may be a stronger predictor than other associated factors, including median family income and poverty status. Further research is warranted on this topic.

Importantly, our multivariable model found *lower* SARS-CoV-2 positivity rates at the neighborhood-level were associated with higher proportions of women aged 15-50 years old receiving public assistance (aOR=0.8). This is a particularly intriguing finding, because it hints at a potential public health

intervention. Individuals receiving these public assistance programs are impoverished, but receipt of government assistantship could enable these individuals to avoid environments that put them at higher risk of infection. This association held even when adjusting for other correlated variables (e.g., poverty, income, education, and various race/ethnicity factors). This finding supports the importance and impact of government and institutional planning initiatives to increase public assistance to at-risk communities.

4.3 Our Approach Can Be Useful for Public Health Intervention Design and Implementation

As we progress through the fall and winter, with further spikes in SARS-CoV-2 positivity rates, public health interventions to mitigate the spread of COVID-19 are crucial. The ability to identify COVID-19 hotspots, both in Philadelphia and across the USA, is critical for the design, development, implementation, and evaluation of these interventions. While contact tracing is an important tool in containing the spread of COVID-19 [16-18], some people remain unable to effectively quarantine or isolate. Therefore, programs to protect people and support contact tracing must be considered. For these programs to be effective, it will be invaluable to understand which areas in a city such as Philadelphia should be targeted. Furthermore, these data can inform where additional testing sites are needed to improve access for those communities that are most at risk.

4.4 Restricting Analysis to Asymptomatic Patients Helps Tease Out Causation from Amongst Associations

Our institution collects information on symptomatology at the time of testing, so we were able to subset our tested population into those with and without reported symptoms. This allowed us to perform two overall sets of analyses, the first involved all tested patients and the second was among only those who were asymptomatic state. Importantly, there are differences in the demographics factors between patients being tested while symptomatic or asymptomatic. Symptomatic patients are coming for care from the communities in the catchment area. On the other hand, asymptomatic patients are enriched for patients receiving specialty care or elective procedures. The asymptomatic population tends to be wealthier and better engaged in the healthcare system. We found 8 neighborhood-level factors appeared associated with SARS-CoV-2 positivity rates among asymptomatic patients. However, when we included all of these factors in a multivariable model there was insufficient evidence of association of any single factor with positivity (**Table 6**). This lack of associations was likely in part due to insufficient power from the lower sample size of asymptomatic subjects and the lower diversity in this population.

4.5 Limitations and Future Work

The limitations of this study include that the data is from March to June of 2020. Inclusion of more recent data would likely provide a richer study of the association of neighborhood-level factors and SARS-CoV-2 infection rates. We used neighborhood-level factors as surrogates for underlying patient factors and thus they cannot fully capture all of the relevant contextual factors for any single individual. This study is also constrained to the neighborhood-level factors collected by the ACS, OpenDataPhilly and other governmental agencies. Other important neighborhood-level factors may have been overlooked; thereby, limiting our results. In addition, our analysis is static and therefore does not account for changes in these factors and their associations over time. Neighborhood-level factors are available at the year-level and do not account for month over month changes in neighborhoods that could be occurring (therefore there is a temporal lag between neighborhood-level factors as reported by the Census Bureau and the true neighborhood-level characteristics). However, this does not limit our approach and those factors, once available, could be easily plugged into our models. Future work will explore the role of the joint relationship between individual-level factors that may contribute to SARS-CoV-2 positivity along with neighborhood-level factors. Finally, as this study was observational, the causal relationships underlying the observed associations could not be directly identified. We performed some multi-variable analyses and investigated the relationship between neighborhood level factors while adjusting for other important neighborhood level factors. This allowed us to identify neighborhood level factors that were strongly associated with SARS-CoV-2 positivity rates in our population while adjusting for other important neighborhood level factors. However, we did not probe any individual factor sufficiently to definitively establish or refute its association with SARS-CoV-2 positivity, especially among other populations where neighborhood-level characteristics may differ.

5. Conclusion

We have developed an informatics approach for integrating multiple neighborhood-level factors from the American Community Survey and opendataphilly.org with SARS-CoV-2 test results. We used these combined set of neighborhood-level factors to assess the spatial association between neighborhood-level factors and SARS-CoV-2 positivity rates amongst all tested patients and asymptomatic tested patients in one large, urban academic healthcare system. We observed that neighborhoods where either the fraction of the population that had graduated highschool education status or the proportion identified as Hispanic/Latinx was higher were associated with higher SARS-CoV-2 positivity rates; conversely, neighborhoods with higher proportions of those identifying as White or receiving public assistance were associated with lower SARS-CoV-2 positivity rates. Overall, we envision that our approach and its results could be used to inform future government and institutional programs by helping to identify causal risk factors (e.g., poverty) underlying COVID-19 'hotspots' and potential interventions (e.g., public assistance programs such as food stamps) that appear to mitigate SARS-CoV-2 spread.

Acknowledgments: We thank the University of Pennsylvania Perelman School of Medicine for generous startup funds and Department of Pathology and Laboratory Medicine for supporting this quality improvement project.

References

1. WHO. WHO Coronavirus Dashboard. <https://covid19whoint/>. 2020;Accessed on August 25, 2020.
2. Wright AL, Sonin K, Driscoll J, Wilson J. Poverty and economic dislocation reduce compliance with covid-19 shelter-in-place protocols. University of Chicago, Becker Friedman Institute for Economics Working Paper. 2020(2020-40).
3. Koh D. Occupational risks for COVID-19 infection. Occupational medicine (Oxford, England). 2020;70(1):3.
4. Gelatt J. Immigrant Workers: Vital to the US COVID-19 Response, Disproportionately Vulnerable. Washington, DC: Migration Policy Institute. 2020.
5. Shai D. Income, housing, and fire injuries: a census tract analysis. Public health reports. 2006;121(2):149-54.
6. Brown EJ, Polsky D, Barbu CM, Seymour JW, Grande D. Racial disparities in geographic access to primary care in Philadelphia. Health Affairs. 2016;35(8):1374-81.
7. Meeker J, Boland MR, editors. The association between neighborhood level exposures and progression to labor. APHA's 2020 VIRTUAL Annual Meeting and Expo (Oct 24-28); 2020: American Public Health Association.
8. Moore JH, Barnett I, Boland MR, Chen Y, Demiris G, Gonzalez-Hernandez G, et al. Ideas for how informaticians can get involved with COVID-19 research. Springer; 2020.
9. Balocchi C, Jensen ST. Spatial modeling of trends in crime over time in Philadelphia. The Annals of Applied Statistics. 2019;13(4):2235-59.
10. Brunsdon C, Comber L. Chapter 12 CALIBRATING SPATIAL REGRESSION MODELS IN R. Code for An Introduction to Spatial Analysis and Mapping in R 2nd edition: bookdown.org; 2019.
11. CensusBureau. Program Income and Public Assistance. 2020;<https://www.census.gov/topics/income-poverty/public-assistance.html -:~:text=Public%20assistance%20includes%20cash%20and,low%2Dincome%20families%20or%20individuals.&text=This%20report%20presents%20data%20on,ACS)%201%2Dyear%20estimates.>(Accessed in August 2020).
12. Martinez DA, Hinson JS, Klein EY, Irvin NA, Saheed M, Page KR, et al. SARS-CoV-2 positivity rate for Latinos in the Baltimore–Washington, DC Region. JAMA. 2020;324(4):392-5.
13. Weinick RM, Jacobs EA, Stone LC, Ortega AN, Burstin H. Hispanic healthcare disparities: challenging the myth of a monolithic Hispanic population. Medical care. 2004:313-20.
14. Page KR, Venkataramani M, Beyrer C, Polk S. Undocumented U.S. Immigrants and Covid-19. New England Journal of Medicine. 2020;382(21):e62.
15. Cordes J, Castro MC. Spatial analysis of COVID-19 clusters and contextual factors in New York City. Spatial and Spatio-temporal Epidemiology. 2020;34:100355.
16. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health. 2020.
17. Cho H, Ippolito D, Yu YW. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. arXiv preprint arXiv:200311511. 2020.
18. Salathé M, Althaus CL, Neher R, Stringhini S, Hodcroft E, Fellay J, et al. COVID-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation. Swiss medical weekly. 2020;150(11-12):w20225-.

Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse

Abraham C. Sanders, Rachael C. White, Lauren S. Severson, Rufeng Ma, MS, Richard McQueen, BS, Haniel C. Alcântara Paulo, Yucheng Zhang, John S. Erickson, PhD, Kristin P. Bennett, PhD, Rensselaer Polytechnic Institute, Troy, New York

Abstract

In this exploratory study, we scrutinize a database of over one million tweets collected from March to July 2020 to illustrate public attitudes towards mask usage during the COVID-19 pandemic. We employ natural language processing, clustering and sentiment analysis techniques to organize tweets relating to mask-wearing into high-level themes, then relay narratives for each theme using automatic text summarization. In recent months, a body of literature has highlighted the robustness of trends in online activity as proxies for the sociological impact of COVID-19. We find that topic clustering based on mask-related Twitter data offers revealing insights into societal perceptions of COVID-19 and techniques for its prevention. We observe that the volume and polarity of mask-related tweets has greatly increased. Importantly, the analysis pipeline presented may be leveraged by the health community for qualitative assessment of public response to health intervention techniques in real time.

1 Introduction

Social media provides a rich corpus of text characterizing in real time the daily happenings and current events within our communities. As such, it has potential utility for individuals and entities wishing to keep their fingers on the pulse of both social and public health issues. Mask-wearing during the COVID-19 pandemic falls into both categories, as the consensus in the scientific community that wearing masks is key to controlling the spread of the SARS-CoV-2 virus[1] has been met with non-negligible resistance for various sociopolitical reasons. Research avenues investigating this mask usage discrepancy are increasingly relevant in light of both the evolution of the coronavirus pandemic into a border-independent global crisis and the extent to which public perceptions of the virus have changed over time.

Background and Related Works: In the pandemic-era reality that has evolved in 2020, social distancing has become the necessary norm, and it is known that social media is playing a bigger role than ever in keeping people connected and informed.[2] Several mainstream social media platforms have seen usage spikes amongst English-speakers since the onset of the pandemic.[3] In keeping with the stimulation of social media activity observed to accompany disease outbreak events, a body of literature has emerged over the past decade that looks specifically at how trends in online activity and discourse can help inform epidemiological models.[4] In conjunction, a suite of programming frameworks and models drawing on data harvested from Twitter have been developed to answer specific research questions about viral trends and their societal impacts.[5–7]

Major Contributions: This analysis aims to provide insight into the broadscale conversation surrounding mask-wearing evolving on Twitter between March and July of 2020, when infection rates initially spiked in the United States, Europe, and other regions throughout the world. To this end, we develop a novel pipeline employing state-of-the-art natural language processing (NLP) techniques in order to systematically characterize Twitter discourse about and public attitudes towards the topic of mask usage during the COVID-19 pandemic. Specifically, we collect and analyze a comprehensive sample of coronavirus-related tweets textually related to mask-wearing. We employ clustering techniques to organize these tweets into fifteen high-level themes and fifteen specific topics within each theme, then perform sentiment analysis on the entire corpus, and also on each theme and topic, across a five-month period. We then apply an abstractive text summarization model using NLP to automatically interpret and describe the subject of the conversation occurring within each theme and topic cluster. We use data visualization and statistical analyses to examine trends in sentiments and divisiveness of the clusters.

Our pipeline is distinct from others recently developed for COVID-19-related information characterization. While other works have primarily drawn from unfiltered Twitter corpora, or alternatively, from manually-annotated datasets specific to a particular hypothesis, we chose to compromise between the two approaches by refining an index of tweets

strictly related to both COVID-19 and masks based on text-based keyword identification. With this semi-selective approach, we highlight the thematic trends that manifest organically in the tweets we have collected, while also ensuring that the global English-speaking conversation surrounding mask-usage during the pandemic is represented.

We find two central, co-occurring trends in the English-speaking Twitterverse by means of the presented pipeline. First, Twitter discourse surrounding mask-wearing within our curated dataset is concluded to grow consistently polarized over time, irrespective of the high-level topic into which it is clustered. Moreover, we find evidence to suggest that sentimentality related to masks and mask-use as expressed on Twitter grew increasingly negative as the pandemic progressed. Cumulatively, we concur that a qualitative, semantic Twitter-based analysis pipeline is capable of revealing striking insights into public responses to the pandemic. We hope that the methods developed here can evolve into tools to help provide rapid real-time assessment of public health measures to inform future interventions.

2 Methods

2.1 Data Collection

We used the Twitter streaming API[8] to collect 189,958,459 original tweets filtered by keywords loosely associated with COVID-19 [1] over a five month period beginning on March 17th, 2020 and ending on July 27th, 2020. We restricted our filter to English-language tweets via the streaming API language filter parameter, and discarded any retweets during this time period. Twitter's API provides access to a representative random sample of approximately 1% of all tweets in near real time, and it has been shown that samples obtained via the API reflect the general content generation patterns of the Twittersphere accurately.[9] We stored all collected tweets in Elasticsearch[10] indices for efficient search and retrieval. Using Elasticsearch, we further filtered our corpus of collected tweets by the criteria that a tweet must include at least one keyword indicating it is strongly associated with COVID-19 and at least one keyword indicating it is strongly associated with mask-wearing. This filter yielded a corpus of 1,013,039 tweets for our analysis.

Although we analyze temporal elements which could be interpreted in a regional context, we chose not to filter our dataset geographically. Users who share their device location represent less than 1% of the Twitter population.[11] Additionally, studies have reported 34% of user-specified profile locations are unusable,[12] and the remainder may not reliably represent tweet origin.[13] We made this decision to avoid any biases that may be present in geotagging Twitter usership and to preclude the risk of inaccuracies inferring tweet location from profile information.

The collected corpus of tweets and the full source code for the data collection analysis pipeline are publicly available at `https://github.com/TheRensselaerIDEA/COVID-masks-nlp`. In compliance with the Twitter content redistribution policy[2], we only provide the tweet IDs corresponding to the collected tweet text used in this work.

Table 1: Filter criteria we used to identify tweets that are related to both COVID-19 and mask-wearing. A tweet must contain at least one keyphrase in both categories to be included.

Keyphrases related to COVID-19	Keyphrases related to mask-wearing
"ncov", "sars-cov-2", "covid", "covd", "covid19", "corona", "virus", "coronavirus", "koronavirus", "wuhancoronavirus", "kungflu", "epidemic", "pandemic", "quarantine", "lockdown", "flatten the curve", "flattenthecurve", "cdc"	"mask", "wearamask", "masking", "N95", "face cover", "face covering", "face covered", "mouth cover", "mouth covering", "mouth covered", "nose cover", "nose covering", "nose covered", "cover your face", "coveryourface"

2.2 Analysis Pipeline

We develop an analysis pipeline to extract, label, summarize, and present the themes, topics and sentiment present in our tweet corpus using state-of-the-art natural language processing tools. While we use it here for analysis of our corpus pertaining to mask-wearing, our methods can be applied to any dataset of text documents. We have included an online supplement [3] containing additional details on implementation decisions and software packages used.

[1] In addition to explicit COVID-19 keywords such as "coronavirus", we include keywords such as "school" and "cancelled" in order to include tweets about a wider array of topics impacted by the pandemic.

[2] Policy can be found at `https://developer.twitter.com/en/developer-terms/agreement-and-policy`

[3] Available at `https://therensselaeridea.github.io/COVID-masks-nlp/paper_supplement.pdf`

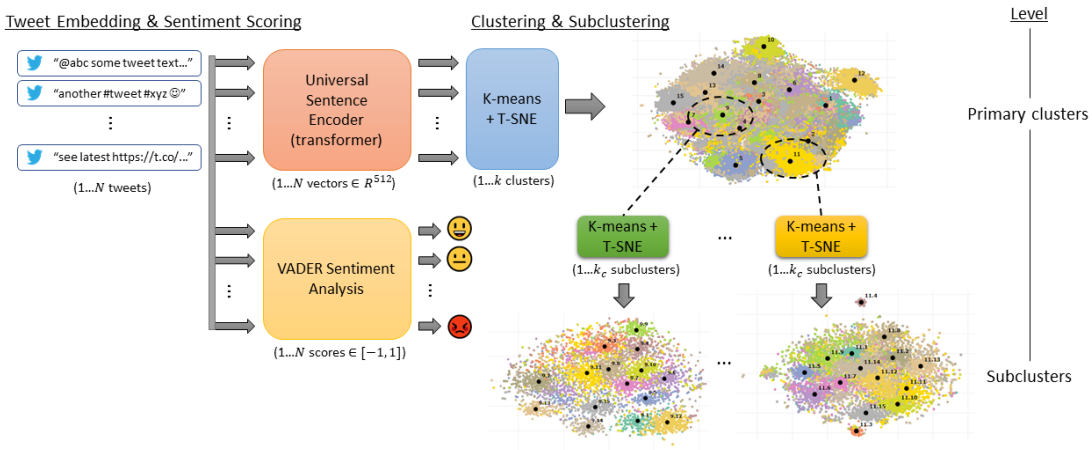


Figure 1: K-means is used to cluster the tweets in their embedding space. A two-level cluster hierarchy is created by applying k-means again to each cluster.

Step 1: Retrieval & Sampling: The first step in the analysis pipeline is the retrieval of a representative random sample of tweets from the corpus. We chose $N = 100,000$ as our sample size for this study, and restricted sampled tweets to those created within the range of March 1st, 2020 to August 1st, 2020 - a sample space of 1,012,815 tweets. Once retrieved, all tweets are cleaned by removing URLs and non-punctuation characters and then normalizing all whitespace character sequences to single spaces.

Step 2: Embedding & Sentiment Scoring: After retrieving and cleaning the sample, each tweet is embedded into a 512-dimensional vector space using the transformer[14] implementation of Google's Universal Sentence Encoder.[15] The vector that represents each tweet is given by the sum of the contextual word representations at each position of the transformer encoder output. Semantically similar tweets are grouped together in the resulting embedding space, where cosine similarity provides a metric of how close two tweets are in meaning.

To assess tweet sentiment, each tweet is also scored using the VADER algorithm - a social-media-centric, lexicon-based sentiment characterization approach.[16] VADER provides a compound polarity score between -1 (most negative) and 1 (most positive). For graphical representations, we use the authors' recommended threshold of ± 0.05 to discretize the score where $s \leq -0.05$ is negative, $-0.05 < s < 0.05$ is neutral, and $s \geq 0.05$ is positive.

Step 3: Clustering & Subclustering: Next, we apply k-means in the embedding space to create a two-level cluster hierarchy - the corpus is grouped into k primary clusters and each primary cluster is then grouped into k_c subclusters. We interpret the primary clusters as representing high-level discussion themes and the subclusters as specific topics within each theme. We re-order the cluster numbers 1 through k and subcluster numbers 1 through k_c by average sentiment score, with 1 being the most negative. To select the optimal number of primary clusters and subclusters, we performed a computational study of the k-means objective function across a range of choices for k and k_c. As documented in our supplement, we selected $k = 15$ and $k_c = 15$ since these values provided a good balance between cluster quality and avoidance of topical redundancy.

We then use t-Distributed Stochastic Neighbor Embedding (t-SNE)[17] to project the clustered embedding space into two dimensions for presentation. In Figure 1, the cluster and subcluster scatterplots use coordinates in $\mathbb{R}^2$ given by t-SNE. The primary cluster plot is color-coded by cluster assignment and the subcluster plots are color-coded by subcluster assignment. The black points represent the cluster and subcluster centers.

We recognize there are many approaches for topic extraction from short text. Our choice to embed tweets with Universal Sentence Encoder and cluster them with k-means in the embedding space was intended to capture context-sensitive representations of tweet text. While our method proves effective, a range of other methods may be used to similar ends. Wang et al. report that k-means combined with sentence embeddings performs comparably to traditional methods such as LDA[18] and TF-IDF on unsupervised Twitter topic modeling, with XLNet[19] and Universal Sentence

Encoder outperforming other models on a variety of metrics.[20]

Step 4: Cluster & Subcluster Labeling: We find keywords that both describe and differentiate the discussion within each cluster and subcluster, and use these keywords as labels. To do this, we compute relative frequencies for words across each cluster, ignoring stopwords and non-alphanumeric characters. Using the relative frequencies, we score each word according to its contribution to the Kullback-Leibler divergence between the word distribution of the cluster and the word distribution of the entire corpus sample: $\text{score}(w) = KL(W_S||W_C) = P(W_S = w)\log\frac{P(W_S = w)}{P(W_C = w)}$. Here, W_C and W_S are the word probability distributions for the corpus sample and sub-sample (cluster), respectively. Subclusters are labeled in the same manner, with the parent cluster taking the place of the corpus sample. Additional illustration of the labeling method is included in the supplement.

A single label representing the corpus sample is computed using the eight words with the highest overall frequencies. For each cluster and subcluster, we select the three words with the highest scores and concatenate them to create theme and topic labels respectively. To avoid reuse of keywords across labels, cluster labels cannot contain keywords that exist in the corpus sample label, and subcluster labels can not contain keywords that exist in the parent cluster label.

Step 5: Cluster & Subcluster Summarization: To augment human interpretations of each cluster and subcluster,

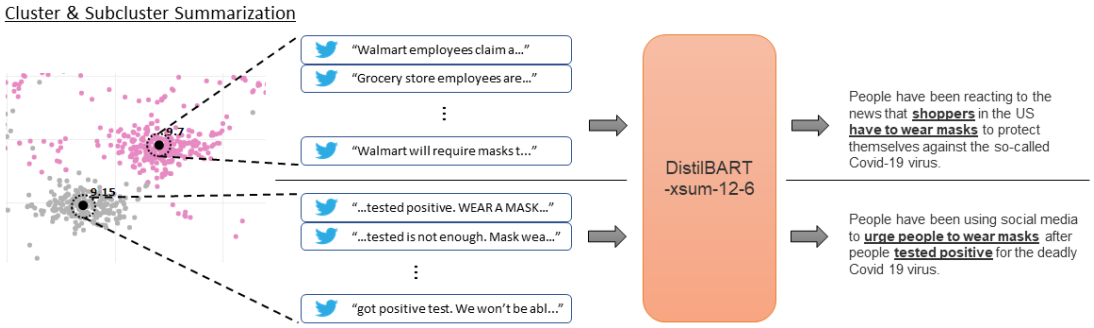


Figure 2: The tweets embedded nearest the subcluster center (shown as a black dot) are used to create the input "article" for DistilBART to summarize.

we generate summaries using DistilBART, an abstractive summarization model from the HuggingFace Transformers[21] package based on Facebook's BART[22] model. While the labels provide a quick description of the type of discussion happening within a cluster or subcluster, the one-to-three sentence summary produced by this process conveys this information in a much more meaningful way. We use a DistilBART instance fine-tuned on the extreme summarization (xsum) task[23] which aims to generate concise summaries of articles without relying on extractive summarization strategies. For each subcluster, we generate the input "article" for DistilBART to summarize by concatenating the text of 20 tweets which are embedded nearest to the subcluster center. For each cluster, we generate the input by concatenating all of the model-generated summaries of its subclusters. We performed a qualitative study to assess two additional strategies for summarizing the clusters, and selected this approach as we found it least prone to misrepresentation of cluster themes. More detail on this study have been included in the supplement.

It should be noted that the xsum dataset used to fine-tune DistilBART is comprised of news articles from the British Broadcasting Corporation (BBC) and their corresponding summaries. Multi-tweet summarization is an admittedly different task domain with noisier, less consistent text. We consider our application of DistilBART in this domain to be proof-of-concept, with the logical next step of fine-tuning on a human-annotated xsum-like dataset for Twitter in keeping with the original methods of Narayan et al.[23]

2.3 Sentiment Analysis

Divisiveness in Sentiment: In order to better understand the sentiment profile of the tweet clusters, we developed a divisiveness score to assess the present level of polarization in tweet sentiment. The score is given by a real number such that polarized samples with little neutral sentiment are given a positive score, while samples with consensus (unimodally concentrated on a single sentiment category) are given a negative score. Otherwise, in the case where sentiment is uniformly distributed across categories, samples have a score of zero. The score is based on the Sarle's Bimodality Coefficient[24] (BC) with an added correction through a weighted average with the BC of the uniform distribution, and then a logit transformation. This weighting counterbalances the large variance of the BC, based on the skewness and kurtosis for small samples,[25] so that such samples with little information are considered to still have uniformly polarized sentiment.

3 Results

We examine the tweet volume and sentiment concerning masks for the entire sample. Table 2 contains sentiment average, overall divisiveness, and trends in divisiveness for each cluster for tweets from March to July 2020. Cluster interpretations in Section 4 further illustrate the nature of the mask discourse.

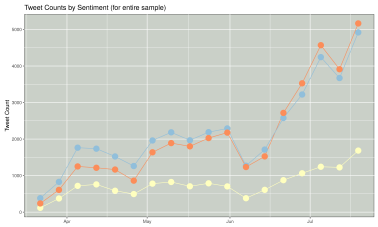
Figure 3a shows the number of negative (red), neutral (yellow) and positive tweets (blue) per week. Clearly the volume and polarity of the discussion have dramatically increased starting in mid-June. Figure 3b shows the labels provided by the keyword analysis for each cluster, ordered from most negative sentiment to most positive sentiment. Figure 3c shows the weekly counts for tweets by sentiment for each week. Clusters 1-3 are the most negative clusters, which, as later detailed in Section 4, respectively discuss the topics of Donald Trump, individuals not wearing masks, and government mask and social distancing mandates.

Table 2: Average sentiment scores, divisiveness scores, and regression line slopes with 95% confidence intervals, and qualitative descriptions of time series trends. Clusters are listed in order of increasing sentiment score.

Cluster	Mean Sentiment	Sentiment 95% CI	Divisiveness Score	Divisiveness LR Slope	Divisiveness LR Slope 95% CI	Trend in Divisiveness Over Time
1	-0.1645	(-0.1767, -0.1522)	1.7472	0.0434	(0.0129 , 0.0740)	Increasing
2	-0.1147	(-0.1263, -0.1031)	2.3017	0.0935	(0.0642 , 0.1227)	Increasing
3	-0.0942	(-0.1071, -0.0811)	2.2086	0.0868	(0.0579 , 0.1157)	Increasing
4	-0.0546	(-0.0657, -0.0434)	2.1962	0.0905	(0.0627 , 0.1184)	Increasing
5	-0.0469	(-0.0589, -0.0347)	1.5292	0.0436	(0.0205, 0.0667)	Increasing
6	-0.0391	(-0.0500, -0.0281)	0.7651	0.0278	(0.0135, 00422)	Increasing
7	-0.0364	(-0.0503, -0.0224)	1.3233	0.0783	(0.0592, 0.0975)	Increasing
8	0.0272	(0.0132, 0.0411)	1.3727	0.0394	(0.0143, 0.0644)	Increasing
9	0.0365	(0.0218, 0.0510)	1.9079	0.0466	(0.0210, 0.0726)	Increasing
10	0.0387	(0.0221, 0.0551)	1.4149	0.0437	(0.0250, 0.0629)	Increasing
11	0.0394	(0.0286, 0.0502)	1.7917	0.0508	(0.0215, 0.0800)	Increasing
12	0.0607	(0.0221, 0.0551)	1.2747	0.0118	(-0.0179 , 0.0416)	*Inconclusive*
13	0.0693	(0.0584, 0.0801)	0.5094	0.0331	(0.0187, 0.04744)	Increasing
14	0.3042	(0.2934, 0.3151)	0.8411	0.0153	(-0.0048, 0.0354)	*Inconclusive*
15	0.3399	(0.3272, 0.3527)	2.4018	0.0694	(0.0242,0.1146)	Increasing

Cluster Divisiveness: To characterize the polarization of each topic cluster and the changes in polarization over time, we perform global and per-week analyses of the divisiveness scores for all clusters. For each cluster we compute divisiveness for each week, then run a linear regression of divisiveness against time; the results are shown in Table 2.

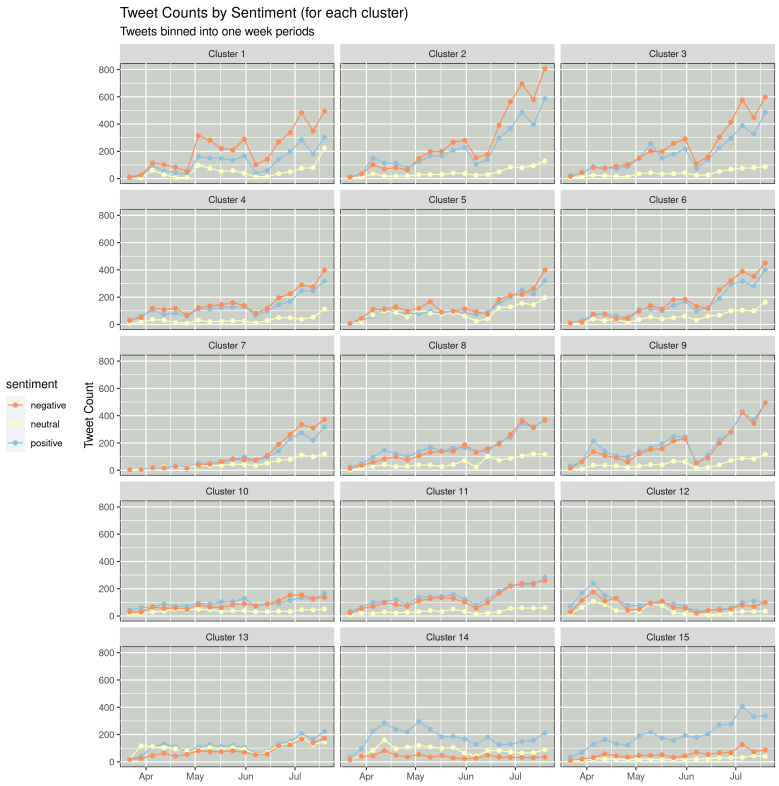
We see that for all clusters, except for Clusters 12 and 14, the confidence intervals for the slope of the fitted lines are entirely positive, indicating an increasing trend in divisiveness over time. However, no clusters display particularly steep trends, with the most significant one being Cluster 13 with a slope equivalent to only 0.0649% of the overall divisiveness score. All clusters are shown to be divisive; however, Clusters 6 and 13 possess the lowest divisiveness scores, while Clusters 2, 3 and 15 are shown to be the most divisive. Cluster 15 in particular is found to have the greatest divisiveness score. However, this result likely comes from a known fault of Sarle's BC when handling heavily



(a) Weekly tweet count by discretized sentiment for the entire sample

1. trump / president / realdonaldtrump
2. vaccine / flu / stop
3. lockdown / social / distancing
4. cdc / public / risk
5. pandemic / global / middle
6. coronavirus / mandate / governor
7. 19 / fucking / real
8. wearamask / maskwearing / masking
9. quarantine / gloves / store
10. corona / hero / gagging
11. droplets / spread / stop
12. n95 / surgical / microns
13. coronavirus / face / make
14. face / facemask / coronavirus
15. hands / wash / stay

(b) Top keywords by cluster



(c) Weekly tweet count by discretized sentiment for each cluster

Figure 3: Sentiment over time, for the entire tweet corpus and for each cluster

skewed distributions.[24] In this case, the divisiveness score is likely incorrectly inflated due to the cluster distribution being heavily skewed towards positive sentiment, shown in Figure 3c. Clusters 2 and 3 then evidently come out to be the most polarizing out of all clusters presented, both also having comparatively large values for the fitted regression line slope with 95% certainty of increasing sentiment divisiveness.

Variance in Sentiment Over Time: A one-way ANOVA was conducted for differences in mask-related sentiment across five consecutive months of early 2020 (March through July) using the complete tweet corpus, $N = 1,013,039$. The ANOVA was performed on the basis of observed normality of residuals, and with the caveat that a Breusch-Pagan test pointed to heterogeneity of variance between months. The test result indicated with significance ($p < 10^{-16}$) the presence of at least one distinct difference in sentiment among the five pandemic months analyzed. A subsequent Bonferroni-corrected pairwise t-test further confirmed statistically significant differences in mean sentiment score across all months studied ($p < 0.001$). In light of this finding, we followed with Dunnett's contrasts[26] to compare the mean sentiment for each group to that of March, the earliest month in our dataset. The results concurred, at $\alpha = .05$, that the mean sentiment scores computed for the months of April through July all differed significantly from that of March, at $p = 0.0143$ for April and $p < 0.001$ for all other months. We re-ran the Dunnett method with the alternative hypothesis that the mean sentiment for each month was greater than or equal to that of March. We failed to reject the null hypothesis of a decrease for each pairwise test. Taken together with the graphical trends found, we interpret this result to suggest an overall decrease in mean sentiment score related to masks and mask-use as of July 2020.

4 Cluster Interpretations

In this section, we select five clusters found to be particularly striking in content. We order the clusters by increasing overall sentiment score, report on the trends in sentiment and divisiveness metrics, and include the automatically-generated summary for each. We then provide manual annotations of the prominent themes that arise, by inspecting

Figure 4: We have made available an interactive document containing the full listing of all clusters, subclusters, and automatically-generated summaries. Our interactive notebook containing results for all clusters can be found at `https://therensselaeridea.github.io/COVID-masks-nlp/analysis/twitter.html`

small samples of tweets lying near each of the fifteen subcluster centers within each cluster. We see that support for mask wearing and cluster sentiment do not necessarily correspond.

Cluster 1: trump / president / realdonaldtrump **(Overall Sentiment : -0.1645 ; Divisiveness : 1.7472)**

DistilBart summary: *People have been reacting to news that President Donald Trump has refused to wear a face mask in public to protect himself from the deadly coronavirus pandemic.*

Interpretation: This cluster (shown in Figure 4) features Twitter users expressing a spectrum of attitudes towards U.S. President Donald Trump. Opinions specifically revolve around Trump's handling of the COVID-19 pandemic in the United States. Distinctly, there exists an evident theme of frustration arising from observations that Trump has refused to wear a mask in public appearances, despite statements from public health officials encouraging the action. In complement, a sizeable positive discussion thread also exists concerning President Trump. A major theme observed here among the pro-Trump tweets is the impression that the media is biased against the president, and that this in turn fosters a public motive to exaggerate the virus.

Cluster 2: vaccine / flu / stop **(Overall Sentiment : -0.1147 ; Divisiveness : 2.3017)**

DistilBART summary: *Following the news that people in the US are being urged to wear face-covering masks to prevent the spread of a new virus that has killed more than 4,000 people in China.*

Interpretation: Cluster 2, "vaccine / flu / stop", is a grim cluster in terms of its overall sentiment, and is distinctly polemical in its semantics. It is found that the majority of tweets sampled from this cluster complain about individuals who don't wear masks. The dominant attitude towards masks is positive, despite the overall negative sentimentality computed for the cluster as a whole. In contrast with the more semantically upbeat "face / hands / stay" cluster (Cluster 15), the theme of death and dying is prevalent. The social nature of disease is a major motif (i.e. "Your actions affect all of us.")

Cluster 3: lockdown / social / distancing **(Overall Sentiment : -0.0942 ; Divisiveness 2.3017)**

DistilBART summary: *Following the news that the US government has ordered people to wear face masks in public to prevent the spread of the deadly Covid-19 coronavirus, people across the world have been reacting to the news on social media.*

Interpretation: Cluster 3 gives an indication of the societal turbulence arising from mask mandates, social distancing enforcement, and similar lockdown-related measures globally. Paradoxically, the overall average sentiment of -0.0941 computed for this cluster is borderline neutral. A strong racial emphasis is evident, with discourse focusing around protests of the #BlackLivesMatter movement, an international phenomenon co-occurring with the coronavirus pandemic mid-year. Several regions of Cluster 3 contain conversations about international responses to the virus, notably around the idea that mask-wearing to prevent the spread of disease agents is a long-standing cultural norm in some regions. In keeping with the slightly negative overall sentiment for this cluster, many of the tweets express sarcasm.

Cluster 12: n95 / surgical / microns **(Overall Sentiment : 0.0693 ; Divisiveness : 1.2747)**

DistilBART summary: *News that a shortage of N95 respirator masks in the US is causing a worldwide shortage has been shared on social media.*

Interpretation: Discourse within Cluster 12 focuses on information about N95 masks and related forms of personal protective equipment (P.P.E.). The evolution of the conversation around the accessibility of medical resources over the timeline of the pandemic is clearly represented. One notable stream of discussion points to the presence of a debate over how effective cloth masks are as guards against infectious agents in comparison to surgical masks. The shortage of respirators experienced by the medical community in the United States is also referenced, as is a change in perspective on N95 masks from the U.S. CDC in the early days of the pandemic.

Cluster 15: hand / wash / stay **(Overall Sentiment : 0.3399 ; Divisiveness : 2.4018)**

DistilBART summary: *Social media users have been sharing their tips and advice on how to prevent the spread of the deadly coronavirus.*

Interpretation: Our most positive cluster, "hand / wash / stay" is composed of tweets sharing tips on prevention measures for stopping the spread of COVID-19. There appears to be highly positive sentiment expressed towards masks and other PPE, and well-meaning admonitions such as "Wash your hands and socially distance!" are frequent. In contrast to other clusters, the Cluster 15 tweets contain little in the way of aggressive, sarcastic or antagonistic semantics. As such, to an extent, this cluster echoes the official messaging of the CDC and similar organizations.

5 Discussion

The objective of our analysis framework was to study the distribution of global mask-related social media discourse, the specific topics within this distribution, their sentimentality trends and how the latter have changed over time. In comparison to the official sources of COVID-19 infection and death rate data, the accessibility and sheer quantity of organic discourse played out over Twitter make this platform an invaluable source of information on public perception of mask usage during the coronavirus pandemic. The cumulative results of our pipeline point to the existence of two central, co-occurring trends in the English-speaking Twitterverse: consistently polarized Twitter discourse surrounding mask-wearing, and an accompanying overall increase in negative sentimentality.

While mask-wearing is a health-related issue, the politicization of mask-wearing is exposed in this investigation. The fact that the U.S. president holds such bearing in the global Twitter conversation about mask-wearing speaks to the degree to which sociopolitical dynamics hold sway over the public perception of the pandemic. The topic-sensitivity of the clustering approach we develop also opens doors for new health-related insights regarding COVID-19's impact.

While our pipeline is effective, opportunities for improvement exist. Any component of the pipeline can potentially be replaced. For example, alternative clustering methods could be used. An open question arising from this research is how well VADER-computed sentiment estimations reflect public opinion in a semantic sense. In this work, we leverage lexicon-based sentiment analysis as a proxy for human attitudes and emotions, but we plan to incorporate a more holistic sentiment representation moving forward (e.g. one capable of detecting expressions of sarcasm).

Two important limitations of our summarization method should be noted. First, the BART-based decoder is a generative language model which creates summaries autoregressively by repeatedly sampling from next-word probability distributions over an entire vocabulary. For this reason, the output summaries are prone to factual inaccuracy in a manner which extractive summarization approaches are not. Second, large or irregularly shaped subclusters may be

poorly represented by the tweets immediately surrounding the subcluster center. In these situations the generated summary may not be applicable to the entire subcluster. We accept these as limitations of the system when used with unannotated data, as is the case in our study. As such, we advise users of the pipeline to regard the summaries as context clues, and then use the notebook provided for further investigation.

6 Conclusion

In light of both the escalation of the pandemic into a global crisis and the extent to which the implications of the virus have changed in the public eye over time, semantic analyses such as we present are increasingly relevant as sources of information to the medical research community for a host of health-related considerations. As we see, mining Twitter data allows for the rapid summarization of opinions about empirically-supported disease prevention measures. Overall, we find that thematic clustering and visualization based on mask-related Twitter data can offer distinct insights into societal perceptions of COVID-19, complementary to findings from more traditional epidemiological data sources. With the aid of abstractive visualizations like the clustering techniques presented, acute estimations of what individuals are actually saying and feeling amidst the viral destruction can be made. As future work, we hope to evolve this pipeline into a valuable tool that can aid health providers and policy makers in understanding public response to health interventions in the ongoing global health crisis. This could include identifying subgroups that are inadequately reached by existing campaigns, as well as predictive modeling of responses to public health messaging to aid health organizations in designing and optimizing outreach campaigns.

Acknowledgements

This study was supported by the Rensselaer Institute for Data Exploration and Applications, the Data INCITE Lab, and a grant from the United Health Foundation.

References

1. D. K. Chu, E. A. Akl, S. Duda, K. Solo, S. Yaacoub, H. J. Schünemann, A. El-harakeh, A. Bognanni, T. Lotfi, M. Loeb *et al.*, "Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis," *The Lancet*, 2020.

2. T. Nabity-Grover, C. M. Cheung, and J. B. Thatcher, "Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media," *International Journal of Information Management*, p. 102188, 2020. [Online]. Available: https://bit.ly/2YzkzIG

3. B. K. Wiederhold, "Social media use during social distancing," 2020.

4. N. E. Kogan, L. Clemente, P. Liautaud, J. Kaashoek, N. B. Link, A. T. Nguyen, F. S. Lu, P. Huybers, B. Resch, C. Havas *et al.*, "An Early Warning Approach to Monitor COVID-19 Activity with Multiple Digital Traces in Near Real-Time," *arXiv preprint arXiv:2007.00756*, 2020. [Online]. Available: https://arxiv.org/abs/2007.00756

5. E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet. Infectious Diseases*, vol. 20, pp. 533 – 534, 2020.

6. S. Zong, A. Baheti, W. Xu, and A. Ritter, "Extracting COVID-19 Events from Twitter," *arXiv preprint arXiv:2006.02567*, 2020. [Online]. Available: https://arxiv.org/abs/2006.02567

7. A. Kruspe, M. Häberle, I. Kuhn, and X. X. Zhu, "Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic," *arXiv preprint arXiv:2008.12172*, 2020.

8. "Consuming streaming data — twitter developer." [Online]. Available: https://bit.ly/32xxtbf

9. Y. Wang, J. Callan, and B. Zheng, "Should we use the sample? Analyzing datasets sampled from Twitter's stream API," *ACM Transactions on the Web (TWEB)*, vol. 9, no. 3, pp. 1–23, 2015. [Online]. Available: https://dl.acm.org/doi/10.1145/2746366

10. "Elasticsearch: The official distributed search analytics engine." [Online]. Available: https://bit.ly/3lqUa9F

11. L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana, "Knowing the tweeters: Deriving sociologically relevant demographics from twitter," *Sociological research online*, vol. 18, no. 3, pp. 74–84, 2013.

12. B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: the dynamics of the location field in user profiles," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 237–246.

13. L. Sloan and J. Morgan, "Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter," *PloS one*, vol. 10, no. 11, p. e0142209, 2015.

14. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

15. D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal Sentence Encoder," *arXiv preprint arXiv:1803.11175*, 2018. [Online]. Available: https://arxiv.org/abs/1803.11175

16. C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in *ICWSM*, 2014.

17. L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

18. M. Hoffman, F. Bach, and D. Blei, "Online learning for latent dirichlet allocation," *advances in neural information processing systems*, vol. 23, pp. 856–864, 2010.

19. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5753–5763.

20. L. Wang, C. Gao, J. Wei, W. Ma, R. Liu, and S. Vosoughi, "An empirical survey of unsupervised text representation methods on twitter data," *arXiv preprint arXiv:2012.03468*, 2020.

21. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *ArXiv*, vol. abs/1910.03771, 2019.

22. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

23. S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," *arXiv preprint arXiv:1808.08745*, 2018.

24. R. Pfister, K. A. Schwarz, M. Janczyk, R. Dale, and J. Freeman, "Good things peak in pairs: a note on the bimodality coefficient," *Frontiers in Psychology*, vol. 4, p. 700, 2013.

25. D. B. Wright and J. A. Herrington, "Problematic standard errors and confidence intervals for skewness and kurtosis," *Behavior Research Methods*, vol. 43, no. 1, pp. 8–17, 2011.

26. S. Lee and D. K. Lee, "What is the proper way to apply the multiple comparison test?" *Korean Journal of Anesthesiology*, vol. 71, no. 5, p. 353, 2018.

Ranking Methodology to Evaluate the Severity of a Quality Gap Using a National EHR Database

Shivani Sivasankar, M.S.,[1,2] An-Lin Cheng, Ph.D.,[1] and Mark Hoffman, Ph.D. [1,2]

[1]University of Missouri-Kansas City School of Medicine, MO; [2]Children's Mercy Hospital, Kansas City, MO

Abstract

Selecting quality improvement projects can often be a reactive process. In order to demonstrate a data-driven strategy, we used multi-site, de-identified electronic health record (EHR) data to prioritize the severity of a quality concern: inappropriate A1c test orders for sickle cell disease patients in two randomly chosen facilities (Facility A & B). The best linear unbiased predictions (BLUP) generated from Generalized Linear Mixed Model (GLMM) was estimated for all 393 facilities with 37,151 SCD patients in the Cerner Health Facts™ (HF) data warehouse based on the ratio of inappropriate A1c orders. Ranking the BLUP after applying the GLMM indicates that the facility A being in the second quartile may not have a quality gap as significant as facility B in the top quartile for this quality concern. This study illustrates the utility of multisite EHR data for evaluating QI projects and the utility of GLMM to enable this analysis.

Introduction

One of the five competencies defined by the Institute of Medicine as essential for all healthcare professionals is the ability to apply principles of Quality Improvement (QI) to evaluate and improve systems performance [1]. The healthcare provider is expected to select appropriate quality indicators and maintain a focus on the areas considered most likely to improve patient care and clinical outcomes [2]. National standards, such as the Centers for Medicare and Medicaid Services (CMS) Core Measures and the National Quality Forum (NQF), provide important guidance for high level quality metrics [3,4]. Often QI initiatives are launched in response to local concerns including error prevention initiatives or efforts to improve time-based metrics [5]. It is not uncommon for the number of potential QI projects to exceed the capacity to complete the work, requiring leaders to prioritize among multiple options. By determining the severity of a quality gap compared to peer organizations, healthcare organizations can strategically prioritize QI projects, focus on the greatest opportunities and allocate time and resources to the more complex, yet critical, qualitative aspects of transforming care to improve population health with patient-centered affordable care [6].

Given the complexity of health care, assessing a quality gap is a dynamic and challenging process. Data is widely available for mandated metrics, enabling organizations to determine whether they are aligned with national averages or deviate substantially from the mean. For example, the National Quality Strategy developed by the NQF provides guidance on prioritizing performance measure gaps for adult immunizations [7]. The CMS has defined core sets of quality measures in specific clinical areas, including cardiology, gastroenterology, HIV and hepatitis C, medical oncology, obstetrics and gynecology, orthopedics, Pediatric Accountable Care Organizations (ACOs), Patient Centered Medical Homes (PCMH), and primary care [8]. However, high quality data to support prioritization of potential QI projects for many other clinical processes are not as widely available.

Other than attaining nationally mandated quality metrics, most quality improvement initiatives are based on local concerns and can often be reactive to a real or perceived crisis [9–11]. Contextualizing the severity of a quality concern can be difficult in the absence of data from peer institutions. Electronic Health Record (EHR) systems can advance the quality of care by providing access to patient health information, monitoring compliance with standard measures of quality so that patients receive guideline recommended care and reduce medical errors through decision support [12]. Individual organizations may use a local data warehouse populated by EHR data to evaluate the quality concern based on previous performance within that institution, for example when often quality improvement (QI) projects are selected in reaction to a recent adverse event [13,14]. While this is frequently useful, local data warehouses lack the broader data necessary to contextualize the severity of a quality concern. This limitation may result in failure to recognize deeper systemic quality gaps in which an organization varies more significantly compared to their peers. Large databases of aggregate, de-identified, data from healthcare providers throughout the U.S. can be useful to compare measures of outcome and practice across several facilities.

Multi-institutional data warehouses (MDW) are distinguished by the consolidation of disparate EHR data sources across several independent, non-affiliated organizations and their use of data structures that are optimized for

queries[15]. Health care facilities can achieve improvement in patient outcomes by participating in MDW [16]. Utilization of a MDW is an effective path to improvement as it encompasses a multifaceted strategy, the ultimate target of which is to decrease the uncertainty inherent to the process. Some of the commonly used aggregate databases include the Nationwide Inpatient Sample (NIS), the Kids' Inpatient Database (KID), the Nationwide Emergency Department Sample (NEDS), the Pediatric Health Information System (PHIS) and the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) [17–21]. These databases contain detailed, longitudinal analysis of hospital encounters and serve as a valuable resource for research and QI. Another such data resource, Cerner Health Facts (HF), has been demonstrated to have frequency of diagnoses codes consistent with the HCUP National Inpatient Survey [22], and has been found useful to evaluate national trends of several health care related research questions [23–26]. HF is distinguished by deeper granularity than the other examples. One challenge in using MDW is accounting for varying characteristics of the contributing facilities. HF includes large and small providers, urban and rural facilities, public and private organizations and geographically diverse contributing sites.

When patients are nested in clusters such as facilities, observations within the same cluster are likely to be correlated. Generalized linear mixed models (GLMM) including both fixed and random effects have been proposed to analyze correlated data which could adjust for facility level variation. Best Linear Unbiased Prediction (BLUP) estimated from the model after adjusting the bias can be used as an index to predict the severity of QI concern. As a proof of concept, in this study we use this methodology to evaluate and rank one of the global quality concerns in health care, inappropriate use of A1c orders for sickle cell disease patients. The glycated hemoglobin test (A1c) test to assist in the diagnoses and management of diabetes may result in false results for sickle cell disease patients due to the abnormal hemoglobin structure and shorter lifespan of their red blood cells [27–31]. Professional organizations such as the American Diabetes Association, National Institute of Diabetes and Digestive and Kidney Diseases and the National Glycohemoglobin Standardization Program recommend the use of alternative tests instead of A1c for patients with SCD [32–34]. We describe a novel application of a multi-institutional EHR data warehouse by resolving the requirement to handle variations between facilities in order to contextualize the severity of inappropriate A1c orders in sickle-cell patients by comparing two random facilities among its peers. We also determine the extent to which facility characteristics contribute to this quality concern.

Materials and Methods

Data source

This study used the de-identified Health Facts data warehouse (Cerner Corporation, Kansas City, MO), which contains longitudinal patient data systematically extracted from the EHR at participating institutions and includes encounter data (emergency, outpatient, and inpatient), patient demographics (age, sex, and race), diagnoses and procedures, laboratory data and facility characteristics (census region, number of beds, acute setting and teaching versus nonteaching status). The release used for this work (2016) consists of 386.2 million encounters and 4.3 billion lab results from 64 million patients in 863 healthcare facilities. All admissions, in patient medication orders and dispensing, laboratory orders and specimens are date and time stamped, providing a temporal relationship between treatment patterns and clinical information. All data are de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA) before being provided to the investigators. Longitudinal relationships between patient encounters within the same health system are preserved. The University of Missouri Kansas City (UMKC) Institutional Review Board has determined that work with Health Facts is considered non-human subjects research.

Extraction of study cohorts from database

All patients with a sickle cell disease (SCD) diagnosis were included in the study cohort based on the International Classification of Diseases, Ninth Revision and Tenth Revision, Clinical Modification codes (ICD-9-CM and ICD-10-CM). These codes were selected based on clinical judgement and the criteria specified in the Phenotype Knowledgebase (PheKB) resource which identifies patients with confirmed SCD diagnosis with a positive predictive value of 99.4% and a sensitivity of 99.4% using ICD-9-CM diagnosis codes [35]. The resulting definition groups (from ICD-9-CM codes) were combined with ICD-10-CM codes to identify the sickle cell disease patient cohort. Encounters before 2011 and after 2016 were excluded from the analysis because Health Facts data architecture was updated in 2008-2009, and the data for 2017 was incomplete. This cohort was evaluated for the presence of A1c orders placed after the first sickle cell diagnosis. Laboratory tests in Health Facts are associated with Logical Observation Identifier Names and Codes (LOINC) values. The LOINC codes which were used to indicate A1c orders were 55454-3, 41995-2, 4548-4, 17855-8, 4549-2, 17856-6. This SCD patient cohort was divided into two groups based on the

presence or absence of A1c encounters: SCD patients with at least 1 A1c encounter and SCD patients without any A1c encounters.

Model building

Two facilities (Facility A and facility B) were randomly chosen to observe the difference in their ranking before and after applying the GLMM. In order to contextualize QI concerns, we described our measures as rates, with the numerator indicating how many times the measure has been met and the denominator indicating the opportunities to meet the measure. The outcome measure for this analysis is the ratio of SCD patients with A1c test orders over the total number of SCD patients in the facility. Contextualizing the two random facilities relative to the other facilities in HF posed two challenges. First, the ranking process needed to be adjusted to address the repeated measures of the facility and facility level differences as each facility can vary from small ambulatory clinics to large hospitals with more than 500 beds. Second, the variation in the denominator (total number of SCD patients) between the facilities need to be accounted for, as each facility encompasses differing SCD patient populations. A single level logistic regression model which predicts outcome from the facility level predictor was extended to multilevel analysis such that each facility had its own intercept in the model. This was used to account for the repeated measures for each facility. The behavior of a facility level outcome was examined as a function of facility level predictors. The logit model specified a linear function at the logit (log odds) scale. The generalized linear mixed model (GLMM) can address the two challenges identified above by including facility level covariates and including longitudinal data of the facility as random effect. First, we identified the facility level attributes available in Health Facts: census region, teaching status, urban or rural status, acute status and bed size range. The second step was to include facility as random effect in the model. The third step was to calculate the Best Linear Unbiased Predictors (BLUP), estimated from the realized values of the random variables that are linear functions of the data. The computed BLUP's are unbiased as the average value of the estimate is equal to the average value of the outcome being estimated and they have minimum mean squared error within the class of linear unbiased estimators [37]. The number of sickle cell patients at each facility were included in the model to account for the different number of SCD patients observed at each facility. This was parametrized into the following model:

$$\text{Outcome} = \gamma_{00} + \gamma_{01}(\text{Number of sickle patients}) + \gamma_{02}(\text{census region}) + \gamma_{03}(\text{teaching facility}) + \gamma_{04}(\text{urban rural status}) + \gamma_{05}(\text{acute status}) + \gamma_{06}(\text{bed size range}) + \mu_{0j}$$

From this model, γ indicates the fixed effects and μ_{0j} indicates variation between facilities which represents the random effect.

Analysis

The analysis was performed using SAS statistical software version 9.4 (SAS Institute, Cary, NC, USA). The model was fit into SAS using the PROC GLIMMIX procedure with the link function equal to logit for proportion outcome [38]. The fixed effects of the number of sickle cell patients, census region, teaching facility, urban or rural status, acute status and bed size range were included in the MODEL statement after the outcome variable. Using this model, facility-specific linear unbiased percentages were generated by year. This represents the percentage of inappropriate A1c orders in each facility every year between 2011 to 2016 after accounting for the random effects inherent within the facilities. The estimated percentages for every hospital were ranked to identify the distribution of the facilities in a quartile. Two facilities were randomly chosen to observe the difference in their ranking before and after applying the GLMM.

Results

Baseline characteristics of data

Among the 863 facilities in the version of Health Facts used for this study, 393 facilities had a sickle cell population. These facilities were distributed across all 4 major geographic regions in the United States: Midwest (79 facilities), Northeast (68), South (162), and West (84). Most of the facilities (239) have a bed size less than 100, while some (133) have bed size between 100 to 500 and a few (21) have more than 500 beds. Most of the facilities (300) are urban while the rest (93) are classified as rural. The study cohort includes 37,151 sickle cell patients. The two randomly chosen facilities were large acute care facilities with a bed size greater than 500. Facility A is an urban facility in the South census region while Facility B is a rural facility in the Northeast census region. There were 1,090 SCD patients treated at facility A and 1,267 patients at facility B. The baseline characteristics of the patient population in the HF cohort and the two facilities are delineated in Table 1. Ten percent of the total patients in the HF SCD cohort (3,931

patient) had at least one A1c encounter. Facility A has a lower proportion of SCD patients with A1c encounters (15%, 168 patients) when compared to facility B (32%, 399 patients).

Table 1. Baseline characteristics of the SCD patient population in the HF cohort and across the six facilities in KCMO

Characteristics	**No. of SCD patients (%)**		
	HF	**Facility A**	**Facility B**
Patients without A1c	33,224(89)	922 (85)	868 (68)
Patients with A1c	3,927(11)	168 (15)	399 (32)
Median Age (IQR)	26 (12-47)	25 (20-38)	33 (24-45)
Gender			
Male	16,235 (44)	505 (46)	487 (38)
Female	20,821 (56)	585 (54)	780 (62)
Race			
African American	23,810 (64)	990 (91)	965 (76)
Caucasian	8,951 (24)	59 (5)	124 (10)
Other	4,390 (12)	41 (4)	178 (14)
Year			
2011	8,331 (22)	262 (24)	381 (30)
2012	8,643 (23)	313 (29)	501 (40)
2013	11,934 (32)	473 (43)	537 (42)
2014	13,294 (36)	478 (44)	562 (44)
2015	11,518 (31)	483 (44)	548 (43)
2016	11,092 (30)	375 (34)	529 (42)

Ranking the two random facilities among other facilities in the HF cohort

Ranking the two facilities based on the ratio of SCD patients with A1c orders shows that facility A is in the first (lower) quartile (< 25th percentile) in the years 2011 to 2013 and in the second quartile (< 50th percentile) in the years 2014 to 2016 while facility B is in the third quartile (< 75th percentile) in all the years between 2011 to 2015 and in the second quartile in the year 2016 (Table 2). The position of the facility A and facility B among the range of the ratio of SCD patients with A1c tests from all facilities in HF for the year 2016 is depicted in Figure 1 where both the facilities are in the second quartile.

Table 2: The percentile values for the quartiles based on the range of the ratio of SCD patients with A1c test in all the facilities and the ratio of SCD patients with A1c test in facility A and facility B for the years 2011 to 2016

Year	Ratio of SCD patients with A1c test				
	HF cohort			Facility A	Facility B
	25th Percentile	50th Percentile	75th Percentile		
2011	0.057	0.095	0.197	0.026	0.162
2012	0.11	0.171	0.333	0.030	0.197
2013	0.115	0.205	0.377	0.098	0.240
2014	0.133	0.25	0.5	0.144	0.274
2015	0.116	0.2	0.394	0.139	0.226
2016	0.108	0.225	0.5	0.129	0.193

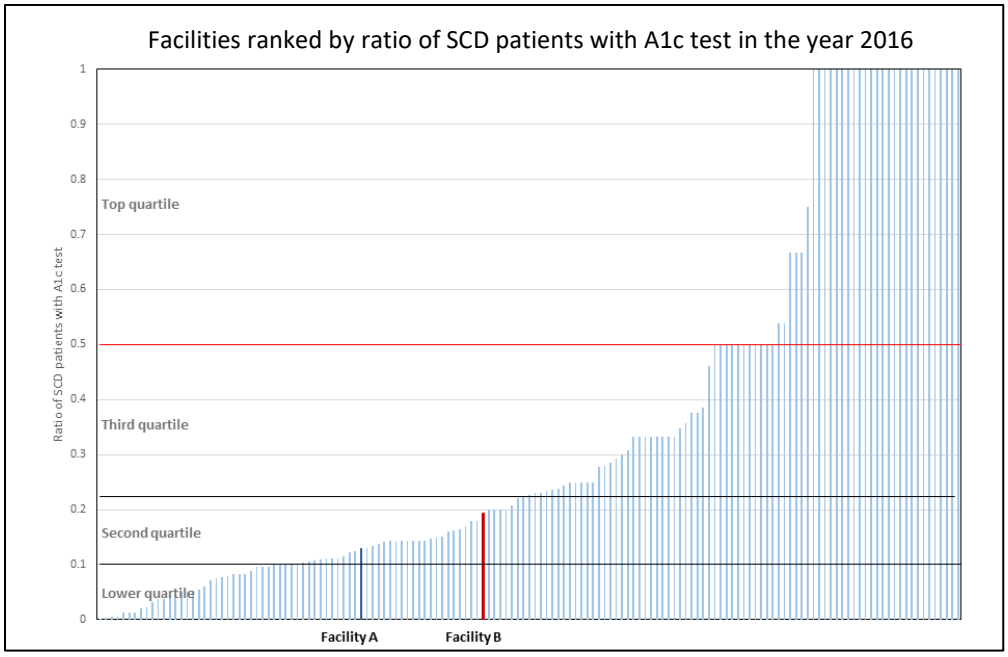


Figure 1. Facility A and facility B are within the second quartile among the range of the ratio of SCD patients with A1c tests from all facilities in HF for the year 2016

The estimated variance of the facility intercept from the empty GLMM model (without covariates) is 2.3501 with a SE of 0.1014. The fixed effect of the intercept has an estimate -2.3501 which represents the grand mean across encounters is statistically significant at p<.0001. After adding the facility level covariates, the estimated variance of the facility intercept is 1.8590 with a SE of 0.2153. By subtracting the total variance between the null model and predictor model it was determined that the predictors explain 21% of the total variance. After adjustment of confounders, facilities in the south census region are 75% less likely to have inappropriate A1c order when compared to the west region (OR, 0.251; 95% CI, 0.150-0.418; p<.00001). Acute care facilities are 72% less likely to have inappropriate A1c orders than Non-Acute care facilities (OR, 0.280; 95% CI, 0.127-0.618; p = 0.0017). Small facilities (bed size <5) are 83% less likely to have inappropriate A1c orders than large facilities (bed size > 500) (OR, 0.173; 95% CI, 0.062-0.486; p = 0.0009) (Table 3).

Table 3. Fixed effects in Generalized Linear Mixed Model (GLMM) for the relationship between facility level predictors and inappropriate A1c test orders in SCD patients

Facility characteristics		**AOR (95% CI)**	**P-value**
Number of SCD patients		1.000 (0.999-1.000)	0.1109
Census Region (Ref- Northeast)			
	Midwest	0.791 (0.450-1.390)	0.4142
	South	0.251 (0.150-0.418)	<.0001
	West	1.108 (0.617-1.988)	0.7318
Teaching status (Ref – Non-Teaching)		0.783 (0.475-1.291)	0.3376
Rural status	(Ref- Urban)	0.726 (0.468-1.126)	0.1525
Acute status	(Ref- Non-Acute)	0.280 (0.127-0.618)	0.0017
Bed Size	(Ref – 500+)		
	<5	0.173 (0.062-0.486)	0.0009
	05-99	1.150 (0.495-2.674)	0.7446
	100-199	1.102 (0.491-2.473)	0.8139
	200-299	1.022 (0.470-2.221)	0.9558
	300-499	0.943 (0.415-2.145)	0.8894

Ranking the BLUP generated for each facility by using the solution for the random effects to estimate the prediction probability for inappropriate A1c orders shows that facility A is less than the 50th percentile value and facility B is greater than the 75th percentile value for all the years between 2011 to 2016 (Table 4). The position of the facility A and facility B among the range of BLUP values estimated through GLMM from all the facilities in HF for the year 2016 is depicted in Figure 2 where facility A remains unchanged and present in the second quartile, which signifies that over-utilization of inappropriate A1c orders is a lower priority in this facility. Whereas, facility B is in the last quartile which signifies that this quality concern is a higher priority in this facility.

Table 4: The percentile values for the quartiles based on the range of the BLUP values estimated from GLMM in all the facilities and BLUP value of facility A and facility B for the years 2011 to 2016

Year	Ratio of SCD patients with A1c test				
	HF cohort			Facility A	Facility B
	25th Percentile	50th Percentile	75th Percentile		
2011	0.041	0.098	0.182	0.092	0.217
2012	0.035	0.093	0.174	0.092	0.211
2013	0.038	0.096	0.018	0.090	0.209
2014	0.037	0.104	0.208	0.090	0.208
2015	0.034	0.095	0.187	0.090	0.209
2016	0.037	0.097	0.185	0.092	0.210

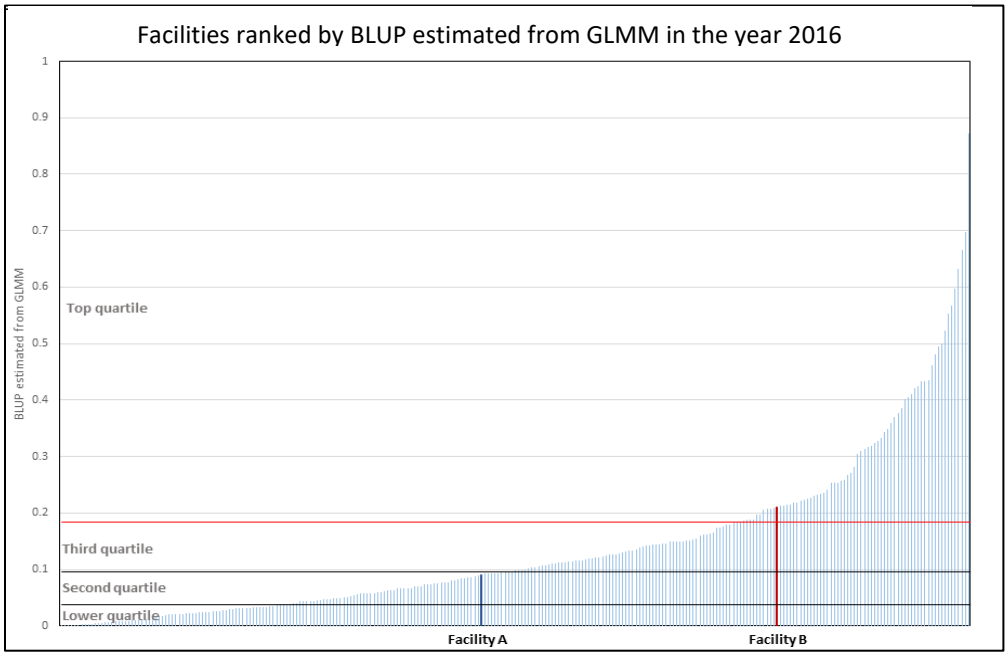


Figure 2. Facility A is in the second quartile and facility B is in the top quartile among the range of BLUP values estimated through GLMM from all the facilities in HF for the year 2016

Discussion

Using comparative multi-site EHR data to prioritize QI projects offers a powerful strategy to help members of a practice understand where their performance falls in comparison to others for topics that are not represented in widely available national data sets. As a proof of concept, we evaluated A1c test orders for sickle cell disease patients in two randomly selected facilities.

There are different techniques which local health provider organizations use to identify and prioritize QI concerns. CDC has provided a list of formalized techniques that can facilitate an orderly process [39]. Other standardized methods such as priority matrix, Hanlon method and analytic hierarchy process techniques use priority scores to

considerably narrow down the priority of the QI concerns [40,41]. However, measuring the A1c tests in SCD patients as a quality concern in the national context using a multi-site EHR provides a complementary external dimension that generates an outward perspective which surpasses preliminary prioritization techniques [42]. It not only identifies quality gaps, but also provides a deep understanding of the level of skill and processes that are required comparable to that of superior performance.

In order to apply this strategy, it was necessary to account for variations between the sites contributing the data set. We noted that many sites in our cohorts had low numbers of eligible patients or low patient volume compared to other facilities. GLMM and BLUP have been demonstrated to be useful for similar challenges in ranking and selection in the context of animal [43], plant breeding [44], to predict an individual's risk of developing cancer [45] and in genetics [37] after taking into account of variation associated with the environmental factors. The quality measurement plan (QMP), which regards probability of distribution of an outcome as the true quality index, is an example of BLUP where non-normal distributions are assumed [46]. It is a useful technique when the ideal ranking involves random effects. Ranking the predicted probabilities of BLUP estimated from GLMM for all facilities offers a novel solution to this issue as it accounts for the variability associated with each level of the outcome by including facility attributes as covariates.

Bias due to the variations between the facilities in the EHR was observed when the facilities were ranked based only on the ratio of SCD patients with A1c orders over the total number of SCD patients within that facility per year. There were many facilities with a higher or lower ratio irrespective of the number of patients in the facility. For example, in the year 2016, 122 facilities had a ratio of 0 which indicates A1c tests are not ordered for SCD patients in those facilities. However, the number of SCD patients present in those 122 facilities ranged between 1 to 194 which should have been considered and appropriately ranked. A facility with 194 SCD patients and without a single inappropriate A1c order should be ranked higher than a facility with just 1 SCD patients and no A1c test order. There were also 26 facilities with a ratio of 1, for which the range of SCD patients was between 1 to 11. Furthermore, there may be considerable variation in the observations due to the differences in the characteristics between facilities such as variation in the type of specialty status, geographic location and size of the facility, which were not accounted for when the unadjusted ratio data was used to rank the facilities. Due to these limitations, ranking the facilities based on the ratio of the outcome could be distorted where overutilization of A1c orders at Facility B may not be considered a high priority quality gap as it is in the third quartile.

The extent to which facility characteristics predict inappropriate A1c orders was determined by the application of GLMM. Facility level predictors correlate with 21% of the total variance. Census region, bed size range and acute status were significant risk factors. The significance of the BLUP over the ratio of outcome can be observed when the GLMM assigns different BLUP values to facilities with the same ratio based on their facility characteristics and number of patients/encounters. Facility A and B with a similar bed size range, were in the same second quartile for the year 2016 when the facilities were ranked based on the ratio of SCD patients with A1c test. This is because of the high number of facilities with a ratio of zero and one which were properly segregated based on the number of SCD patients and facility characteristics. For example, in the year 2016, a facility with only 1 SCD patient without any A1c test has the same ratio of 0, as does a facility with 194 SCD patients without any A1c tests. However, the BLUP value for the latter facility is 0.00013 and is ranked higher than the prior facility with a BLUP value of 0.113. From another example, among two facilities with a similar SCD population (180 SCD patients) and no A1c encounters, the larger facility (bed size > 100) is ranked higher than the smaller facility (bed size < 100) in all six years which demonstrates that this model accounts for facility level characteristics to assign ranks.

Ranking the predicted probabilities estimated from BLUP of all the facilities shows that the QI initiative for A1c test orders for sickle cell patients in facility B should have a higher priority (fourth quartile) than compared to facility A (second quartile), if national context was the primary factor in the prioritization (Figure 2). This strategy addresses the limitations associated with the previous model by adjusting the outcome measure between and within facilities. The outcome measures are meaningful only when adequately adjusted for facility level characteristics and the differences between the patient and procedural characteristics of the facilities. In this application, the data structures are longitudinal (i.e., across multiple facility encounters) and the analyses involve repeated measures at several years which are modelled by GLMM. It accounts for the non-normal distribution of the dependent variable, its restricted range, the relation between mean and variance and includes both fixed and random effects of the varying covariates [47]. The BLUP methodology uses the solution for the random effects in GLMM and produces the best estimator of probabilities to predict the outcome variable in a facility [48].

Based on the GLMM analysis, Facility B would prioritize reduction of A1c tests in SCD patients while facility A may not. In addition to facility B, there are 68 facilities in the HF cohort for the year 2016 which are ranked in the bottom quartile with respect to inappropriate A1c orders, suggesting that these two quality concerns are prevalent in many facilities. Further review of this data could provide insights into the institutional and provider behaviors associated with these ordering patterns. Additional confounders may include differences in experience and practice in providing A1c testing among Health Facts facilities, or perhaps a higher proportion of patients at risk for diabetes are acquired at these bottom quartile facilities than other healthcare entities. However, provider characteristics and comorbidities of the SCD patients were not included in this model. Another limitation is analyzing de-identified EHR data is challenging as there might be errors in capturing the complete SCD cohort. There is no way to verify the accuracy of diagnostic codes (ICD-9-CM, ICD-10-CM) used in Health Facts which are for administrative purposes and hence, are unreliable when compared to gold standard clinician extraction [49,50]. Our findings suggest the need for interventions to reduce or eliminate the use of A1c tests in SCD patients for the diagnoses or management of diabetes.

Conclusion

Analyzing multi-institutional EHR data, after adjusting the bias due to the covariates by using GLMM to generate BLUP's, provided a useful way to evaluate the severity of QI project (Figure 2). The GLMM in this study offered several advantages. First, the linear unbiased predicted percentages from GLMM adjusted the bias due to covariates. Second, GLMMs were able to model the longitudinal structure of the data. Third, by including the facility as a random variable in the mixed model, we were able to generalize the inference of the fixed effect to the population. Additionally, one favorable property of the BLUP generated from the GLMM is shrinkage towards the mean, which anticipates regression of the facility characteristics to the mean of the outcome observed by every facility. Finally, we developed an analytic pipeline which can be easily adopted to different measurements of outcome variables, such as ordinal or nominal outcome. While our focus was prioritizing QI projects between facilities, the methods described here are useful for any comparison of diverse facilities in a multi-contributor EHR data warehouse.

Acknowledgments

This work was funded by CDC Cooperative Agreement NU47OE000105-01-01. We wish to acknowledge the contributions of Suman Sahil from University of Missouri-Kansas City as well as Dr. Kamani Lankachandra from Truman Medical Center for early involvement in this project

References

1. Lohr KN. Medicare: a strategy for quality assurance. Vol. 1. National Academies Press; 1990.
2. Plebani M, Sciacovelli L, Marinova M, Marcuccitti J, Chiozza ML. Quality indicators in laboratory medicine: a fundamental tool for quality and patient safety. Clin Biochem. 2013;46(13–14):1170–4.
3. Medicare C for, Baltimore MS 7500 SB, Usa M. Overview [Internet]. 2018 [cited 2019 Jan 30]. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/index.html
4. NQF: Improving Healthcare Quality [Internet]. [cited 2019 Jan 30]. Available from: https://www.qualityforum.org/Setting_Priorities/Improving_Healthcare_Quality.aspx
5. Muething SE, Goudie A, Schoettker PJ, Donnelly LF, Goodfriend MA, Bracke TM, et al. Quality Improvement Initiative to Reduce Serious Safety Events and Improve Patient Safety Culture. Pediatrics. 2012 Aug;130(2):e423–31.
6. Institute for Healthcare Improvement: The IHI Triple Aim [Internet]. [cited 2019 Jan 30]. Available from: http://www.ihi.org:80/Engage/Initiatives/TripleAim/Pages/default.aspx
7. NQF: Priority Setting for Healthcare Performance Measurement: Addressing Performance Measure Gaps for Adult Immunizations [Internet]. [cited 2019 Jan 30]. Available from: http://www.qualityforum.org/Publications/2014/08/Priority_Setting_for_Healthcare_Performance_Measurement__Addressing_Performance_Measure_Gaps_for_Adult_Immunizations.aspx
8. Medicare C for, Baltimore MS 7500 SB, Usa M. Core Measures [Internet]. 2017 [cited 2019 Jan 30]. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/Core-Measures.html
9. Safety I of M (US) C on DS for P, Aspden P, Corrigan JM, Wolcott J, Erickson SM. Adverse Event Analysis [Internet]. National Academies Press (US); 2004 [cited 2018 Oct 20]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK216102/

10. Heavner JJ, Siner JM. Adverse Event Reporting and Quality Improvement in the Intensive Care Unit. Clin Chest Med. 2015 Sep;36(3):461–7.
11. Rafter N, Hickey A, Condell S, Conroy R, O'Connor P, Vaughan D, et al. Adverse events in healthcare: learning from mistakes. QJM Int J Med. 2015 Apr 1;108(4):273–7.
12. Rumball-Smith J, Shekelle PG, Bates DW. Using the Electronic Health Record to Understand and Minimize Overuse. JAMA. 2017 17;317(3):257–8.
13. Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing Observations from Electronic Medical Record Flowsheets in an i2b2 based Clinical Data Repository to Support Research and Quality Improvement. AMIA Annu Symp Proc. 2011;2011:1454–63.
14. Himes BE, Dai Y, Kohane IS, Weiss ST, Ramoni MF. Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records. J Am Med Inform Assoc. 2009 May 1;16(3):371–9.
15. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data. Med Care. 2012 Jul;50(Suppl):S49–59.
16. Evans RS, Lloyd JF, Pierce LA. Clinical Use of an Enterprise Data Warehouse. AMIA Annu Symp Proc. 2012 Nov 3;2012:189–98.
17. NIS Database Documentation [Internet]. [cited 2019 Jan 31]. Available from: https://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp
18. KID Database Documentation [Internet]. [cited 2019 Jan 31]. Available from: https://www.hcup-us.ahrq.gov/db/nation/kid/kiddbdocumentation.jsp
19. NEDS Overview [Internet]. [cited 2019 Jan 31]. Available from: https://www.hcup-us.ahrq.gov/nedsoverview.jsp
20. PHIS [Internet]. [cited 2019 Jan 31]. Available from: https://www.childrenshospitals.org/phis
21. ACS National Surgical Quality Improvement Program [Internet]. American College of Surgeons. [cited 2019 Jan 31]. Available from: https://www.facs.org/quality-programs/acs-nsqip
22. DeShazo JP, Hoffman MA. A comparison of a multistate inpatient EHR database to the HCUP Nationwide Inpatient Sample. BMC Health Serv Res. 2015 Sep 15;15:384.
23. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed Res Int. 2014;2014:781670.
24. Petrick JL, Nguyen T, Cook MB. Temporal trends of esophageal disorders by age in the Cerner Health Facts database. Ann Epidemiol. 2016 Feb;26(2):151-154.e4.
25. Yang H, Chaudhari P, Zhou Z-Y, Wu EQ, Patel C, Horn DL. Budget impact analysis of liposomal amphotericin B and amphotericin B lipid complex in the treatment of invasive fungal infections in the United States. Appl Health Econ Health Policy. 2014 Feb;12(1):85–93.
26. Shorr AF, Myers DE, Huang DB, Nathanson BH, Emons MF, Kollef MH. A risk score for identifying methicillin-resistant Staphylococcus aureus in patients presenting to the hospital with pneumonia. BMC Infect Dis. 2013 Jun 6;13:268.
27. Smaldone A. Glycemic control and hemoglobinopathy: when A1C may not be reliable. Diabetes Spectr. 2008;21(1):46–9.
28. Lippi G, Targher G. Glycated hemoglobin (HbA1c): old dogmas, a new perspective? Clin Chem Lab Med. 2010 May;48(5):609–14.
29. Thom CS, Dickson CF, Gell DA, Weiss MJ. Hemoglobin Variants: Biochemical Properties and Clinical Correlates. Cold Spring Harb Perspect Med [Internet]. 2013 Mar [cited 2018 Oct 11];3(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3579210/
30. Bry L, Chen PC, Sacks DB. Effects of hemoglobin variants and chemically modified derivatives on assays for glycohemoglobin. Clin Chem. 2001 Feb;47(2):153–63.
31. Speeckaert M, Van Biesen W, Delanghe J, Slingerland R, Wiecek A, Heaf J, et al. Are there better alternatives than haemoglobin A1c to estimate glycaemic control in the chronic kidney disease population? Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc. 2014 Dec;29(12):2167–77.
32. Association AD. Standards of medical care in diabetes—2014. Diabetes Care. 2014;37(Supplement 1):S14–80.
33. Sickle Cell Trait & Other Hemoglobinopathies & Diabetes (For Providers) | NIDDK [Internet]. National Institute of Diabetes and Digestive and Kidney Diseases. [cited 2018 Feb 16]. Available from: https://www.niddk.nih.gov/health-information/diagnostic-tests/sickle-cell-trait-hemoglobinopathies-diabetes

34. NGSP: Factors that Interfere with HbA1c Test Results [Internet]. [cited 2017 May 24]. Available from: http://www.ngsp.org/factors.asp
35. Michalik DE, Panepinto JA| Medical College of Wisconsin| Sickle Cell Disease| PheKB [Internet]. 2017 [cited 2020 Dec 15]. Available from: https://phekb.org/phenotype/615
36. Robinson GK. That BLUP is a Good Thing: The Estimation of Random Effects. Stat Sci. 1991;6(1):15–32.
37. Schabenberger O. Introducing the GLIMMIX procedure for generalized linear mixed models. SUGI 30 Proc. 2005;196.
38. CDC - National Public Health Performance Standards - STLT Gateway [Internet]. 2018 [cited 2019 Feb 15]. Available from: https://www.cdc.gov/od/ocphp/nphpsp/documents/Prioritization.pdf
39. Wang H, Xie M, Goh TN. A comparative study of the prioritization matrix method and the analytic hierarchy process technique in quality function deployment. Total Qual Manag. 1998;9(6):421–30.
40. Rowan L, Walsh R, Durrant L. Creating national emergency pediatric health care priorities through a modified nominal group process and Hanlon method. In: Proceedings of the 132nd Annual American Public Health Association Meetings. 2004.
41. Hutton R. Effective benchmarking through a prioritization methodology. Total Qual Manag. 1995;6(4):399–412.
42. Tavernier A. Advantages of BLUP animal model for breeding value estimation in horses. Livest Prod Sci. 1988;20(2):149–60.
43. Piepho HP, Möhring J, Melchinger AE, Büchse A. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica. 2008 May 1;161(1):209–28.
44. Martínez-Ávila JC, Guillén-Ponce C, Earl J, García-Cortés LA. Hereditary Lifetime Cancer Risk Assessment Modeling: A Case Study in Breast Cancer. Int J Mol Genet Gene Ther. 2016;1(2).
45. Hoadley B. The quality measurement plan (QMP). Bell Syst Tech J. 1981;60(2):215–73.
46. McCulloch CE, Neuhaus JM. Generalized linear mixed models. Wiley StatsRef Stat Ref Online. 2014;
47. Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;423–47.
48. Quan H, Li B, Saunders LD, Parsons GA, Nilsson CI, Alibhai A, et al. Assessing Validity of ICD-9-CM and ICD-10 Administrative Data in Recording Clinical Conditions in a Unique Dually Coded Database. Health Serv Res. 43(4):1424–41.
49. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring Diagnoses: ICD Code Accuracy. Health Serv Res. 40(5p2):1620–39.

Pseudo-data generation for the extraction of Problems, Treatments and Tests

Jeff Smith, M.S.[1], Evan French, William Cramer, MBA, M.S., PMP[1], Özlem Uzuner, PhD[2], and Bridget T. McInnes, Ph.D.[1]
[1]Virginia Commonwealth University, Richmond, VA, USA
[2]George Mason University, Fairfax, VA, USA

Abstract

One of the primary challenges for clinical Named Entity Recognition (NER) is the availability of annotated training data. Technical and legal hurdles prevent the creation and release of corpora related to electronic health records (EHRs). In this work, we look at the impact of pseudo-data generation on clinical NER using gazetteering utilizing a neural network model. We report that gazetteers can result in the inclusion of proper terms with the exclusion of determiners and pronouns in preceding and middle positions. Gazetteers that had higher numbers of terms inclusive to the original dataset had a higher impact.

1 Introduction

Named Entity Recognition (NER) is a subtask of Natural Language Processing (NLP) that identifies entities fitting a predetermined category from text. NER typically sits in the middle of NLP pipelines as the information extracted is sent to downstream tasks such as question answering[1], word sense disambiguation[2], and automatic text summarization[3].

Classical NER sought to identify entities from categories such as person, place, and location. Many early datasets utilized news sources as their primary corpora[456]. Traditional machine learning techniques for identifying sequences, such as conditional random fields (CRF) and hidden markov models (HMM) were used as classifiers. Modern NER now compromises of many domains, are domain specific, and utilize neural network architectures with contextual word or character embeddings to label sequences of entities[78]. In this study, we examine clinical NER focusing on the extraction of medical entities from the unstructured narratives in electronic health records (EHRs). The proper identification of these entities is critical for downstream tasks that rely on them such as developing a clinical knowledge extraction system[9]. Unfortunately, NER models have shown that they do not generalize well and require domain specific adaptation[10]. In 2001, Poibeau and Kosseim[11] were one of the first groups to document the issues of performing NER across different domains. Changes in the structure of text, the vocabulary used, and the types of entities being extracted require differing approaches which do not easily carry over between one another. Clinical NER is no exception to this. EHRs tend to be heavily unstructured, absent of complete sentences, and make heavy use of domain specific abbreviations and jargon.

One of the primary challenges in any supervised machine learning task is training data. Large amounts of annotated training data are required to be efficient and generalize the problem space well. However, these datasets have to be annotated by hand and require significant man-hours to assemble and normalize between annotators[12].A potential method of bypassing some of the challenges of generating datasets is the use of pseudo-data, also known as artificial training data. Pseudo-data has been used historically to generate additional samples in cases where there is a large class imbalance. There are two main types of pseudo-data, synthetic and sampled. Synthetic pseudo-data is data that has been produced through statistical or rule based methods that attempts to mimic samples already available in the dataset. A popular algorithm for generating synthetic data is SMOTE, or synthetic minority over-sampling technique[13], which uses a K-nearest neighbors algorithm to create new samples between existing ones. Sampled pseudo-data is data that is extracted from corpora or datasets and is labeled through semi-supervised[14] or unsupervised[15] methods. In this work, we look at the impact of pseudo-data generation on clinical NER using gazetteering utilizing a neural network model. We examine precision, recall, and F_1 score on a variety of gazetteers using the MIMIC-III Clinical Care Database as a pseudo-data source.

2 Data

In this section, we describe the corpora and gazeteers used in this study.

Corpora

2010 i2b2 Clinical Dataset. The 2010 i2b2 Clinical Dataset[16] is an annotated set of clinical data comprised of discharge summaries from three healthcare systems, Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center. The dataset was annotated for concept extraction, assertion classification and relation classification. The dataset contains annotations for three concepts, problems, tests, and treatments. Table 1 lists the exact details for the dataset. In particular, the training set contains 170 annotated documents with 16,525 annotations over 149,541 tokens; and the test data set contains 256 annotated documents with 31,161 annotations over 267,249 tokens.

Table 1: 2010 i2b2 Clinical Dataset Information

i2b2	Text			Annotations		
Set	**Files**	**Sentences**	**Tokens**	**Problem**	**Test**	**Treatment**
Train	170	16315	149541	7073	4608	4844
Test	256	27625	267249	12592	9225	9344

MIMIC-III. The MIMIC Critical Care Database[17] (MIMIC-III) is a freely available de-identified dataset comprising of electronic health records (EHRs) for over 40,000 critical care patients from the Beth Israel Deaconess Medical Center. It was collected between 2001 and 2012 and includes documents ranging from caregiver notes to imaging reports and prescribed medications. In particular, the caregiver notes section is comprised of 2,083,180 individual documents from a variety of internal departments with a total token count of 487,639,049. The caregiver discharge notes were used by this study as the primary data source for pseudo-data generation.

Gazetteers

Table 2 the number of annotations in the gazetteers for the three entities used for pseudo-data generation.

Table 2: Gazetteer Annotation Information

Source	ICD10CM	ICD10PCS	CPT	WebMED	FDA Drug List	Southern Cross
Problem	86,168	-	-	-	-	-
Test	-	303	4,813	17,676	-	-
Treatment	-	59	8,571	-	42,639	2,599

ICD10. The International Statistical Classification of Diseases and Related Health Problems (ICD) is a medical classification list generated by the World Health Organization that standardizes coding for medical terms across the globe. In the United States, ICD was modified into ICD-CM and ICD-PCS. ICD-CM is the clinical modification of ICD for classifying diagnosis and reasons for hospital visits. It is split into 21 chapters of differing categories. There were 5,562 entries in the system that were used as a source for problem annotations. ICD-PCS is the modification of ICD for standardizing procedure coding systems and is mainly used for billing and reporting. All of the tests and treatments used within the United States are codified in this database. There were 1,065 tests and 9,165 treatments coded in the database that were used for annotation sources.

CPT. Current Procedural Terminology (CPT) is another medical code set used in the United States and maintained by the American Medical Association. It contains information for surgical and diagnostic procedures and is released annually. It is mainly used by clinicians to itemize treatments. There were 1,346 tests and 3,958 treatments coded in the database that were used as annotation sources for this study.

FDA Approved Drug List. The Food and Drug Administration (FDA) Approved Drug List is a current listing of all drugs that have been given approval in the United States. Drugs are listed by both their brand name and generic chemical name along with application method and distributing company. A list of current drugs was downloaded and used as a source of potential treatment annotations for our system. A total of 8,906 treatments were present in the list.

WebMD Medical Tests. WebMD is an American company that provides news and information relating to medical care and diagnosis and was founded in 1996. The company maintains the WebMD website and features glossaries for

medical terms. One glossary is a list of medical tests and information about them. The list of terms from the glossary[18] was extracted and contained 625 tests.

Southern Cross Surgery List. Southern Cross is a healthcare system based out of New Zealand that maintains their own coding database for all surgeries performed at their locations. This list of surgeries was published online in 2014[19] and contains 43 pages of procedure codes along with text representations for each which totaled 761 treatments. This list was extracted and used as a source of treatment annotations for our system.

3 Method

We downloaded the MIMIC-III database, extracted discharge summaries, and ran them through pre-processing. All gazetteer sources were downloaded and also pre-processed. Gazetteer annotations were matched to text in the MIMIC-III discharge summaries and set aside as annotation files. A visual guide to the pipeline can be seen in Figure 1.

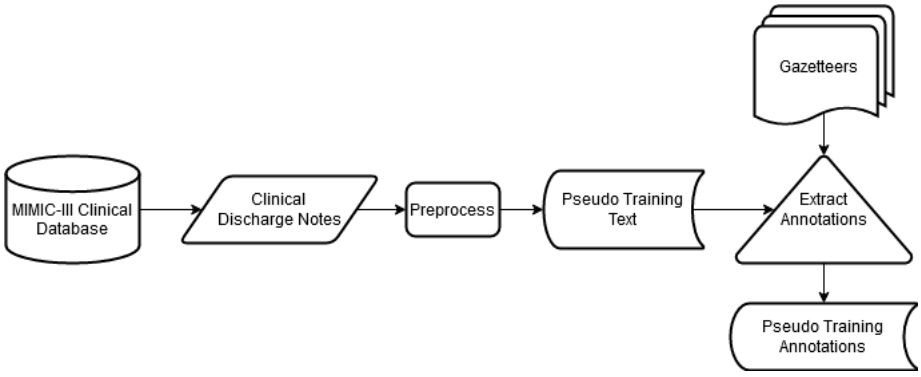


Figure 1: Visual Pipeline Representing Extraction of Pseudo Annotations from the MIMIC-III corpus.

Preprocessing. Text for discharge summaries was extracted from the 'NoteEvents' table where 'Category' was set to 'Discharge summary'. Pre-processing started with an initial step of combining all de-identified terms into single terms that could be easily turned into features, including numbers, dates, and times. Punctuation was then modified to match the format that the i2b2 dataset was in and sentences less than 8 tokens were removed. 400,000 sentences were then randomly selected. Gazetteer sources underwent manual review to extract terms and remove unneeded words for better matching. Annotations from each gazetteer were then scanned for in the extracted text and annotation files were generated. For multi-class experiments, annotation files from different sources were merged. In cases where there were overlapping annotations of different classes, each annotation were removed.

Model. In this study, we use a Bi-directional Long Short-Term Memory (BiLSTM) Recurrent Neural Network. The feature vectors are arranged in order by token and fed into the network along with the sentence length. Each sentence is read by two separate LSTM cells that read in alternate directions. The output is concatenated and fed into a linear layer with dropout. The data is then classified using a conditional random field (CRF). In this study, the training occurred for 15 epochs in each experiment unless otherwise stated. This number was chosen by training the network over a large amount of epochs numerous times and selecting the point where additional training produced negligible results. The BiLSTM+CRF architecture was pre-trained on the MIMIC-III annotations for n epochs; and then fine-tuned over the i2b2 training data for the remaining $15 - n$ epochs. The pre-training mimics the method used by Giorgi and Bader when pre-training with a noisy dataset[20]. Training utilized batch back propagation with an Adam optimizer at a learning rate of 0.005. Every model was initialized and trained independently.

Feature Representation. Instances were preprocessed to remove symbols and convert numbers, dates, and times into standardized values. We ran each instance through the SpaCy POS/DEP parser to get POS and DEP tags for each token and mapped them to locations in a one-hot vector. We then ran each token through a Word2Vec model that was trained on the MIMIC-III clinical notes database. The output of this model represented the first 200 values of the vector. The POS/DEP tags were appended and represented the rest of the vector for each token. The final shape of the resulting matrix was [sentences, tokens, features]. In this study, our features vector parameters consist of: Word2Vec

embeddings (length 200), is a number, date, or time, part of speech tags (length 19), and dependency tags (length 52).

Evaluation. We evaluate our method using precision, recall and F_1 score. The precision is a measure of how accurate each prediction was; recall is a measure of how many of the total tokens were predicted; and F_1 is the harmonic mean between the two. We evaluate our system at both the token and entity level. For the entity level evaluation we use two methods: strict and lenient. Strict requires a true positive to have exact span and tag matches. Lenient requires a partial match of the span with the correct tag. If two or more tags match a phrase, it will only be counted as true positive once, and the additional phrase will not be counted as a false.

4 Results

In this section, we describe the experiments we conducted on gazetteer annotating and the reasoning behind them. The experiments are split into: 1) single-class experiments and 2) multi-class experiments, where multi-class attempts to label all entities within a single model. The first set of experiments we performed take each individual gazetteer source and train a model with an inverse epoch structure. That is, a set number of epochs is defined and the gazetteer source is trained on for a given number of epochs. Then for the rest of the epochs, the model is trained on the i2b2 training data. This provides a way to determine what kind of effect each gazetteer is having on the model in an isolated environment. The next experiment takes every gazetteer source and trains them together on a multi-class model. This allowed us to determine if there were cumulative effects from including all of the gazetteers. Finally, we conducted two experiments only taking the best performing gazetteers. The first followed the same pattern as the previous experiments. The second kept the number of epochs for the i2b2 training static and changed the number of epochs we pre-training was conducted on the pseudodata.

Single class Gazetteer Pretraining

In this section, we present the results of our inverse epoch structure experiments over each individual gazetteer source. This evaluation provides a way to determine the effect each gazetteer is having on the model in an isolated environment. We also discuss the crossover results for each of the gazetteers shown in Table 3. The table shows the number of entities that were identified in MIMIC-III and the i2b2 training and test data for each of the gazetteers.

Table 3: Gazetteer Annotation Term Crossover Analysis

Class	Gazetteer	# Types	# Types in Mimic (%)	# Types Matching i2b2 Train (%)	# Types Matching i2b2 Test (%)
Problem	ICD10CM	4349	1872 (4.30E-01)	620 (1.17E-01)	678 (9.44E-02)
Test	CPT	1154	127 (1.10E-01)	64 (1.21E-02)	67 (9.33E-03)
	ICD10PCS	467	31 (6.64E-02)	12 (2.27E-03)	12 (1.67E-03)
	WebMD	675	345 (5.11E-01)	136 (2.57E-02)	141 (1.96E-02)
Treatment	CPT	1385	98 (7.08E-02)	50 (9.46E-03)	50 (6.96E-03)
	ICD10PCS	790	29 (3.67E-02)	21 (3.98E-03)	23 (3.20E-03)
	FDA	7230	2213 (3.06E-01)	338 (6.40E-02)	405 (5.64E-02)
	Southern	860	199 (2.31E-01)	103 (1.95E-02)	111 (1.55E-02)

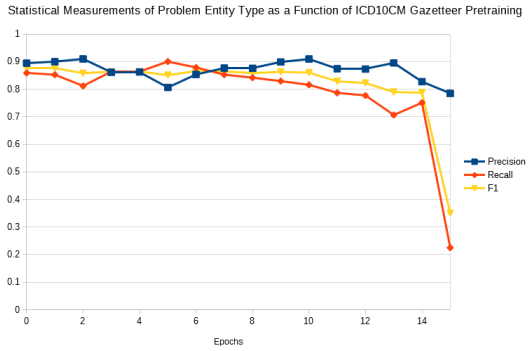


Figure 2: Statistical Measurements of Problem Entity Type as a Function of ICD10CM Gazetteer Pretraining

Problem Entity. Figure 2 shows the token level precision, recall, and F_1 scores of the problem entity type when we pretrained a single-class model on annotations from ICD10CM. The precision of the model rises above baseline when

trained 1 or 2 epochs then declines until 10 epochs where it remains roughly the same before finally diving 10 points. The recall follows an inverse pattern where it initially declines, then by 5 epochs goes above the baseline. After six epochs, the network has a continuous decline before the final dropoff. When compared to the multi-class model, the precision and recall exhibit the same general pattern but with a larger range in numbers. Out of the 4349 annotations available in ICD10CM, only 620 and 678 of them were found in the i2b2 test and train datasets respectively. This represents the highest percentage found in any of the gazetteers.

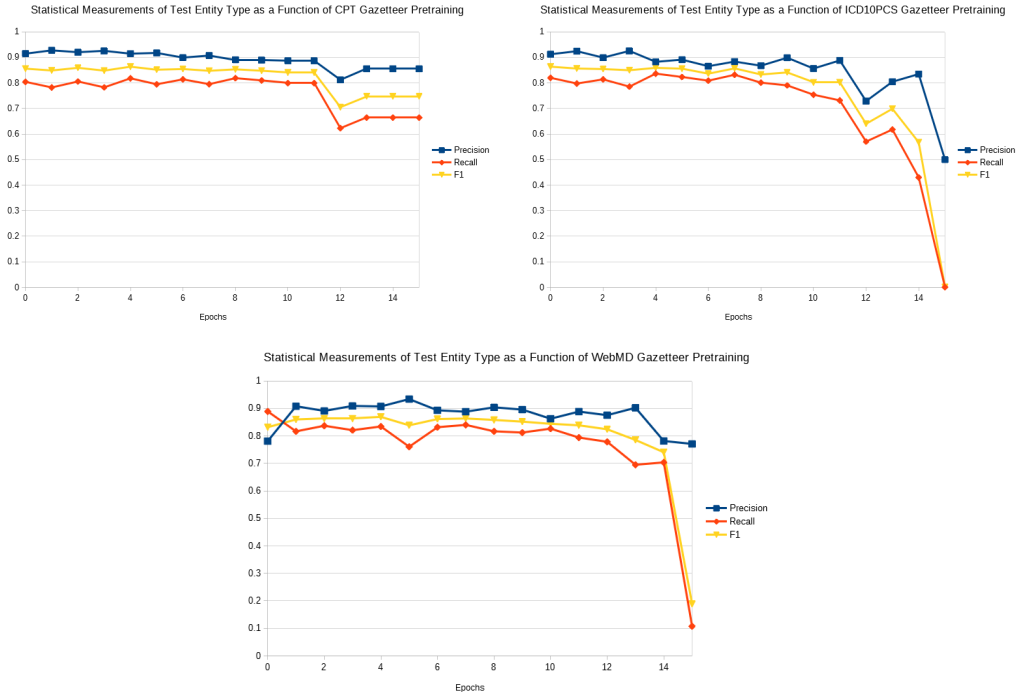


Figure 3: Statistical Measurements of Test Entity Type

Test Entity. Figure 3 shows the token level precision, recall, and F_1 scores of the test entity type when we pretrained a single-class model on annotations from CPT, ICD10PCS and WebMD gazeteers. For the CPT results, the precision and recall oscillate below and above baseline as the number of trained epochs increases. Out of the 1154 annotations available in CPT for the test annotation group, only 64 and 67 were found in the i2b2 test and train datasets respectively. For the ICD10PCS results, precision and recall also oscillate around baseline. Between 4 and 7 epochs we get an increase in recall at the cost of precision. Out of the 467 annotations available in ICD10PCS for the test annotation group, only 12 were found in both the i2b2 test and train datasets. The WebMD results show that the initial inclusion of the annotations from one epoch and onward result in a fairly large (10%) increase in precision. The recall inverses and falls, though not enough to offset the precision gains. Out of the 675 annotations available in WebMD, only 136 and 141 were found in the i2b2 test and train datasets respectively.

Treatment Entity. Figure 4 show the token level precision, recall, and F_1 scores of the treatment entity types. For the CPT results, the precision and recall oscillate slightly around baseline before suffering a steep dropoff. Out of the 1385 annotations available in CPT for the treatment annotation group, only 50 were found in both the i2b2 test and train datasets. For the ICD10PCS results the precision and recall both hover slightly below baseline when trained for a few epochs. Like previous experiments, the recall tanks at the end. Out of the 790 annotations available in ICD10PCS for the treatment annotation group, only 21 and 23 were found in the i2b2 test and train datasets respectively. For the FDA Drug List results the inclusion of the FDA Drug List has a large impact (>10%) on recall when included in a single epoch and onward until past 10 epochs. There is a small hit to precision (2%) initially and a larger hit (5-7%) from 3 epochs to 12. Out of the 7230 annotations available in the FDA Drug List, only 338 and 405 were found in the i2b2 test and train datasets respectively. For the Southern Cross Surgery List results when they are included for a

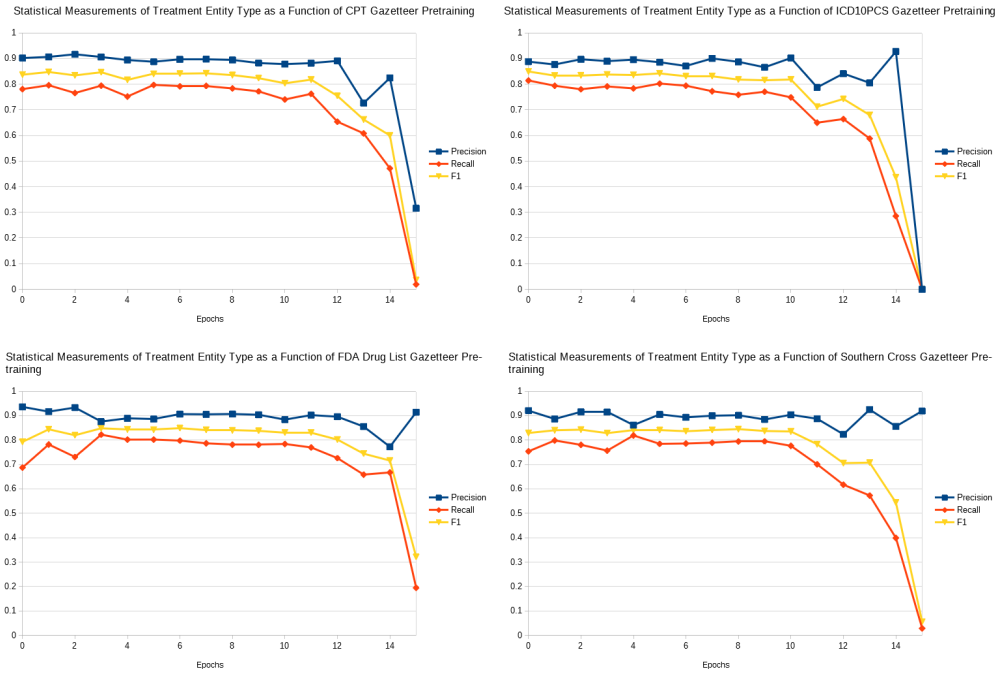


Figure 4: Statistical Measurements of Treatment Entity Type

single epoch or four or more, precision gains a moderate amount (5%) while recall declines a similar amount. This pattern holds until 10 epochs of pretraining where the recall dives towards 0. Out of the 800 annotations available in the Southern Cross Surgery List, only 103 and 111 were found in the i2b2 test and train datasets respectively.

Multi-Class Gazetteer Pretraining In this section, we evaluate our model by combining the gazetteer sources and training them in a multi-class environment. This allows us to determine the cumulative effects of including all sources.

Comprehensive Multi-Class Gazetteer Pretraining. Figure 5 shows the token level precision, recall, and F_1 scores of the problem, treatment, and test labels respectively when we merged the annotations from all the gazetteers and trained them on a multi-class model. The experiment was run three times and the results averaged. With the problem and treatment entity types, pretraining for one to five epochs results in a small (2%) increase in recall with a nearly equivalent loss in precision. Training for longer periods of time results in a steep decrease in recall which also results in a steep decline in F_1 score. Precision remains fairly high. The test entity type loses recall and gains precision at the same level as the previous entity types when we incorporate the gazetteer pretraining.

Static Multi-Class Gazetteer Pretraining. Figure 6 shows the token level precision, recall, and F_1 scores of the problem, treatment, and test labels respectively when we merged the annotations from only the best the gazetteers and trained them on a multi-class model. The specific gazetteers we used were ICD10CM (Test), WebMD Test List, FDA Drug List, and the Southern Cross Surgery List. For problem and test, the precision rises for several epochs when including the gazetteer data before eventually dropping off. The baseline treatment precision remains the highest but doesn't drop off drastically. The recall for the problem and treatment types also rise a couple of points whereas the recall for the test type suffers. In all three instances, the highest F_1 score is observed on the 5/10 ratio.

Selective Static Multi-Class Gazetteer Pretraining. Figure 7 shows the token level precision, recall, and F_1 scores of the problem, treatment, and test labels respectively when we merged the annotations from only the best the gazetteers and trained them on a multi-class model with a set number of epochs on the i2b2 dataset. The specific gazetteers we used were the same as in the selective multi-class experiment. For the problem type, precision increases all the way to the max number of epochs on the gazetteer data. Recall stops increasing after the 6th epoch and drops off rapidly afterwards. In most of the chart, an inverse pattern between recall and precision can be observed. For the test type,

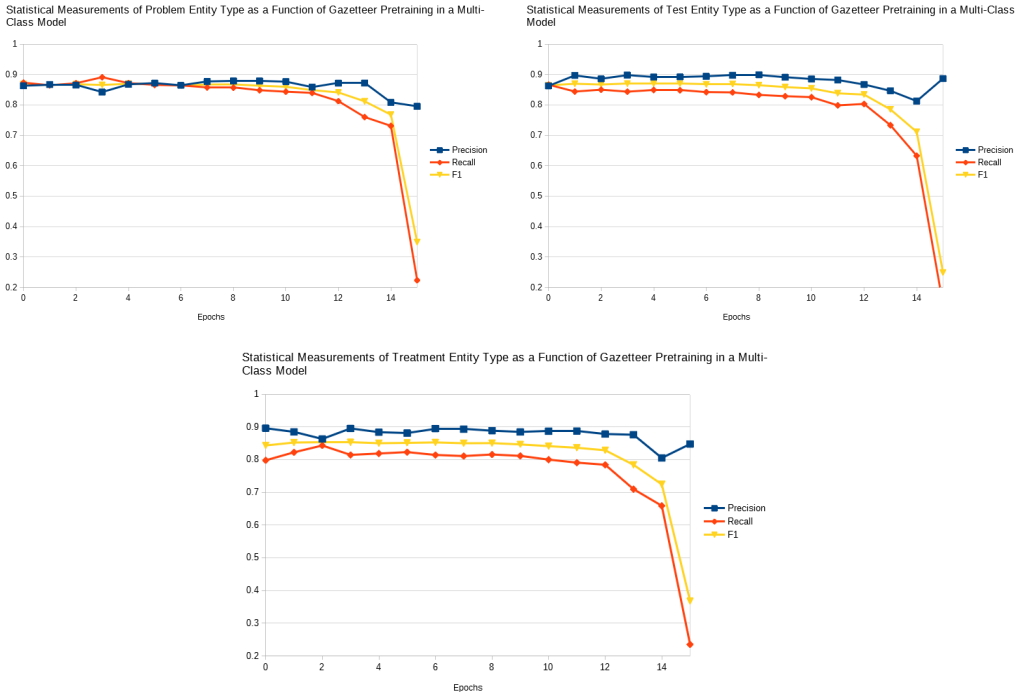


Figure 5: Statistical Measurements of Problem (Top), Test (Middle) and Treatment (Bottom) Entity Types as a Function of Gazetteer Pretraining in a Comprehensive Multi-Class Model

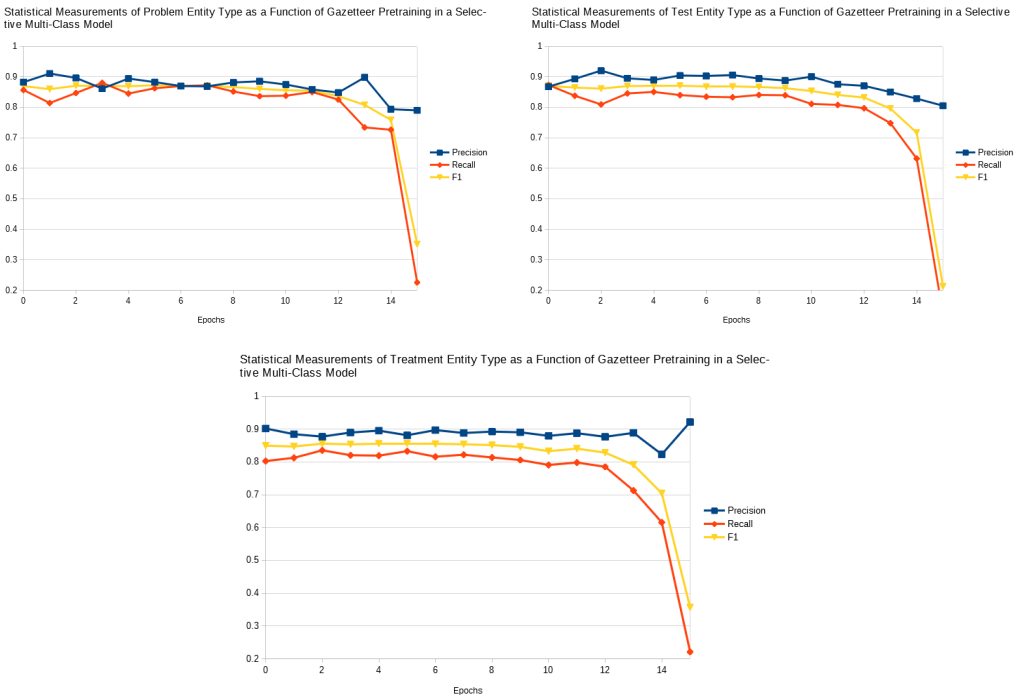


Figure 6: Statistical Measurements of Problem (Top), Test (Middle) and Treatment (Bottom) Entity Types as a Function of Gazetteer Pretraining in a Selective Multi-Class Model

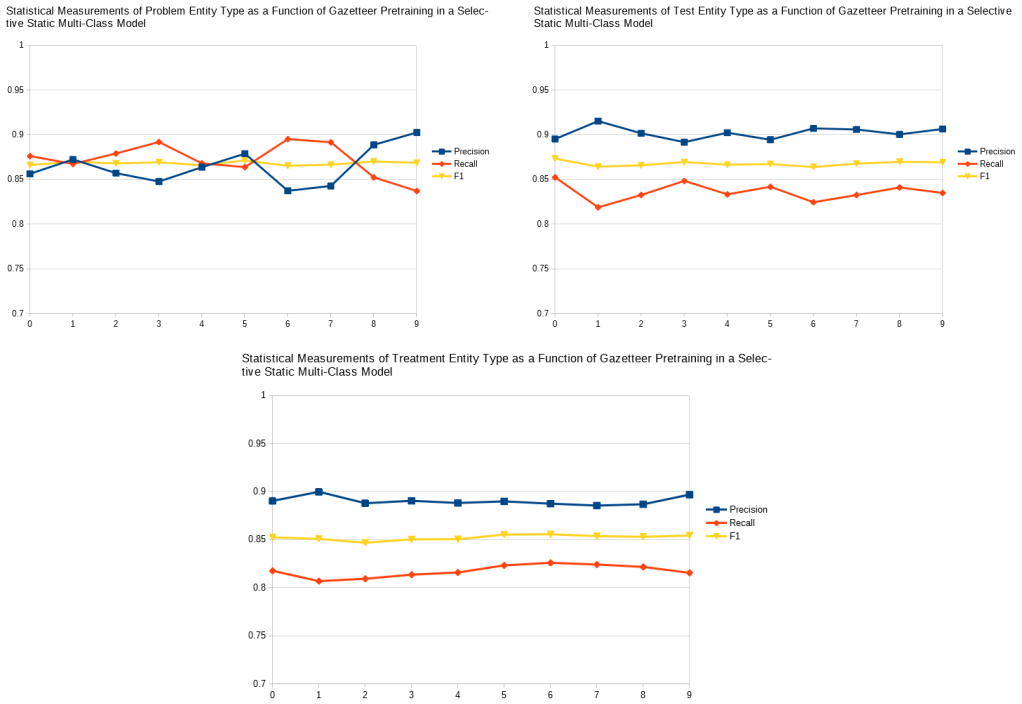


Figure 7: Statistical Measurements of Problem (Top), Test (Middle) and Treatment (Bottom) Entity Types as a Function of Gazetteer Pretraining in a Selective Static Multi-Class Model

precision increases for one epoch after the baseline but never again. Recall never increases. The test type benefits most from no gazetteer data. For the treatment type, precision increases slightly on the first epoch then hovers around baseline. Recall increases towards the latter epochs and falls back down at max epochs. The highest F_1 scores were observed around 5/6 epochs except for test.

Overall Multi-Class Gazetteer Results.

Table 4 shows the entity level strict and lenient precision, recall and F_1 scores for our multi-class baseline and selective epoch results. The results show that the F_1 scores, problem and treatment were able to exceed the baseline however the results show that test was not.

Table 4: Gazetteer Annotation Multi-Class Phrase Strict (Lenient) F_1 Measurement

Source	Class	Precision	Recall	F_1
Baseline	Problem	0.696(**0.919**)	0.631(0.815)	0.662(0.864)
(Epoch 15)	Test	**0.793**(**0.902**)	**0.762**(0.862)	**0.777**(**0.881**)
	Treatment	0.706(0.882)	0.681(**0.834**)	0.693(0.857)
Selective	Problem	**0.718**(0.91)	0.655(0.821)	**0.685**(**0.864**)
(Epoch 5)	Test	0.787(0.902)	0.754(0.86)	0.77(0.88)
	Treatment	**0.774**(**0.913**)	0.705(0.824)	**0.738**(**0.866**)
Selective Static	Problem	0.702(0.899)	**0.657**(**0.831**)	0.679(0.863)
(Epoch 5)	Test	0.742(0.885)	0.739(**0.868**)	0.74(0.877)
	Treatment	0.764(0.904)	**0.709**(0.831)	0.735(0.866)

5 Discussion and Limitations

Throughout all of the gazetteering experiments, a common trend we observed was a trade-off between precision and recall. In most cases, the trade-off was high enough to increase the F_1 score. The problem and treatment entities were improved over baseline.

Precision and recall being reciprocal of each other is not a new phenomenon with respect to machine learning. In the context of our experiments, one potential reason for the trade-off is the use of determiners and pronouns. In the i2b2 dataset, many of the annotations include determiners and pronouns in the beginning or middle. Some examples are: "an outpatient holter monitor", "his chest x-ray", and "a few fine crackles at the left base". With the annotations generated by gazetteering through MIMIC-III, these pronouns and determiners are not selected. This could result in a case where training too much on the gazetteering data results in the inclusion of proper terms with the exclusion of pronouns and determiners.

Some gazetteers outperformed others by several points. To try and understand the reason why, we analyzed the number of terms that overlapped between the gazetteers and the datasets (Table 3). What we found is that gazetteers that had a higher number of terms in the datasets had a higher impact. This is a logical conclusion, as more relevant data will results in better outcomes. What was surprising is the amount of crossover some gazetteers had. ICD10PCS and CPT codes had a large number of unique terms but very little crossover. This indicates that the vocabulary used in billing codes does not match what is used by medical professionals. Public lists, such as the WebMD test list and the Southern Cross surgery list, had a high level of crossover. The FDA drug list also had a high level of crossover. This makes sense as drugs are a common treatment for almost any condition.

There were a couple of identified limitations. The first is that we found at the end of our work was the coverage of our Word2Vec model. Only 86% of the terms in the training and test annotations were found in the model. This limitation could provide an explanation for the upper cap that the models demonstrated. The second is the way that gazetteer annotations were generated compared to the original annotations. The training and test annotations contained determiners pronouns at preceding and mid positions in many cases whereas the gazetteer annotations did not.

6 Conclusions and Future Work

In this work, we explored using gazetteering as a pseudo-data generation technique to improve performance in a deep neural network architecture. We showed that using gazetteers, there is a trade-off between precision and recall depending on the entity type and gazetteer used. This allows for more finely tuned results depending on intent.

There are a number of follow-up studies that can be considered. To generate sentence structures similar to that of the training and test documentation, term swapping could be employed. This is where terms from gazetteer sources are swapped in place of original terms in training annotations and trained on. Generating new embedding models from different sources or merging multiple embeddings could also be considered for better coverage. Other sources for gazetteers could also be considered that better align with the language used by practitioners writing clinical notes. Regenerating the gazetteer annotations and including preceding and mid determiners and pronouns could potentially provide better results.

References

[1] Toral A, Noguera E, Llopis F, Munoz R. Improving question answering using named entity recognition. In: International Conference on Application of Natural Language to Information Systems. Springer; 2005. p. 181–191.

[2] Ciaramita M, Altun Y. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2006. p. 594–602.

[3] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Ohe K. Text2table: Medical text summarization system based on named entity recognition and modality identification. In: Proceedings of the BioNLP 2009 Workshop; 2009. p. 185–192.

[4] Sang EF, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050. 2003;.

[5] Grishman R, Sundheim B. Design of the MUC-6 evaluation. In: Proceedings of the 6th conference on Message understanding. Association for Computational Linguistics; 1995. p. 1–11.

[6] Chinchor N, Brown E, Ferro L, Robinson P. Named Entity Recognition Task Definition (version 1.4). Online at: http://www nist gov/speech/tests/ie-er/er_99/doc/ne99_taskdef_v1_4 pdf accessed. 1999;6.

[7] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:150801991. 2015;.

[8] Santos CNd, Guimaraes V. Boosting named entity recognition with neural character embeddings. arXiv preprint arXiv:150505008. 2015;.

[9] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 09;17(5):507–513. Available from: https://doi.org/10.1136/jamia.2009.001560.

[10] Marrero M, Urbano J, Sánchez-Cuadrado S, Morato J, Gómez-Berbís JM. Named entity recognition: fallacies, challenges and opportunities. Computer Standards & Interfaces. 2013;35(5):482–489.

[11] Poibeau T, Kosseim L. Proper name extraction from non-journalistic texts. In: Computational Linguistics in the Netherlands 2000. Brill Rodopi; 2001. p. 144–157.

[12] Friedman C, Johnson SB. Natural language and text processing in biomedicine. In: Biomedical Informatics. Springer; 2006. p. 312–343.

[13] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–357.

[14] Zhu X, Goldberg AB. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning. 2009;3(1):1–130.

[15] Zaït M, Messatfa H. A comparative study of clustering methods. Future Generation Computer Systems. 1997;13(2-3):149–159.

[16] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011;18(5):552–556.

[17] Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3:160035.

[18] WebMD Medical Tests A-Z - Find information on medical tests from A to Z. WebMD;. Available from: https://www.webmd.com/a-to-z-guides/tests.

[19] List of Surgical Procedures; 2014. Available from: https://www.southerncross.co.nz/Portals/0/Society/EFulfillment/Product/List_of_Surgical_Procedures.pdf.

[20] Giorgi JM, Bader GD. Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics. 2018;34(23):4087–4094.

Development of an Informatics Algorithm to Link Seasonal Infectious Diseases to Birth-Dependent Diseases Across Species: A Case Study with Osteosarcoma

Sarah Tadlock[1], Charles A Phillips, MD, MSHP[2], Margret L Casal, Dr.med.vet, PhD[3], Marc S Kraus[3], DVM, Anna R Gelzer, Dr.med.vet, PhD[3], Mary Regina Boland, MA, MPhil, PhD, FAMIA[1]

[1]Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania, Philadelphia, PA, USA; [2]Cancer Center, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [3]Department of Clinical Studies and Advanced Medicine, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA

Abstract

Many diseases have been linked with birth seasonality, and these fall into four main categories: mental, cardiovascular, respiratory and women's reproductive health conditions. Informatics methods are needed to uncover seasonally varying infectious diseases that may be responsible for the increased birth month-dependent disease risk observed. We have developed a method to link seasonal infectious disease data from the USA to birth month dependent disease data from humans and canines. We also include seasonal air pollution and climate data to determine the seasonal factors most likely involved in the response. We test our method with osteosarcoma, a rare bone cancer. We found the Lyme disease incidence was the most strongly correlated significant factor in explaining the birth month-osteosarcoma disease pattern ($R=0.418$, $p=2.80X10^{-23}$), and this was true across all populations observed: canines, pediatric, and adult populations.

1. Introduction

1.1 Importance of Birth Month on Disease Risk

Prior research has established that birth seasonality affects disease risk for diseases and conditions falling into four broad categories (1): mental, cardiovascular, respiratory and women's reproductive health conditions (2). The etiologies underlying the relationship between birth seasonality (and or month) and each disease vary depending on the type of disease. For example, exposure to air pollution during the early stages of development (first trimester in humans) has been implicated in increased risk of cardiovascular disease in humans (3) and a similar finding was found in dogs (4). Additionally, climate factors have been implicated in the birth month - disease relationship. For example, climate factors (heat and humidity) are responsible for an increase in the dust mite population have been implicated in increased risk of asthma during certain birth months validating early work in this area (1,5).

1.2 Seasonality of Infectious Diseases May Play a Role

Another seasonally varying factor that may play a role in birth month - disease relationships is infectious diseases. Infectious diseases fall into two main categories based on their underlying causes: bacterial and viral. Prior work had investigated the relationship between flu in Asia and the USA across six sites to uncover relationships between flu and birth season-disease relationships (3). However, the results of this work revealed that flus often differ across the globe, and therefore, the specific strains of a flu may be important in understanding its specific role in the etiology of birth month dependent diseases.

1.3 Literature Gap: Country-Specific Information on Infectious Diseases Needed

A major gap of our prior work and the literature is the lack of detailed exploration of the role of infectious diseases, both bacterial and viral, on birth month dependent diseases. To properly address this issue, we need to catalog all of the seasonal bacterial and viral infectious diseases and conditions that occur in the USA. This must be done regional (e.g., for USA and separately for Asia, and so forth) to properly capture the local seasonality of each infectious disease and virus in that particular region. Therefore, the first step of this informatics algorithm is to mine the publicly available data from the Centers for Disease Prevention and Control (CDC) to establish a set of seasonally varying infectious diseases and viruses. These will form our dataset of potential factors in birth month dependent diseases.

1.4 Case Study: Osteosarcoma

Osteosarcoma is a form of cancer that originates in bone. In the United States, approximately 800 new osteosarcoma cases are diagnosed in humans each year, about half of which occur in patients under twenty

years old (6). Osteosarcoma incidence is largely bimodal, with 10% of new cases occurring in patients aged 60 and older (7). In older adults, osteosarcoma is often linked to an existing, underlying condition (8), many of which are explicitly bone-related. Although rare, we found that osteosarcoma varied by birth month in both dogs and humans. The birth month-disease risk curve was the same in dogs and humans indicating that the culprit exposure was likely perinatal (i.e., at the time of birth). If the exposure were prenatal, we would have expected to see a shift in the birth month-disease curves between dogs and humans due to their different gestational lengths (2 vs. 9 months) (9).

1.5 Purpose of Study

The purpose of this study is to develop an informatics method to link diseases caused by seasonal viruses and bacteria within the United States to birth month-dependent diseases. Our method utilizes publicly available data and links prior methods that utilize seasonally varying pollution and climate factors. Our method finds the most significant seasonal factor associated with a birth month dependent disease whether it be a seasonal infection, pollutant, or climate factor (e.g., rainfall).

2. Materials and Methods

This process collected data from the Children's Hospital of Philadelphia (CHOP), Penn Medicine, and Penn Vet, health systems, and public data available from US government sources including the CDC, NOAA, and EPA. The collection and analysis methods are visualized in **Figure 1**.

2.1 Obtaining Seasonal Infection Data

2.1.1 National Data

We created a list of potentially seasonal viruses and bacteria to investigate. Shiga toxin-producing *Escherichia coli*, legionellosis, listeriosis, Lyme disease, pertussis, measles, mumps, rubella, salmonellosis, tetanus, and tuberculosis (named as listed in the CDC tables) were included. We obtained annual data categorized by month over a ten-year period from web-based CDC resources.

Data from 2009-2015 was collected from the CDC's Morbidity and Mortality Weekly Report (MMWR) Summary of Notifiable Infectious Diseases and Conditions — United States (10). Data from 2016-2018 was collected from the CDC's Wide-ranging Online Data for Epidemiologic Research (WONDER) Nationally Notifiable Infectious Diseases and Conditions, United States: Annual Tables (11). These resources presented the total number of reported cases for each month in a series of columns, with the last column as the total. Between the 'December' and 'Total' columns was a space for cases that were included in the total for the year, but not categorized within a month. However, this space remained blank, presuming that all counted cases in the year were reported within a specific month. The tabular, numeric values were compiled and saved in disease-specific tables for further analysis. The CDC also has a national outbreak database tool, the National Outbreak Reporting System (NORS) Dashboard (12). This tool provides insight into foodborne illnesses outbreaks including E. coli, listeriosis, and salmonellosis; allowing the user to sort data by year, setting, and vector. These data were helpful for identifying strong categories for further investigating case counts across certain food groups, although the aggregate case counts were more definitive than the stratified.

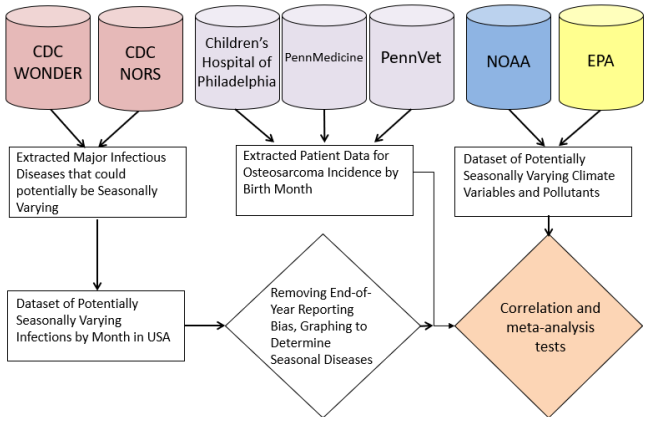


Figure 1. Overview of Our Method to Determine and Analyze Seasonal Diseases and Factors Related to Osteosarcoma Birth Season.

2.1.2 Regional Data

Lyme disease is prevalent in the northeast and northcentral parts of the United States (13). Because the osteosarcoma CHOP dataset comes from patients diagnosed in Pennsylvania, which has thousands of Lyme disease cases each year, state-level data were collected from the Pennsylvania Department of Health (14). Lyme disease seasonality data by month in Pennsylvania were only available from a 2017 report, unlike all other diseases analyzed, which had 10 years of data to average. County-level data from 2009-2018 in Pennsylvania were modified only by labelling case counts under their respective 5-digit Federal Information Processing Standard codes before analysis.

To investigate potential geographic similarities in Lyme disease and osteosarcoma distribution, osteosarcoma case counts were collected by state (15).

2.2 Obtaining Seasonal Pollution and Climate Data

Because all osteosarcoma cases were diagnosed in Philadelphia, Pennsylvania, county climate and pollutant data for the city were collected from NOAA (16). Six climate variables, measured by month, were factored into the analysis, including: mean sunshine hours, high and low temperatures, relative humidity, precipitation days, and precipitation inches. Five pollutant variables, measured daily and averaged by month, were included; SO2, CO, NO2, PM 2.5, and ozone Air Quality Index values (17). **Table 1** includes the data sources for all regional and national infection, climate and pollution data.

Table 1. Sources of All Exposure Variables

Exposure	Study Site	Collection Site	Collection Period	Source	Ref.
Infectious Disease					
E. coli	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
	Multi-state	State-based, reported to CDC	2009-2019	CDC	(18)
Legionellosis	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
Listeriosis	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
	Multi-state	State-based, reported to CDC	2009-2019	CDC	(19)
Lyme	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
	Pennsylvania	Pennsylvania, USA	2017	PA DOH	(14)
Measles	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
Mumps	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
Pertussis	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
Rubella	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
Salmonellosis	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
	Multi-state	State-based, reported to CDC	2009-2019	CDC	(20)
Tetanus	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
Tuberculosis	USA	State-based, reported to CDC	2009-2018	CDC Wonder	(11)
Climate					
All 6 Sunlight/Moisture Variables	Philadelphia County, Pennsylvania	Philadelphia County, Pennsylvania	1981-2010	NOAA	(16)
Pollutants					
All 5 Pollutant Variables	Philadelphia County, Pennsylvania	Philadelphia County, Pennsylvania	2000-2019	EPA	(17)

2.1.3 Assessing Infections' Potential to Affect Bone from Literature

Across the diseases studied, Lyme had the most literature related to bone conditions in various populations. A 1980 study of a man "whose disease appears to be tick‐transmitted" reported joint destruction like that seen in rheumatoid arthritis (21). A Dutch study found "subluxation of the toe joint and periostitis of the bones of the lower limb" (22), while a study in rodents found that Lyme disease "induces trabecular bone loss" (23). In children, orthopedic Lyme can present itself as swelling and pain in the large joints, like in the knees, and potential abnormalities like osteopenia and arthritis, though the latter two symptoms are not unique to pediatric Lyme (24). Lyme disease in dogs can present itself with joint swelling and lameness (25).

2.2 Clinical Data for Osteosarcoma

We obtained dog and human data broken down by biological sex due to the importance of hormonal growth patterns on bone growth. In addition, we separated pediatric human patients from adult human data due to the difference based on age for bone growth patterns that have been implicated in osteosarcoma.

2.2.1 Dog Data

We obtained data from Penn Vet on pet dogs that were diagnosed with osteosarcoma. We obtained this data from a cleaned clinical data repository specifically designed for research purposes (26). Our prior research

also found that certain dog breeds were associated with greater risk of osteosarcoma with the top five dog breeds being the Anatolian Shepherd Dog, Greyhound, Irish Wolfhound, Saint Bernard and Bullmastiff (26). In addition, there is known to be an association between osteosarcoma and being of female sex (as a biological variable). Therefore, we separated our data based on biological sex, which resulted in a set of 286 male dogs with osteosarcoma and 292 female dogs with osteosarcoma. We compared the birth month distribution in the male and female dogs with osteosarcoma against the male and female dogs in a larger cohort treated between 2000 and 2017 (26). The majority of dogs both with and without osteosarcoma were spayed/neutered.

2.2.2 Human Data

We obtained data on pediatric osteosarcoma from CHOP. There were 131 children treated at CHOP with osteosarcoma that were from the general Philadelphia area. These osteosarcoma patients were compared against other patients treated at CHOP for other pediatric tumors (i.e., non-osteosarcoma tumors). We also obtained adult human data from females (recorded as biological females at PennMedicine) treated between 2010 and 2017. We compared 402 adult females with osteosarcoma against our entire female-only adult cohort of 771,954 patients.

2.3 Statistical Analysis

2.3.1 Determining Seasonality of Infections

Despite all reported cases being sorted under a specific month, there were consistent increases in cases during December, seemingly contrary to expected counts from November and January data. This may be a result of individual states receiving and reporting an influx of cases at the end of the calendar year and reporting them all to the CDC as occurring during December (11). To normalize the effects of potential over-estimation in December, a hypothetical December value was created by taking the average of January and November cases. The hypothetical December considers that December lies between January and November and assumes the number of cases would do the same. We created overlaid scatterplots of average case counts by month with lines in one chart and an average case count by month with standard deviations for each disease to conduct an initial visual assessment of seasonality, in **Figure 2**. Pennsylvania Lyme only has the one scatterplot, because only data from 2017 was available.

2.3.2 Stratification by Sex

The adult human and dog data were stratified by sex to examine the connection between hormonal regulation and bone growth. In adolescents, osteosarcoma often occurs around the time of a growth spurt, when osteoblasts are rapidly forming to make new bone. Growth spurts are caused by hormonal changes in adolescents, which concurrently spur sexual maturation (27). While there are no definitive causes of osteosarcoma, genetics may play a role (28), and incidence in adolescents, particularly males, and post-menopausal women suggest that a hormonal distinction may affect osteosarcoma occurrence across sexes.

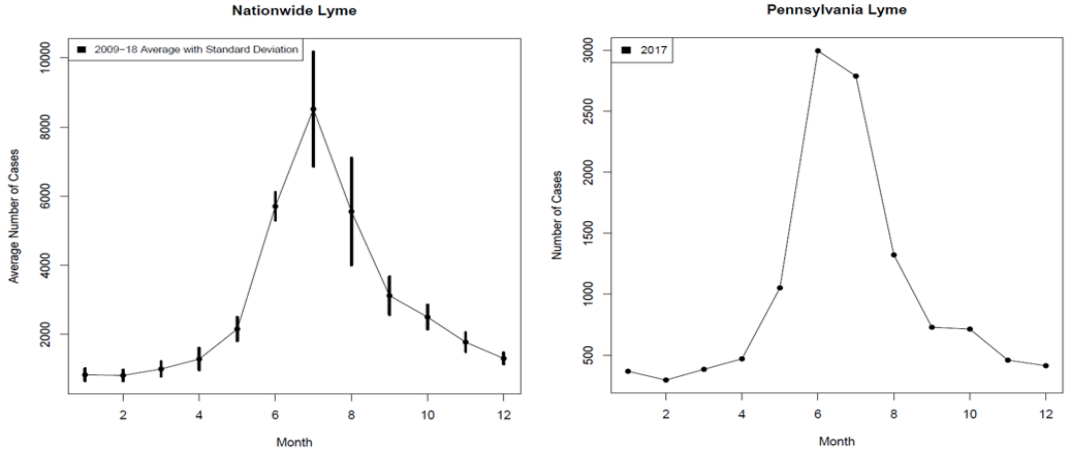


Figure 2. Curves for Nationwide and Pennsylvania Lyme Disease. Nationwide Data was Averaged from 2009-18, while Pennsylvania-specific Data was Exclusive to 2017.

2.3.3 Correlation of Osteosarcoma and Meta-Analysis of Osteosarcoma Incidence - Infections Across Multiple Populations

We then ran t-tests on all eleven disease datasets after their hypothetical December modification against the four osteosarcoma datasets. Observing the average trends of monthly disease cases, we pursued further

analysis on eight diseases, including: E. coli, legionellosis, aggregate listeriosis, Lyme, Lyme in Pennsylvania, rubella, salmonellosis, and tuberculosis.

With these eight groups, we used correlation testing to make plots of the disease's average monthly cases against each osteosarcoma dataset, and exported a matrix of the dataset size, correlation value, and p-values.

We used the metacor R package to generate forest plots with the DerSimonian-Laird function. There are plots for national data of each of the eight diseases that appeared to be seasonal, along with Lyme disease monthly trends in Pennsylvania in 2017. The forest plots included meta-analysis of all eight diseases with children, adult females, and sex-stratified dog osteosarcoma datasets.

We again used the metacor R package DerSimonian-Laird function to extract the summary values for each of the eight disease and osteosarcoma correlations and generated a Manhattan plot of the -log(p) * sign(R) to demonstrate the correlation of the disease with osteosarcoma incidence. These processes were repeated with the osteosarcoma correlations with climate and pollutant variables, respectively, and then compiled into a total plot of factors and osteosarcoma correlations in **Figure 3.**

2.3.4 Determining Effects of Vaccination on Seasonally Varying Infections Across Species

Except for rubella, all other seasonal diseases investigated are bacterial infections. In the US, the only common human vaccine administered for any of the diseases was MMR, the combination vaccine preventing measles, mumps, and rubella (29). Among dogs, the only common vaccine is for canine Lyme disease. While dogs cannot get legionellosis, they are susceptible to leptospirosis, another disease frequently linked to bacteria living in outdoor bodies of water, or rodents in urban areas (30). Dogs are also not capable of contracting measles, but can get canine distemper, another illness within the measles family (31). **Table 4** includes detailed vaccine schedules for humans and dogs.

3. Results

3.1 Seasonality of Infectious Diseases in USA

Our method found that eight infectious diseases were seasonal while four were not seasonally varying. Some of these were found to be seasonal in the literature (**Table 2),** however, given the presence of vaccines for both dogs and humans, many previously seasonal diseases are no longer seasonal. Therefore, it is important to investigate the current seasonality of infectious bacterial and viral diseases and conditions. A couple of diseases that we investigated are not reported to infect dogs (**Table 2**). However, we included these diseases (e.g., Legionnaires) due to the presence of related infections that do affect dogs (e.g, Leptospirosis).

Table 2. Seasonality of Common Infectious Diseases

Infectious Disease	Seasonal in Current Data (2009-2018)?	Historically Seasonal (from literature)?	Affects Bone in Literature?	Affects Dogs	Similar Canine Virus/Bacteria	Ref.
E. coli	Yes	Yes	Yes	Yes		(32–34)
Legionellosis	Yes	Yes	Yes	No	Leptospirosis	(35–37)
Listeriosis	Yes	Yes	Yes	Yes		(38–40)
Lyme	Yes	Yes	Yes	Yes		(21,41,42)
Measles	No	Yes	Yes	No	Canine distemper	(31,43,44)
Mumps	No	Yes	No	Yes		(45,46)
Pertussis	No	No	No	No	*Bordetella bronchiseptica*	(47,48)
Rubella	Yes	Yes	Yes	No		(49,50)
Salmonellosis	Yes	Yes	Yes	Yes		(51–53)
Tetanus	No	Yes	Yes	Yes		(54–56)
Tuberculosis	Yes	Yes	Yes	Yes		(57–59)

3.2 Seasonal Infection, Climate and Pollutant Data

Importantly, many infectious diseases could be correlated to birth month dependent diseases not because the infection is related to the birth month dependent disease, but rather because of shared climate factors that may be underlying the relationship. To help address this issue, we also investigate the effects of climate and pollutant factors on the osteosarcoma - birth month relationship, in a manner similar to prior work (3). We next gathered government data on six climate variables and five pollutants in Philadelphia county in **Table 1**. These additional factors are important to consider because of the established disease

seasonality. The variation in months and seasons is a result of various climate and pollutant interactions which affect breeding conditions for bacteria or virus carriers like humidity.

Table 3. Results for Correlation with Osteosarcoma Birth Seasonality

Factor	R	Nominal P-value	Significant After Bonferroni Correction?
Seasonal Infection			
E. coli	0.189	0.004	No
Legionellosis	0.156	0.054	No
Listeriosis	0.069	0.191	No
Lyme	0.209	4.31×10^{-4}	Yes
PA Lyme	0.418	2.80×10^{-23}	Yes
Rubella	0.005	0.481	No
Salmonellosis	0.109	0.103	No
Tuberculosis	0.291	0.011	No
Seasonal Climate Factor			
Mean Sunshine	0.324	0.001	Yes
High Temp	0.241	0.002	Yes
Low Temp	0.233	8.96×10^{-4}	Yes
Relative Humidity	0.108	0.277	No
Precipitation Days	-0.005	0.473	No
Precipitation Inches	-0.008	0.441	No
Seasonal Pollutant Factor			
SO2	-0.121	0.128	No
CO	-0.226	0.041	No
NO2	-0.247	0.002	Yes
PM 2.5	0.107	0.108	No
Ozone	0.281	0.008	No

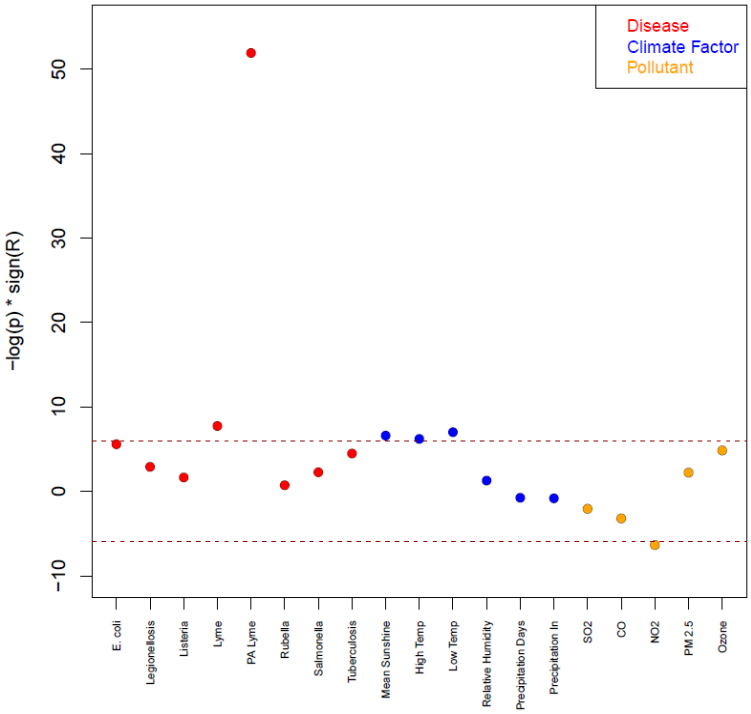


Figure 3. Overall Modified Manhattan Plot Showing All Seasonal Factors Relationship with Osteosarcoma Birth Season. The y-axis shows the log(p-value) * sign of the correlation. The positive correlations are on the top part and the negative correlations on the bottom of the figure. Red dashed line indicates the Bonferroni cutoff.

3.3 Meta-Analysis of Seasonal Factors and Osteosarcoma Risk Across Populations

Nineteen variables spanning seasonal disease, climate factors, and pollutants were assessed for correlation with osteosarcoma (results in **Table 3** and **Figure 3)**. We found six of the nineteen variables were significantly associated with Osteosarcoma after adjusting for multiple testing using Bonferroni correction (**Table 3)**. These six significant associations are strong candidates for understanding the relationship between Osteosarcoma and birth seasonality. These include national Lyme disease, Pennsylvania Lyme disease, Mean Sunshine hours, high and low temperature and NO2 exposure (shaded in grey in **Table 3)**. Pennsylvania Lyme disease had the strongest positive correlation and a breakdown of the association in each of the four datasets is shown in **Figure 4**. NO2 exposure was the only exposure that was anti-correlated with Osteosarcoma and was also the only pollutant that was significant.

3.4 Assessing the Effect of Vaccination on Infection - Osteosarcoma Risk Signal

Of the seasonal diseases, only human rubella and canine Lyme disease have widespread vaccine usage in the United States. An individual rubella vaccine was developed by 1969, and was soon combined into the Measles, Mumps, and Rubella (MMR) vaccine in 1971 (60). As of the 2017-2018 school year, school-aged children in Pennsylvania are required to have two doses of the MMR vaccine, barring medical or religious exemptions (61). In general, people aged 18 or older and born after 1956 needs one dose the MMR vaccine if they have not already had rubella (62). However, the bimodal age distribution of osteosarcoma suggests that an at-risk population aged 60 and older, whose age group comprises 10% of osteosarcoma cases, may not have had the MMR vaccine in any

dosage. The dataset of children with osteosarcoma are far more likely to have been vaccinated using two doses against rubella for school within their lifetimes, whereas the adult females may have been born before the vaccine was available and/or enforced. However, it is also important to note that within the last decade, there were fewer than ten reported rubella cases in the United States each year.

4. Discussion

4.1 Links Between NO2, Estradiol, and Cancer

Nitrogen dioxide (NO2) was the only finding that was significantly anti-correlated with Osteosarcoma (**Figure 3).** Importantly, nitric oxide synthase induced by 2-methoxyestradiol has led to cytotoxicity and apoptosis followed the generation of intracellular NO2 (63). NO2 is known to be toxic, and capable of causing cell and DNA damage, triggering cell death (63). Nitro-oxidative stress encourages NO2 to impact DNA (63). While NO2 has been linked to causing cell and DNA damage, it has also been found to kill metastatic osteosarcoma cells (63). As a known toxin, NO2 has the potential to kill healthy cells, encourage their mutation into potentially cancerous cells, and even eventually kill the cancerous cells themselves (63). Estradiol is a form of estrogen, whose levels peak and plateau within a females' reproductive years (64). A study found that 2-methoxyestradiol caused apoptosis in osteosarcoma cells (65). The increase in estradiol around puberty and decrease around menopause might indicate changes in NO2 production, and various cellular consequences, like DNA mutations. When estradiol levels are maintained throughout the reproductive years, osteosarcoma incidence also plateaued.

4.2 Links Between Lyme Disease, Bone, and Cancer

Lyme disease can have lasting effects on bone, including arthritis and reduced density. While only 50% of Lyme patients develop the characteristic rash, orthopedic symptoms may lead to a more conclusive diagnosis (24). Like osteosarcoma, Lyme disease often causes joint swelling and pain and knees are commonly affected for both conditions. Patients infected with Lyme disease (especially chronic infection) increases susceptibility to cellular changes due to a chronically weakened immune system, which leaves them vulnerable to developing illnesses (66). Importantly, our literature review found that two diseases (mumps and pertussis) did not affect bone in the literature (**Table 2**) and therefore, we would not expect these diseases to be involved in explaining the osteosarcoma birth seasonal relationship. Our method revealed that one of the two was not seasonal (pertussis) and therefore not included in our meta-analysis and the other (mumps) was not correlated with osteosarcoma (**Figure 3, Table 3).**

4.3 Our Method Optimized for Hypothesis Generation

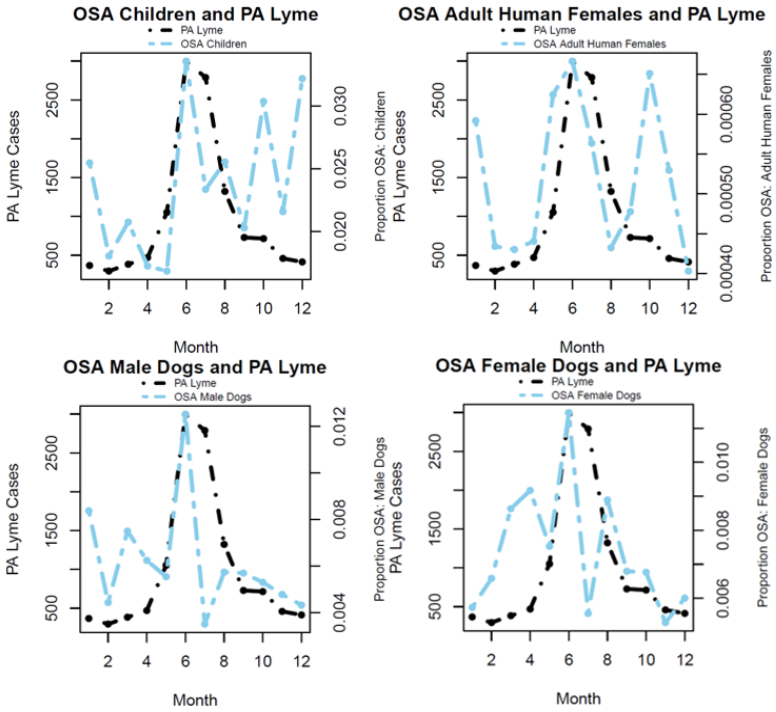


Figure 4. Pennsylvania Lyme and Osteosarcoma (OSA) Dataset Correlations.

The methods described here are useful in hypothesis generation, but not for investigating causality. Investigating relationships between osteosarcoma and various seasonal factors is a necessary step in identifying strong candidates for further analysis, including confounding variables, and interactions of multiple factors.

4.4 Limitations

Limitations of this study include that we do not know who among our population was vaccinated (either dogs or humans). We are also assuming that the national data on seasonal infections collected between 2009-2018 is constant throughout the entire time period. This is most correct for the pediatric patients, but for older women, they may have been exposed to a different seasonal infection pattern.

Generalizability of our method to other regions beyond the contiguous 48 states in the United States of America may be limited by prevalence of infectious diseases at those locations and availability of vaccination at a given location. For example, yellow fever vaccination is more common in states such as Hawaii, but is not broadly administered in the contiguous USA states (67).

Table 4. Vaccine Availability

Disease	Cause	Population / Year Vaccine Became Available	Schedule	Source	Ref.
E. coli	Bacteria	n/a	n/a	IntechOpen	(68)
Legionellosis	Bacteria	n/a	n/a	CDC	(69)
Listeriosis	Bacteria	n/a	n/a	Frontiers in Cellular and Infection Microbiology	(70)
Lyme	Bacteria	Dog Only / 2009	First dose: 8 weeks, Booster after 1 year, Boosters every 3 years	Merck Animal Health	(71)
PA Lyme	Bacteria	Dog Only / 2009	First dose: 8 weeks, Booster after 1 year, Boosters every 3 years	Merck Animal Health	(71)
Rubella	Virus	Human Only / 1971 (MMR)	First dose: 12-15 months, Second dose: 4-6 years	CDC, CHOP	(29) (60)
Salmonellosis	Bacteria	n/a*	n/a	Human Vaccines & Immunotherapeutics	(72)
Tuberculosis	Bacteria	n/a*	n/a	CDC	(73)

* Both salmonellosis and tuberculosis have vaccines in the US, although neither are commonly administered. The salmonella serotype typhi causes typhoid fever, and this is one serotype covered by a vaccine (72). Foodborne salmonella infection is often manageable, and not vaccinated against. BCG is a tuberculosis vaccine rarely administered in the US because of low infection counts (74). However, it is often given to children in other countries where tuberculosis risk is higher.

5. Conclusion

We developed an algorithm to harness publicly available data on infectious diseases to assess the seasonality of infectious diseases and to incorporate these seasonal patterns into a larger analysis framework that includes pollutant and climate factors that seasonally vary. This enables analysis of multiple factors to assess their role in birth seasonal relationships, one example being osteosarcoma. We found significant correlations with monthly national and Pennsylvania-based Lyme disease counts, Philadelphia mean sunshine hours, and a significant anti-correlation with Philadelphia NO2 air quality values. We adjusted all of our findings' significance using Bonferroni to address multiple testing. These results warrant further investigation to elucidate the underlying biological mechanisms. In addition, our method could be applied to other birth month dependent diseases that could have a seasonal infectious disease as a potential mechanism to explain the findings.

Acknowledgments: We thank the Perelman School of Medicine for generous funding to support this project.

References

1. Boland MR, Shahn Z, Madigan D, Hripcsak G, Tatonetti NP. Birth month affects lifetime disease risk: a phenome-wide method. J Am Med Informatics Assoc. 2015;22(5):1042–1053.
2. Boland MR, Fieder M, John LH, Rijnbeek PR, Huber S. Female Reproductive Performance and Maternal Birth Month: A Comprehensive Meta-Analysis Exploring Multiple Seasonal Mechanisms. Sci Rep. 2020;10:Article 555.
3. Boland MR, Parhi P, Li L, Miotto R, Carroll R, Iqbal U, et al. Uncovering exposures responsible for birth season – disease effects: a global study. ournal Am Med Informatics Assoc. 2017;25(3):275–288.
4. Boland MR, Kraus MS, Dziuk E, Gelzer AR. Cardiovascular Disease Risk Varies by Birth Month in Canines. Sci Rep. 2018;8:Article 7130.
5. KORSGAARD J, DAHL R. Sensitivity to house dust mite and grass pollen in adults: Influence of the month of birth. Clin Exp Allergy. 1983;
6. Johns Hopkins Medicine. Osteosarcoma [Internet]. Health. 2020. Available from: https://www.hopkinsmedicine.org/health/conditions-and-diseases/sarcoma/osteosarcoma
7. American Cancer Society. Key Statistics for Osteosarcoma [Internet]. 2020. Available from: https://www.cancer.org/cancer/osteosarcoma/about/key-statistics.html
8. American Cancer Society. Osteosarcoma Risk Factors [Internet]. 2018. Available from: https://www.cancer.org/cancer/osteosarcoma/causes-risks-prevention/risk-factors.html

9. Burke A. How Long Are Dogs Pregnant? [Internet]. 2016. Available from: https://www.akc.org/expert-advice/dog-breeding/how-long-are-dogs-pregnant/#:~:text=The normal gestation period in,is often hard to determine.
10. CDC. Summary of Notifiable Infectious Diseases and Conditions — United States. MMWR. 2015.
11. CDC. Nationally Notifiable Infectious Diseases and Conditions, United States: Annual Tables. WONDER. 2018.
12. CDC. NORS Dashboard [Internet]. National Outbreak Reporting System (NORS). 2018. Available from: https://wwwn.cdc.gov/norsdashboard/
13. CDC. Reported Cases of Lyme Disease — United States, 2018 [Internet]. Lyme Disease. 2018 [cited 2020 Jul 7]. Available from: https://www.cdc.gov/lyme/datasurveillance/maps-recent.html
14. Division of Infectious Disease Epidemiology. 2017 Lyme and Other Tickborne Diseases Surveillance Report. 2019.
15. CDC. United States and Puerto Rico Cancer Statistics, 1999-2016 Incidence Request [Internet]. WONDER. 2016 [cited 2020 Jul 6]. Available from: http://wonder.cdc.gov/cancer-v2016.html
16. NOAA. Climate data for Philadelphia [Internet]. 2020 [cited 2020 Jul 22]. Available from: https://en.wikipedia.org/wiki/Philadelphia#Climate
17. EPA. Air Data - Multiyear Tile Plot [Internet]. Outdoor Air Quality Data. 2020. Available from: https://www.epa.gov/outdoor-air-quality-data/air-data-multiyear-tile-plot
18. CDC. Reports of Selected E. coli Outbreak Investigations [Internet]. E. coli Outbreaks. 2020 [cited 2020 Jun 17]. Available from: https://www.cdc.gov/ecoli/outbreaks.html
19. CDC. Listeria Outbreaks [Internet]. Listeria Outbreaks. 2020 [cited 2020 Jun 17]. Available from: https://www.cdc.gov/listeria/outbreaks/index.html
20. CDC. Reports of Selected Salmonella Outbreak Investigations [Internet]. Salmonella Previous Outbreaks. 2020. Available from: https://www.cdc.gov/salmonella/outbreaks.html
21. Steere AC, Brinckerhoff CE, Miller DJ, Drinker H, Harris Jr. ED, Malawista SE. Elevated levels of collagenase and prostaglandin e2 from synovium associated with chronic lyme arthritis. Arthritis Rheumatol. 1980;23(5):591–9.
22. Houtman PM, Tazelaar DJ. Joint and bone involvement in Dutch patients with lyme borreliosis presenting with acrodermatitis chronica atrophicans. Neth J Med. 1999;54(1):5–9.
23. Tang TT, Zhang L, Bansal A, Grynpas M, Moriarty TJ. The Lyme Disease Pathogen Borrelia burgdorferi Infects Murine Bone and Induces Trabecular Bone Loss. Infect Immun. 2017;85(2).
24. Davidson RS. Orthopaedic complications of Lyme disease in children. Biomed Pharmacother. 1989;43(6):405–8.
25. American Veterinary Medical Association. Lyme disease: A pet owner's guide [Internet]. 2020. Available from: https://www.avma.org/resources/pet-owners/petcare/lyme-disease-pet-owners-guide
26. Boland MR, Casal ML, Kraus MS, Gelzer AR. Applied Veterinary Informatics: Development of a Semantic and Domain-Specific Method to Construct a Canine Data Repository. Sci Rep. 2019;Article 18641.
27. Center for Academic Research and Training in Anthropogeny. ADOLESCENT GROWTH SPURT [Internet]. 2020. Available from: https://carta.anthropogeny.org/moca/topics/adolescent-growth-spurt#:~:text=In humans%2C the hormones responsible,the adolescent life history stage.
28. Children's Hospital of Philadelphia. Osteosarcoma (Bone Cancer in Children) [Internet]. 2020. Available from: https://www.chop.edu/conditions-diseases/osteosarcoma
29. CDC. MMR Vaccination [Internet]. 2020. Available from: https://www.cdc.gov/vaccines/vpd/mmr/public/index.html#:~:text=CDC recommends all children get,days after the first dose.
30. American Veterinary Medical Association. Leptospirosis [Internet]. 2020. Available from: https://www.avma.org/resources/pet-owners/petcare/leptospirosis
31. Racine E. American Kennel Club [Internet]. Can Dogs Get Measles? 2019. Available from: https://www.akc.org/expert-advice/health/can-dogs-get-measles/#:~:text=Fortunately%2C the short answer is,the same family as measles.
32. Crémet L, Broquet A, Brulin B, Jacqueline C, Dauvergne S, Brion R, et al. Pathogenic potential of Escherichia coli clinical strains from orthopedic implant infections towards human osteoblastic cells. Pathog Dis. 2015;73(8).
33. Freeman JT, Anderson DJ, Sexton DJ. Seasonal peaks in Escherichia coli infections: possible explanations and implications. Clin Microbiol Infect. 2009;15(10).
34. Clark M. E. Coli Infection In Dogs: Symptoms, Causes, And Treatments [Internet]. 2020 [cited 2020 Dec 8]. Available from: https://dogtime.com/dog-health/53345-e-coli-infection-dogs-symptoms-causes-treatments
35. McClelland MR, Vaszar LT, Kagawa FT. Pneumonia and Osteomyelitis Due to Legionella longbeachae in a Woman with Systemic Lupus Erythematosus. Clin Infect Dis. 2004;38(10).
36. FALCONI TMA, CRUZ MS, NAUMOVA EN. The Shift in Seasonality of Legionellosis in the U.S. Epidemiol Infect. 2018;146(14):1824–33.
37. Wag Walking. Can Dogs Get Legionnaires' Disease? [Internet]. 2020 [cited 2020 Dec 8]. Available from: https://wagwalking.com/wellness/can-dogs-get-legionnaires-disease
38. Charlier C, Leclercq A, Cazenave B, Desplaces N, Travier L, Cantinelli T, et al. Listeria monocytogenes-associated joint and bone infections: a study of 43 consecutive cases. Clin Infect Dis. 2012;54(2):240–8.
39. CDC. Get the Facts about Listeria [Internet]. 2020 [cited 2020 Dec 8]. Available from: https://www.fda.gov/animal-veterinary/animal-health-literacy/get-facts-about-listeria
40. MacGowan AP, Bowker K, McLauchlin J, Bennett PM, Reeves DS. The occurrence and seasonal changes in the isolation of Listeria spp. in shop bought food stuffs, human faeces, sewage and soil from urban sources. Int J Food Microbiol. 1994;21(4):325–34.
41. Moore SM, Eisen RJ, Monaghan A, Mead P. Meteorological Influences on the Seasonality of Lyme Disease in the United States. Am J Trop Med Hyg. 2014;90(3):486–496.
42. Meyers H. Lyme Disease in Dogs: Symptoms, Tests, Treatment, and Prevention [Internet]. 2020 [cited 2020 Dec 8]. Available from: https://www.akc.org/expert-advice/health/lyme-disease-in-dogs/
43. ScienceDaily. Measles virus plays role in Paget's disease of bone, researchers say [Internet]. Science News from research organizations. 2011. Available from: https://www.sciencedaily.com/releases/2011/01/110114155338.htm#:~:text=Measles virus plays role in Paget's disease of bone%2C researchers say,-Date%3A January 16&text=Summary%3A,bone%2C

according to new research.
44. Micaela Elvira Martinez. The calendar of epidemics: Seasonal cycles of infectious diseases. PLOS Pathog. 2018;14(11).
45. Heuer V. Mumps in Dogs [Internet]. 2009 [cited 2020 Dec 8]. Available from: https://www.petmd.com/dog/conditions/infectious-parasitic/c_dg_mumps
46. Shah AP, Smolensky MH, Burau KD, Cech IM, Lai D. Seasonality of primarily childhood and young adult infectious diseases in the United States. Chronobiol Int. 2006;23(5):1065–82.
47. CDC. Pertussis [Internet]. Immunology and Vaccine-Preventable Diseases – Pink Book. 2015 [cited 2020 Dec 8]. p. Pertussis. Available from: https://www.cdc.gov/vaccines/pubs/pinkbook/downloads/pert.pdf
48. Wag Walking. Can Dogs Get Pertussis? [Internet]. 2020 [cited 2020 Dec 8]. Available from: https://wagwalking.com/wellness/can-dogs-get-pertussis
49. Reed GB. Rubella bone lesions. J Pediatr. 1969;74(2):208–13.
50. World Health Organization. Rubella [Internet]. 2020 [cited 2020 Dec 8]. Available from: https://www.who.int/biologicals/areas/vaccines/rubella/en/
51. S. McAnearney, McCall D. Salmonella Osteomyelitis. Ulster Med J. 2015;84(3):171–2.
52. US FDA. Get the Facts about Salmonella [Internet]. 2020 [cited 2020 Dec 8]. Available from: https://www.fda.gov/animal-veterinary/animal-health-literacy/get-facts-about-salmonella#:~:text=Salmonellosis is uncommon in dogs,other pets in the household.
53. Magossi G, Bai J, Cernicchiaro N, Jones C, Porter E, Trinetta V. Seasonal Presence of Salmonella spp., Salmonella Typhimurium and Its Monophasic Variant Serotype I 4,[5],12:i:-, in Selected United States Swine Feed Mills. Foodborne Pathog Dis. 2019;16(4).
54. Cedars-Sinai. Tetanus [Internet]. 2020. Available from: https://www.cedars-sinai.org/health-library/diseases-and-conditions/t/tetanus.html
55. Barnette C. Tetanus in Dogs [Internet]. VCA Hospitals. 2020 [cited 2020 Dec 8]. Available from: https://vcahospitals.com/know-your-pet/tetanus-in-dogs#:~:text=Although tetanus can be seen,toxin than humans and horses.
56. Clark W. Health J, Zusman J, Sherman IL. Tetanus in the United States, 1950-1960. Am J Public Health. 1964;
57. Carlos Pigrau-Serrallach, Rodríguez-Pardo D. Bone and joint tuberculosis. Eur Spine J. 2013;22(4):556–66.
58. Ampel NM. Tuberculosis Is Seasonal in the U.S. New Engl J Med J Watch. 2012;
59. Thoen CO. Tuberculosis in Dogs [Internet]. 2018 [cited 2020 Dec 8]. Available from: https://www.merckvetmanual.com/dog-owners/disorders-affecting-multiple-body-systems-of-dogs/tuberculosis-in-dogs
60. Children's Hospital of Philadelphia. Vaccine History: Developments by Year [Internet]. Vaccine Education Center. 2019 [cited 2020 Jul 13]. Available from: https://www.chop.edu/centers-programs/vaccine-education-center/vaccine-history/developments-by-year#:~:text=In 1963 the measles vaccine,the MMR vaccine in 1971.
61. PA DOH. School Vaccination Requirement Fact Sheet [Internet]. School Immunizations. 2017 [cited 2020 May 8]. Available from: https://www.health.pa.gov/topics/Documents/School Health/SIR8.pdf
62. U.S. Department of Health & Human Services. Rubella (German Measles) [Internet]. vaccine.gov. 2020 [cited 2020 May 8]. Available from: https://www.vaccines.gov/diseases/rubella#:~:text=for your child.-,Adults,doses of the rubella vaccine.
63. Gorska-Ponikowska M, Ploska A, Jacewicz D, Szkatula M, Barone G, Bosco G Lo, et al. Modification of DNA structure by reactive nitrogen species as a result of 2-methoxyestradiol–induced neuronal nitric oxide synthase uncoupling in metastatic osteosarcoma cells. Redox Biol. 2020;32:101522.
64. Dimitrakakis C, Bondy C. Androgens and the breast. Breast cancer Res. 2009;11(5).
65. Maran A, Zhang M, Kennedy A., Sibonga J., Rickard D., Spelsberg T., et al. 2-methoxyestradiol induces interferon gene expression and apoptosis in osteosarcoma cells. Bone. 2002;30(2):393–8.
66. Envita Medical Center. How Lyme Disease and Chronic Infections Can Lead to Cancer [Internet]. 2020. Available from: https://www.envita.com/lyme-disease/lyme-disease-and-chronic-infections-can-lead-to-cancer
67. Hawai'i Pacific Health Straub Medical Center. Travel Medicine [Internet]. 2020 [cited 2020 Dec 18]. Available from: https://www.hawaiipacifichealth.org/straub/services/travel-medicine/
68. Larzábal M, Cataldi AA, Vilte DA. Human and Veterinary Vaccines against Pathogenic Escherichia coli. In: The Universe of Escherichia coli. IntechOpen; 2019. p. 1–20.
69. Cooley LA. Legionellosis (Legionnaires' Disease & Pontiac Fever) [Internet]. Travelers' Health. 2019 [cited 2020 May 8]. Available from: https://wwwnc.cdc.gov/travel/yellowbook/2020/travel-related-infectious-diseases/legionellosis-legionnaires-disease-and-pontiac-fever
70. Calderón-González R, Frande-Cabanes E, Bronchalo-Vicente L, Lecea-Cuello MJ, Pareja E, Bosch-Martínez A, et al. Cellular vaccines in listeriosis: role of the Listeria antigen GAPDH. Front Cell Infect Microbiol. 2014;4.
71. Merck Animal Health. Nobivac Lyme Vaccine [Internet]. Nobivac. 2020. Available from: https://www.merck-animal-health-usa.com/nobivac/nobivac-canine-vaccine-lyme
72. MacLennan CA, Martin LB, Micoli F. Vaccines against invasive Salmonella disease: Current status and future directions. Hum Vaccin Immunother. 2014;10(6):1478–1493.
73. CDC. Tuberculosis (TB) Vaccination [Internet]. Vaccines and Preventable Diseases. 2009. Available from: https://www.cdc.gov/vaccines/vpd/tb/index.html
74. CDC. Questions and Answers About TB [Internet]. Tuberculosis (TB). 2012. Available from: https://www.cdc.gov/tb/publications/faqs/qa_latenttbinf.htm#Latent4

Explainable AI-based clinical decision support system for hearing disorders

Katarzyna A. Tarnowska, Ph.D. , Brett C. Dispoto, B.S., Jordan Conragan, B.S.
San Jose State University, San Jose, CA

Abstract

In clinical system design, human-computer interaction and explainability are important topics of research. Clinical systems need to provide users with not only results but also an account of their behaviors. In this research, we propose a knowledge-based clinical decision support system (CDSS) for the diagnosis and therapy of hearing disorders, such as tinnitus, hyperacusis, and misophonia. Our prototype eTRT system offers an explainable output that we expect to increase its trustworthiness and acceptance in the clinical setting. Within this paper, we: (1) present the problem area of tinnitus and its treatment; (2) describe our data-driven approach based on machine learning, such as association- and action rule discovery; (3) present the evaluation results from the inference on the extracted rule-based knowledge and chosen test cases of patients; (4) discuss advantages of explainable output incorporated into a graphical user interface; (5) conclude with the results achieved and directions for future work.

Introduction

Background Tinnitus commonly referred to as "ringing in the ears", is a phantom auditory disorder estimated to affect about 15% of the global population, with patients existing on a wide spectrum of symptom severity[1]. Although there is currently no known cure for tinnitus, there do exist management techniques; in particular, tinnitus retraining therapy (TRT) has been shown to yield high success rates in clinical trials[2]. TRT is based on the neurophysiological model of tinnitus and is administered as a combination of clinical counseling and sound therapy[3]. The goal of TRT is the habituation of tinnitus, meaning that patients who have undergone successful treatment will still be aware of the phantom auditory signal, but will no longer be agitated by its existence. Despite its clinical success, the infrastructure surrounding tinnitus retraining therapy has a few obstacles to overcome until the treatment can be widely and efficiently adopted by medical practitioners. TRT is a highly personalized treatment that requires patients to receive care over a substantial period of time in order to be successful. Presently, physicians must be deeply experienced in TRT in order to provide effective treatment. It is relatively a niche treatment and it takes years of practice to establish the expertise necessary. Throughout treatment, an effective physician will assess patient progress and modify treatment protocol accordingly. Physicians generally make these decisions based on their knowledge of TRT heuristics along with a deeper knowledge of the neurophysiological model of tinnitus: both of which are lacking in both general-care and ear-nose-throat/audiology practitioners.

Objectives For the reasons outlined above, a data-driven knowledge-based clinical decision support system (CDSS) for the diagnosis and treatment of tinnitus was proposed[4]. This work implements and evaluates the proposed strategy by developing a graphical user interface (GUI) and a knowledge base (KB) with the inference component tested against patient cases. The requirements for a CDS system were determined with the following objectives:

- make secondary use of tinnitus electronic health records with the goal of discovering novel actionable patterns in TRT with regard to treatment outcomes,
- present interpretable and user-friendly real-time clinical advice in TRT delivery, and
- integrate seamlessly with an audiology clinical workflow increasing the likelihood of use.

In turn, it is expected that a system meeting the above criteria will help meet the broader clinical goals:

- wider adoption of an effective tinnitus retraining therapy,
- improved efficiency in the management of tinnitus symptoms, and
- improved effectiveness of treatment outcomes resulting in better care for tinnitus patients.

Methods

Explainable AI for interpretability Despite the promise of assisting human decision making through data-driven approaches, system users often find it challenging to explain and interpret the AI-algorithms behind the transformations of the system's input into a recommendable output. Physicians holding medical responsibility can hardly trust the system's results without an explanation of its underlying decision-making process. To address this issue, we propose an explainable approach for clinical system development. The CDS provides natural-language justification for its output, which is relatively simple to understand by humans. The methods follow the *precision medicine* and *personalized medicine* paradigms to help audiologists better diagnose the category of a hearing problem and treat it in an effective quantifiable manner.

Three-tier implementation for scalability The prototype implementation of the proposed explainable CDS strategy for tinnitus diagnosis and treatment follows a three-tier system architecture (see Figure 1). It is expected that the proposed Java-based platform for CDS will result in wider adoption of the generic CDS with seamless integration into EHR systems, as well as more maintainable, organized, and flexible changes in the CDS component. The so-called *Model-View-Controller* system design pattern was utilized to develop a CDS prototype, called eTRT (*electronic Tinnitus Retraining Therapy*). Data *Model* and back-end rule-based logic are separated from the front-end human-computer interaction module (the *View*), and the synchronization between the Model and the View data is handled through the *Controller*.

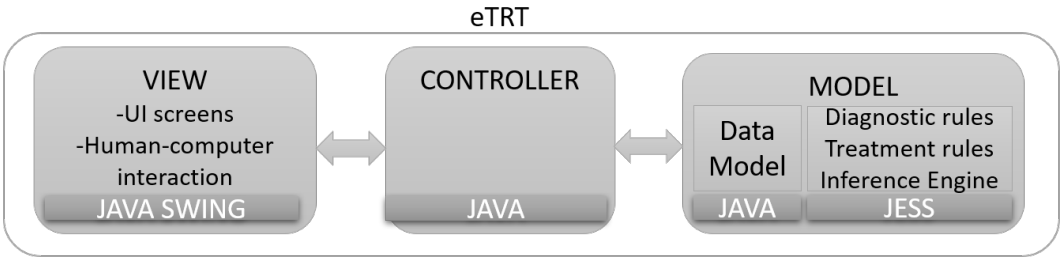


Figure 1: Three-tier system architecture for eTRT - a prototype CDS for tinnitus diagnosis and therapy. The user interface layer (the *View*) was developed in a graphical toolkit for Java Swing. Data *Model* is a separate system component and includes a knowledge base with a rule engine implemented in the Java expert system shell (JESS). The *Controller* synchronizes the View and the Model.

Graphical interface for usability The user interface (UI) of the system is of critical importance, being the only mode of interaction between the physician and the underlying CDS model. Before the system is applied in the health-care environments in a usable manner, a user-friendly graphical UI has to be designed and implemented. Within the proposed solution, the CDS component integrates into clinical TRT workflow as depicted in Figure 2. The prototype GUI was developed in a cross-platform graphical toolkit for Java, called Swing, with customary component extensions for screen development. The developed UI supports clinical processes in:

- Storing and managing the data related to:
  - Tinnitus patients - demographics, medical history, audiology evaluations, and structured interviews[6];
  - TRT visits - diagnoses, treatment applied (sound therapy and counseling), and the outcome evaluation with standardized forms, such as *Tinnitus Handicap Inventory* (THI)[7] and *Tinnitus Functional Index* (TFI)[8].
- Providing evidence-based diagnostic and treatment decision support with explanations and quantifiable predictive outcomes.

The sample GUI screens of the CDS supporting an audiologist in consultation with a new tinnitus patient consultation are presented in Figure 3.

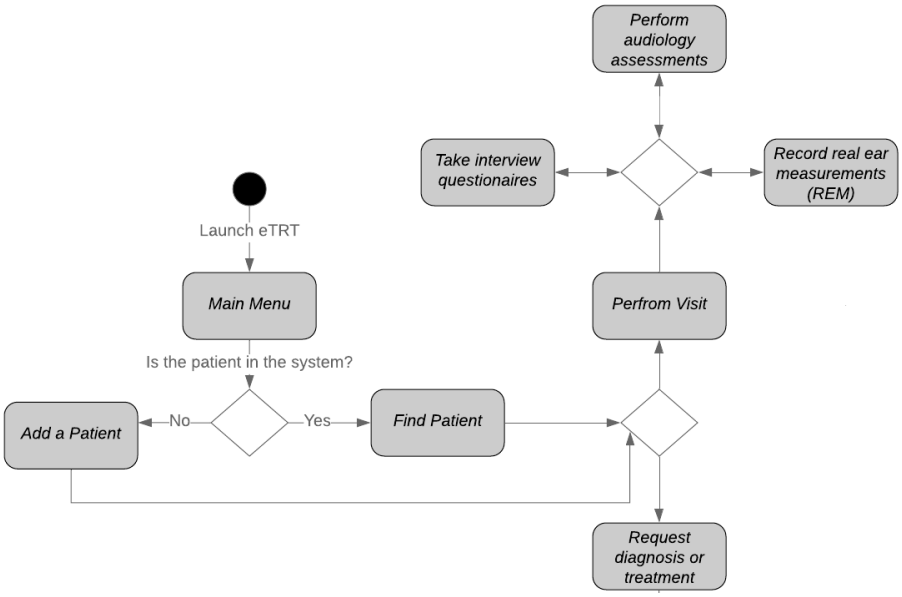


Figure 2: The design of human-computer interaction that integrates eTRT in a clinical workflow.

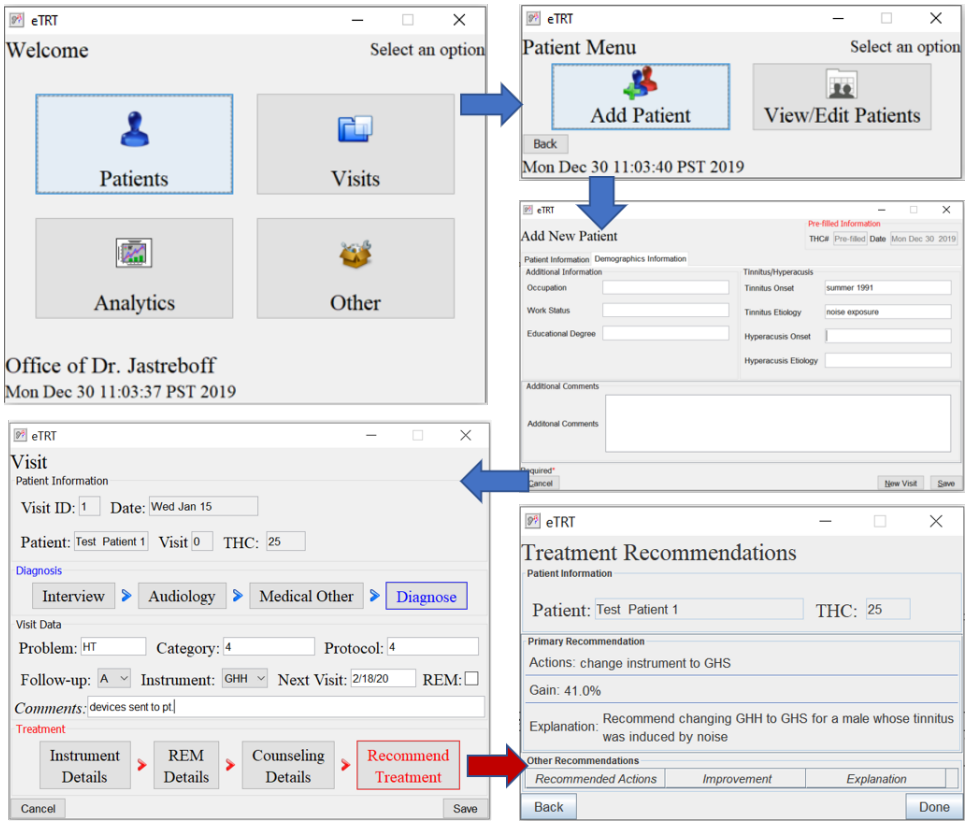


Figure 3: Sample GUI-based clinical TRT workflow with eTRT system: (1) new patient data entry; (2) new visit data entry; (3) treatment recommendations inferred from the current patient/visit data and eTRT knowledge base.

Rule-based knowledge representation for flexibility Within the logic layer, a flexible and interpretable rule-based data model was incorporated which is expected to augment the clinician's knowledge to provide better healthcare for tinnitus. The clinical decision rules are automatically exported from the output files of the data mining software (LISp-Miner) and encoded into CDSS using the Encoder component (see Figure 4). This architecture allows for an efficient update of the knowledge base once new raw data on treatment outcomes becomes available (self-adaptation). The preliminary knowledge discovery used a clinical dataset describing 555 unique patients and 3000 TRT visits, recorded by Dr. Jastreboff over several years of clinical practice at Tinnitus and Hyperacusis Center of Emory University School of Medicine[5]. The dataset was anonymized in a clinic before sharing, therefore no IRB protocol was involved. The dataset was pre-processed, including data cleansing, data transformation, and feature extraction, described in more detail in[5]. The proposed machine learning models for decision support in the diagnosis and treatment include *association rules* and *action rules*. The extracted rules were validated with Dr. Jastreboff as a domain expert in TRT.

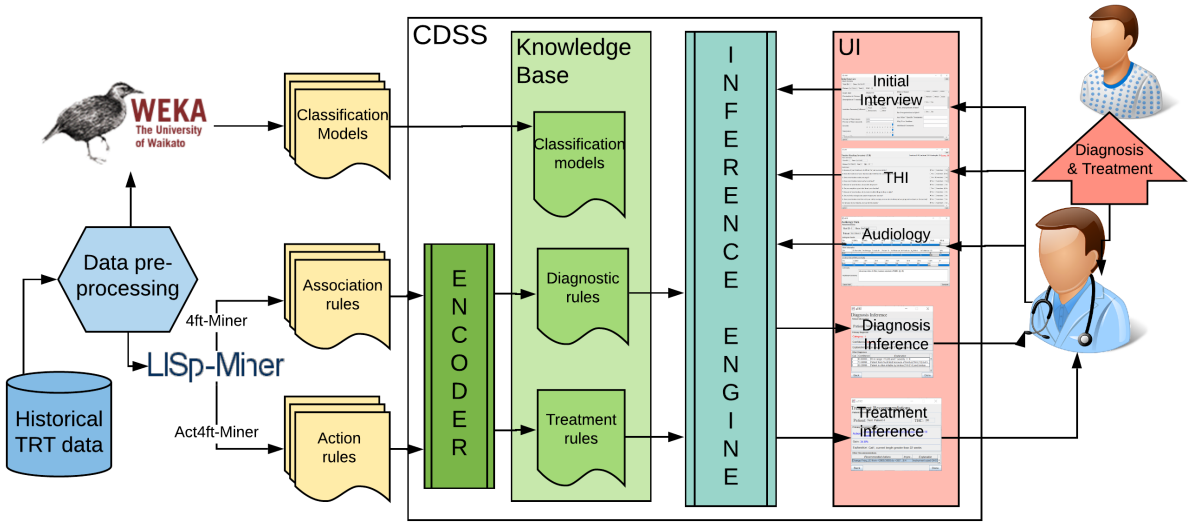


Figure 4: The high-level architecture of a data-driven CDSS for tinnitus with the *Encoder* component automatizing knowledge engineering, *Knowledge Base* with *Inference Engine* providing decision support in diagnosis and treatment.

Diagnostic rules TRT protocol differentiates five categories of the hearing problem: category 0 indicates tinnitus as a minimal problem; 1–tinnitus as a significant problem; 2–tinnitus as significant and hearing loss as existing; 3–tinnitus irrelevant, but hyperacusis (decreased sound tolerance) is significant; 4 is characterized by prolonged tinnitus/prolonged exacerbation of hyperacusis. The accurate categorization is critical, as it determines the further treatment protocol, and therefore the effectiveness of TRT. The diagnosis process defined by the TRT protocol was modeled with a decision rule concept, which is defined as the following:
A **decision rule** is a rule r in the form $(\phi \Rightarrow \delta)$, where ϕ is called *antecedent* (or assumption) and δ is called *descendant* (or thesis). Each rule is characterized by *support* and *confidence*. *Support(r)* is defined as the number of objects matching the rule's antecedent. *Confidence(r)* is the relative number of objects matching both the rule's antecedent and descendant of the rule. Listing 1 presents a sample diagnostic rule encoded in eTRT using JESS syntax. The diagnostic pattern consists of the premises and the conclusion, a confidence metric, and a natural-language explanation.

Listing 1: A sample diagnostic rule encoded in eTRT using JESS syntax.

```
(defrule C1-HLpr-Hpr-Tpr
        (Interview {hpr >= 0 && hpr < 0.5 & hlpr >= 0 && hlpr < 0.5 &&
        tpr >= 6 && tpr < 8}) => (add (new Diagnosis 1 85% "hyperacusis and
        hearing loss indicated as low, but tinnitus indicated as high")))
```

Treatment rules The treatment decisions in TRT are supported with actionable knowledge encoded into eTRT. The action rule concept is a novel way in machine learning proposed by Ras and Wieczorkowska[10]. Since its introduction in 2000, the application of action rules was proposed, among others, for business, medicine, and music indexing[10,11,12,13]. Action rule is defined as the following logical term:

Action rule r is a term $[(\omega) \wedge (\alpha \rightarrow \beta) \Rightarrow (\theta \rightarrow \psi)]$, where $(\omega \wedge \alpha) \Rightarrow \theta$ and $(\omega \wedge \beta) \Rightarrow \psi$ are classification rules, ω is a conjunction of stable attribute values, $(\alpha \rightarrow \beta)$ shows changes in flexible attribute values, and $(\theta \rightarrow \psi)$ shows the desired effect of the action. Now we give an example assuming that a is a stable attribute, b is a flexible attribute, and d is a decision attribute. Terms $(a, a_2), (b, b_1 \rightarrow b_2), (d, d_1 \rightarrow d_2)$ are examples of *atomic actions*. Expression $r = [(a, a_2) \wedge (b, b_1 \rightarrow b_2)] \Rightarrow (d, d_1 \rightarrow d_2)$ is an example of an action rule saying that if value a_2 of a in a given object remains unchanged and its value of b will change from b_1 to b_2, then its value of d is expected to transition from d_1 to d_2. Listing 2 presents a sample action rule extracted from tinnitus datasets and encoded into eTRT using JESS syntax. In this rule, the stable part is defined by gender (male) and etiology (noise), which constitutes the fixed patient's profile. The flexible (changeable) attribute is the instrument used in sound therapy and the recommendation indicated by the rule's conclusion is changing the instrument's model from GH hard to GH soft to improve THI score with 80% confidence. The explanation of the rule is encoded in natural language by the *Encoder* component.

Listing 2: A sample action rule for treatment encoded in eTRT using JESS syntax.

```
(defrule Gm-NTI-GHH
        (Patient {gender == "M" && tEtiology == "NTI"})
        (Instrument {it == "GHH"})
        => (add (new Treatment "GHH -> GHS" 80 "change GHH to GHS
        for a male whose tinnitus was induced by noise exposure")))
```

Action rule modeling is especially promising in the field of medical data, as a doctor can examine the effect of treatment decisions on a patient's improved state. This technique is also particularly useful for building knowledge-based decision support systems since it provides actionable advice needed by practitioners.

Inference engine There are hundreds of such rules extracted from the dataset in the knowledge discovery process. Each rule represents a small piece of the expert's knowledge available from the clinical dataset through machine learning. Rules, such as these presented in Listings 1 and 2, are numerous which presents a challenge for encoding them in the eTRT knowledge base. The manual encoding would be very inefficient and error-prone. Therefore, the automatic conversion was implemented as the Encoder component, which extracts rules from the LISp-Miner output files, parses them, and translates them into JESS syntax, including explanations encoded in natural language. Once rules are encoded into KB by Encoder, the *inference* or *deduction* controls the application of the rules to the patient cases. The method for the inference applied within this research is based on an efficient pattern-matching algorithm called Rete[14], which is provided by the JESS framework. The deduction algorithm matches the current patients' data entered into the system with the machine-learned diagnostic/treatment rules in the knowledge base. If the left-hand side of the rule (antecedent) is matched with the current patient, the right-hand side (consequent) of a rule is executed. Consequent clauses decide about the diagnosis/treatment decision suggested and displayed to the physician via GUI.

Results

Association patterns discovery In the experiments on decision rule discovery LISp-Miner 4ft-Miner module was used[9]. In the pattern formulation, the descendant of the rule (the decision attribute) was defined as the TRT category (0-4). For the antecedent part, attributes describing a patient, such as demographics etiology, pharmacology, audiometry, severity, and effect on life were chosen. The examples of the found associations between the variables describing the tinnitus patient and the TRT category are provided in Table 1. These rules are interpreted as follows:

- *If* hyperacusis H_{pr} and hearing loss HL_{pr} not indicated as problems, but tinnitus T_{pr} indicated a problem - *then* a patient falls under Category 1 with 85% confidence.
- *If* audiometric values of L_2 (audiogram at 2kHz for the left ear) is greater or equal to 50 and R_6 (audiogram at 6kHz for the right ear) is less or equal to 75, *then* a patient falls under Category 2 with 87% confidence.

Table 1: Sample association rules between patient characteristics and the TRT category.

Diagnostic association rule	**Confidence**
$H_{pr}(<0;0.5)) \wedge HL_{pr}(<0;0.5)) \wedge T_{pr}(<6;8)) \Rightarrow Category(1)$	85%
$L_2 \geq 50 \wedge R_6 \leq 75 \Rightarrow Category(2)$	87%

Actionable patterns discovery In the experiments on actionable pattern discovery LISp-Miner Act4ft module was used. The actionable pattern was formulated by choosing the patient's profile (e.i. age, gender, etiology) in the stable (fixed) part, and treatment methods as flexible (or changeable in the course of treatment). Treatment methods in TRT include sound therapy with different categories, types, and models of instrument, real-ear measurements (REM) supporting instrument fitting set to optimal numerical parameters, and counseling delivered within individual therapy. The goal of the actionable pattern (or the desired effect) is to decrease the severity of tinnitus as indicated by the *total score* of tinnitus handicap inventory. To indicate the change, the additional, temporal attribute was added and its values were imputed as calculated distance-based percentage change metric between visits. Table 2 presents examples of extracted action rules, which indicate changes in the settings of the instruments, or changing the length of a particular treatment that lead to patient's improvement. These rules are interpreted as follows:

Table 2: Examples of discovered action rules for recommending treatment in TRT.

Treatment action rule	**Conf.**
$T_{side}(yes) \wedge OMTI(yes) : (Ins_{vis(01)}GHH \rightarrow V) \wedge FU(0 \rightarrow T) \Rightarrow Ch(better)$	82%
$Ins(SG) : (Mix_{RSL}(<11;12) \rightarrow <9;10)) \Rightarrow Ch(better)$	100%
$FU(A) \wedge Ins_{vis(01)}(GHI) \wedge Freq_{LE}(<3000;3150)) : (treat(<5;6) \rightarrow <6;8)) \Rightarrow Ch(better)$	88%

- *If* tinnitus was induced by other medical condition (*OMTI*), and as a side effect of taking medications (T_{side}), *then* changing the sound generator model GH hard (GHH) to the Viennatone model (V) at the first visit and changing the follow-up contact to the telephone-based (T) improves patient with 82% confidence.
- *If* the current treatment involves sound generator *SG*, *then* changing mixing point for the right ear Mix_{RSL} from $< 11; 12)$ to $< 9; 10)$ improves a patient's state with 100% confidence.
- *If* the current treatment involves audiology ($FU(A)$) with the GHI instrument, and frequency in the left ear measured by REM -$Freq_{LE}$ - in the range of $< 3000; 3150)$ *then* prolonging that treatment from 5-6 weeks to 6-8 weeks brings improvement with 88% confidence.

Inference evaluation The evaluation objective is to determine whether the prototype system does what it was intended to and at an adequate level of accuracy. The system is expected to generate accurate, patient-specific, and interpretable clinical suggestions. This will encourage efficient and effective use of tinnitus retraining therapy for the management of hearing disorders. The evaluation method includes:

1. Identifying a set of representative test cases of patients from the dataset not used for building the model.
2. Running inference on the chosen test cases entered into the system (see Figures 5–7).
3. Performing quantitative and qualitative evaluation of the system based on the results from the above.

The metrics used for this evaluation of the system include:

- Accuracy - the number of correct predictions versus the total number of predictions. The predictions were compared with the actual diagnosis/treatment decisions from Dr. Jastreboff, who is considered the "gold standard" in TRT, as the founder and years-long practitioner of the method[2].
- Coverage - the number of test cases matched against the knowledge base.
- Interpretability - if the recommendations were provided with human-understandable explanations.

Test cases The goal was to identify the smallest possible *representative* set of test cases. A test patient for each etiology and each category of the hearing problem was selected from the test dataset (see Table 3). The chosen test cases reflect the heterogeneity of the hearing problem and patient profile.

Table 3: Patient test cases - patient profile, etiology of their hearing problem, the actually diagnosed category and the actual treatment protocol as determined by the TRT founder.

Test case	Patient profile	Etiology	Diagnosis	Treatment protocol
1	male, age 38	noise exposure	Category 4	Category 4
2	male, age 49	ear surgery	Category 3	Category 3
3	female, age 77	hearing loss	Category 2	Category 2
4	male, age 53	stress-related	Category 1	Category 1
5	male, age 36	car accident	Category 0	Category 1

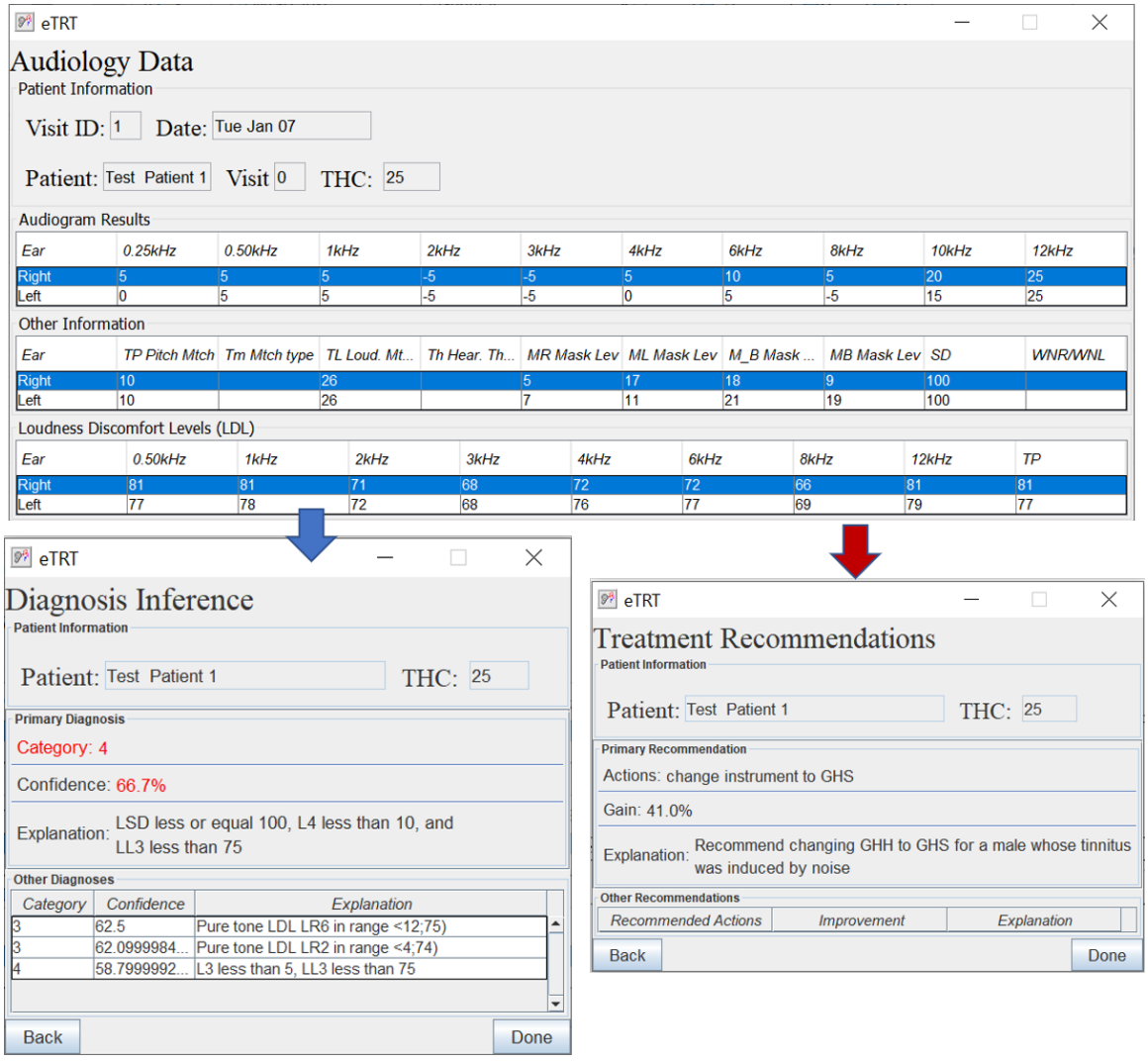


Figure 5: The diagnostic/treatment inference results for test case 1 (noise-based, middle-aged male) based on audiometry: (1) primary diagnosis of category 4 with 66.7%, and (2) treatment recommendation for changing the instrument type with the expected decrease in tinnitus severity by 41 percentage points.

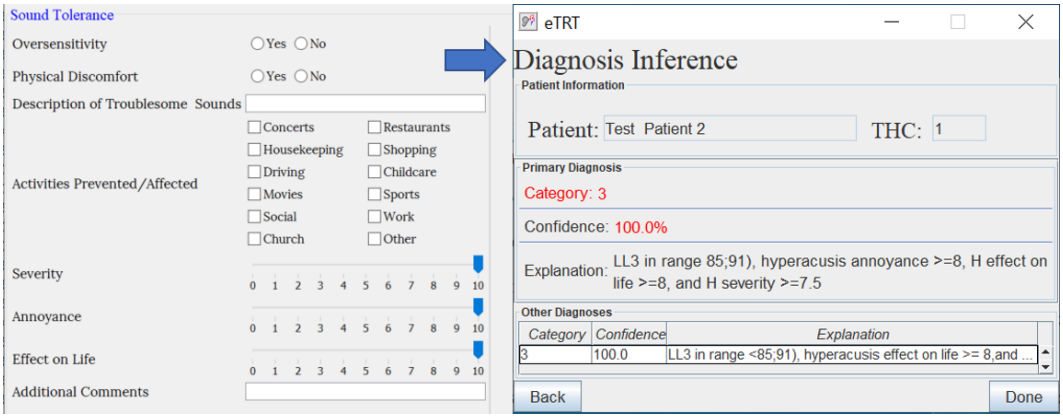


Figure 6: The diagnostic inference for test case 2 based on audiometry/initial interview. The explanation for the category 3 with 100% confidence included a high score for hyperacusis as a problem.

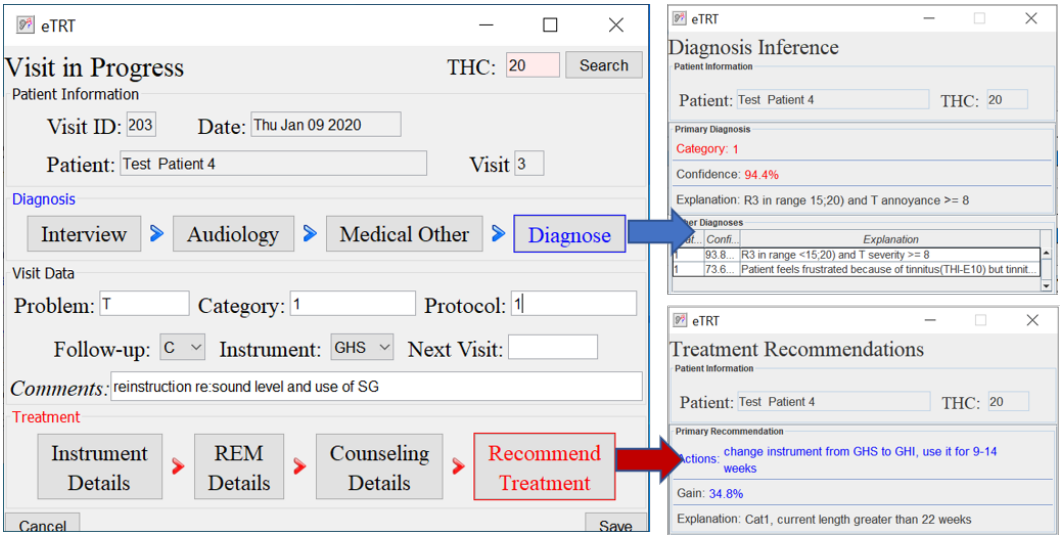


Figure 7: The diagnostic/treatment inference results for test case 4: (1) category 1 was inferred based on the audiometry results and initial interview (annoyance over tinnitus high); (2) recommendation included the change of the sound instrument from GH soft and shorten its application time to 9-14 weeks with an expected gain of 34.4 percentage points.

Result summary Tables 4 and 5 provide the diagnostic and treatment inference results for all test cases. The diagnosis prediction was 80% accurate and covered 100% of cases. The average confidence in the primary diagnosis inference was 83.51 %. The only incorrect prediction was for test case 5. After closer investigation, this case was annotated by the physician as a "discrepancy in information" in interview data, and "inconsistent results" is audiological evaluation, which are the reasons that misled the predictive model (as an "outlier" data point). Moreover, the actual protocol followed was the same as for the category predicted by the system. The treatment recommendations were generated for 3 out of 5 patient test cases. The other two cases were not covered, that is, no action rule was matched with the patient profile, due to a limited number of rules encoded manually in KB at the time of testing. For all the tested cases, both the diagnostic and the treatment recommendations were explained with a human-comprehensible message/reason. The explanations were provided by means of the premises of the rules in KB that were matched against the current patient's profile/visit data. The predictions' probabilities were quantified by means of the matched rules' confidence metric.

Table 4: Actual versus predicted TRT category for the test patients, the confidence, and the explanation behind the prediction.

Test case	Actual	Predicted	Pred. Conf.	Explanation
1	Cat 4	Cat 4	66.7%	LSD <= 100, L4 < 10, and LL3 < 75
2	Cat 3	Cat 3	100%	LL3 in $< 85; 91)$, Hyper. Annoy >=8, H_{Effect} >=8, and H Sev >=7.5
3	Cat 2	Cat 2	96.2%	LR8 >= 999, R6 >= 75, and T_{sv} >= 8
4	Cat 1	Cat 1	94.4%	LL3 in $< 15; 20)$ and Tin. annoy. >= 8
5	Cat 0	Cat 1	60.3%	patient often irritable by tinnitus (E14) and tinnitus makes him anxious (E22)

Table 5: Recommendation generated for the test cases 1,4, and 5. Each recommendation is supported by a predicted gain in the patient's improvement and explanation based on the patient's personalized profile. Due to a limited knowledge base at the time of testing, no action rules matching profiles in test cases 2 and 3 were found, but are expected once a full KB is built.

Test case	Recommended Action(s)	Gain	Explanation
1	Change instrument from GHH to GHS	41 pp	a male whose tinnitus was induced by noise
4	Change instrument from GHS to GHI, use it for 9-14 weeks	34.8 pp	Cat1, instr. duration greater than 22 weeks
5	Change Freq LE from $< 2800; 3000)$ to $< 2670; 2800)$ in REM	8.4 pp	Instrument used GHS

Discussion

The application of novel approaches in actionable data mining helps uncover links between clinical variables, with the goal to develop optimal strategies for tinnitus management. This is expected to open new horizons for TRT, which does not have a stagnant protocol but continues to evolve based on information gathered from treatments of patients[2]. The proposed CDS strategy was implemented and preliminarily evaluated in the eTRT prototype. The next step is the usability study with the target users and later within real patient consultation. Its future deployment in the clinical setting is expected to improve TRT delivery, tinnitus management, and clinical outcomes, such as THI total score and habituation time. The system's output is transparent, and as such, will increase its acceptance in the clinical setting. The clinical decision-makers offered the predictive models, which cannot answer questions, such as "Why did you predict this patient to fall into this category?" or "Why did you recommend these treatment actions for this patient?" will be hesitant to use the system. The lack of transparency is a problem for the clinician who wants to understand the way the model works to help them improve their service. While AI technologies are powerful, the adoption of these algorithms in health care has been slow because doctors and regulators cannot verify their results. For fields such as health care, where mistakes can have catastrophic effects, the "black-box" aspect of AI makes it difficult for doctors and regulators to trust it. Explainable clinical systems have the opportunity to disrupt the health care sector because of their ability to diagnose and produce results efficiently. The trend of explainable AI has grown in recent years and looks set to continue.

From the design phase of the eTRT prototype, through implementation up to the evaluation, the researchers' efforts have focused on the human-computer interaction and the AI explainability. The result is the prototype system that utilizes the explainable AI (XAI) approach by providing justification for its results and is human-understandable. Through rule-based inference, intelligent decisions are made and the automatic decision-making process is traced. The advantage of the approach is not only being similar to human reasoning but also involving a minimal set of predictor variables to infer diagnosis or treatment and therefore reducing cognitive load on the clinical user. The end result is developing a prototype that not only generates new insights but is also more credible. The final outcome of this research is expected to be the realization of "precision medicine" or "personalized medicine" in tinnitus practice, which takes into account individual variability in demographics and etiology of hearing disorder for each patient.

Conclusion

Our contribution is implementing a novel evidence-based CDS strategy for a niche medical condition. This promising tool will help physicians optimize diagnosis and treatment for tinnitus, and will be particularly useful for practitioners, not that experienced or familiar with TRT. The developed prototype CDSS supports clinical decision-making with a high degree of accuracy, covers a wide range of heterogeneous cases, and provides natural language-based interpretation within GUI. The system also provides a degree of certainty for the suggestions and alternatives for a primary diagnostic/treatment hypothesis. The limitations of this study were identified for the future work directions, which includes the following:

- Usability study involving actual users of the system (audiology clinicians) to determine usefulness and acceptance of the system in the clinical setting and testing the system in a real-time patient consultation;
- Machine learning tuning and further model calibration to optimize predictive accuracy and coverage;
- Expanding the knowledge base with more rules and including data from more TRT clinical experts and clinics;
- Developing a strategy to integrate the CDS component within electronic health records (EHR) systems of audiology clinics.

Although this work is specific to audiology, the proposed data-driven AI-based approach to developing a knowledge-based clinical decision support system is applicable to a wide range of disease, where the lack of experts blocks the delivery of effective treatment for the majority of patients.

References

1. American Tinnitus Association. Understanding the facts [Internet], Available from: www.ata.org/understanding-facts. [Cited 29 Jul 2020]
2. Jastreboff PJ. 25 years of tinnitus retraining therapy. HNO. 2015 Apr;63(4):307-11.
3. Jastreboff PJ, Hazell JWP. Tinnitus retraining therapy: implementing the neurophysiological model. Cambridge: Cambridge University Press; 2004.
4. Tarnowska KA, Ras ZW, Jastreboff PJ. Decision support system for diagnosis and treatment of hearing disorders. The case of tinnitus. Series in Computational Intelligence 685, Springer; 2017.160 p.
5. Tarnowska KA, Ras ZW, Jastreboff PJ. Mining for actionable knowledge in tinnitus datasets. In: Thriving Rough Sets. International: Springer; 2017. p.367-395.
6. Jastreboff MM, Jastreboff PJ. Questionnaires for assessment of the patients and treatment outcome, Sixth International Tinnitus Seminar, 1999.
7. Newman CW, Wharton JA, Jacobson GP. Retest stability of the tinnitus handicap questionnaire. Ann Otol Rhinol Laryngol 1995;104:718-23.
8. American Academy of Audiology. Tinnitus functional index. [Internet] Available from: https://www.audiology.org/news/tinnitus-functional-index [Cited 29 Jul 2020]
9. Simunek M. LISp-Miner control language description of scripting language implementation. Journal of Systems Integration. 2014;5(2):28–44.
10. Ras ZW, Wieczorkowska A. Action-rules: how to increase profit of a company. In: Principles of data mining and knowledge discovery. Springer; 2000. p.587–592.
11. Tarnowska K, Ras Z, Daniel L. Recommender system for improving customer loyalty. Studies in Big Data 55, Springer; 2019. 124 p.
12. Wasyluk H, Ras ZW, Wyrzykowska E. Application of action rules to HEPAR clinical decision support system. Experimental and Clinical Hepatology Bd. 2008;4(2):46–48.
13. Ras ZW, Wieczorkowska A. Advances in music information retrieval. Studies in Computational Intelligence, Springer; 2010.
14. Forgy CL. Rete: A fast algorithm for the many pattern/many object pattern match problem. Artificial intelligence. 1982;19(1):17–37.

Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization

Byron C. Wallace, PhD[1], Sayantan Saha, BS[1],
Frank Soboczenski, PhD[2], Iain J. Marshall, MD, PhD[2],
[1]Northeastern University, Boston, MA; [2]King's College London, London

Abstract

We consider the problem of automatically generating a narrative biomedical evidence summary from multiple trial reports. We evaluate modern neural models for abstractive summarization of relevant article abstracts from systematic reviews *previously conducted by members of the Cochrane collaboration, using the* authors conclusions *section of the review abstract as our target.*[1] *We enlist medical professionals to evaluate generated summaries, and we find that summarization systems yield consistently fluent and relevant synopses, but these often contain factual inaccuracies. We propose new approaches that capitalize on domain-specific models to inform summarization, e.g., by explicitly demarcating snippets of inputs that convey key findings, and emphasizing the reports of large and high-quality trials. We find that these strategies modestly improve the factual accuracy of generated summaries. Finally, we propose a new method for* automatically *evaluating the factuality of generated narrative evidence syntheses using models that infer the directionality of reported findings.*

1 Introduction

Biomedical *systematic reviews* aim to synthesize all evidence relevant to a given clinical query[1,2]. Such reviews typically comprise both quantitative and narrative summaries of the evidence. The former is most often a statistical meta-analysis of the results reported in the constituent trials, which in turn informs the natural language interpretation provided in the latter. In Cochrane reviews,[2] brief narrative summaries communicating the main review findings are provided in structured abstracts in the *Authors' conclusions* section. Below (left) is an example from a review of the evidence concerning the use of inhaled antibiotics for cystic fibrosis[3]. We also provide the summary generated by one of the automated models we evaluate (right), given the abstracts of the included papers.

Authors' conclusions Inhaled antibiotic treatment probably improves lung function and reduces exacerbation rate, but a pooled estimate of the level of benefit is not possible. The best evidence is for inhaled tobramycin. More evidence, from trials of longer duration, is needed to determine whether this benefit is maintained and to determine the significance of development of antibiotic-resistant organisms.	***Automatically generated summary*** Inhaled antibiotics are effective in the treatment of Pseudomonas aeruginosa pulmonary infection in CF patients. However, there is a need for further randomised controlled trials to assess long-term safety and efficacy of inhaled antibiotics in patients with cystic fibrosis. Further trials are also needed to assess the effects of antibiotic treatment on morbidity and mortality.

Table 1: Example *Author conclusions* from a Cochrane systematic review abstract (left) and an automatically generated summary (right), conditioned on the set of clinical trial abstracts that informed the corresponding review.

Narrative summaries of clinical trials are invaluable for practitioners because they provide a concise, readable summary of all evidence relevant to the clinical question that motivated the corresponding review. These summaries are not verbatim distillations of the original trial reports, but can be considered as *critical* summaries. The review authors should consider strengths and weaknesses of the source trials, see through any 'spin' from the clinical trial authors, and emphasize the strongest evidence. The process of generating these summaries manually is lengthy and laborious.

[1]We make this dataset available at: `https://github.com/bwallace/RCT-summarization-data`.
[2]`https://www.cochrane.org/`

Consequently, summaries will not always be available for arbitrary clinical questions (even when relevant trial reports exist). Moreover, even where available they will often be out of date. A system that could automatically summarize clinical trials literature would be capable of summarizing all evidence, on-demand.

In this work we evaluate state-of-the-art multi-document neural abstractive summarization models that aim to produce narrative summaries from the titles and abstracts of published reports of relevant randomized controlled trials (RCTs). We train these models using the *Authors' conclusions* sections of Cochrane systematic review abstracts as targets, and the titles and abstracts from the corresponding reviews as inputs. We evaluate models both quantitatively and qualitatively, paying special attention to the *factuality* of generated summaries.

Related Work

Automatic Summarization and Question Answering for EBM This paper extends a thread of prior work on summarization for EBM[4–6]. Demner-Fushman and Lin led a seminal effort on automatic question answering (QA) from the literature to aid EBM[4]. This work on QA is adjacent to traditional summarization: In their approach they aimed to extract snippets from individual articles relevant to a given question, rather than to *generate* an abstractive summary of relevant abstracts, as is our aim here. Follow-up work on (extractive) QA over clinical literature has further demonstrated the promise of such systems[7]. For recent efforts in this vein, we point the reader to the latest BioASQ challenge iteration[8], which included a biomedical QA task. While related, we view the task of extractive biomedical QA as distinct from the more focussed aim of generating abstractive narrative summaries over relevant input abstracts to mimic narratives found in formal evidence syntheses (Table 1).

Directly relevant to this setting of biomedical systematic reviews, Molla[5,9] introduced a dataset to facilitate work on summarization in EBM that comprises 456 questions and accompanying evidence-based answers sourced from the "Clinical Inquiries" section of the Journal of Family Practice. Sarkar *et al.*[6] surveyed automated summarization and EBM, respectively, highlighting the need for domain-specific multi-document summarization systems to aid EBM. In contrast to our approach, these prior efforts used comparatively small corpora, and pre-dated the current wave of the neural summarization techniques that have yielded considerable progress in language generation and (abstractive) summarization[10–12].

Neural Abstractive Summarization Automatic summarization is a major subfield in NLP[13,14]. Much of the prior work on summarization of biomedical literature has used *extractive* techniques, which directly copy from inputs to produce summaries. However, narrative evidence synthesis is an inherently *abstractive* task — systems must generate, rather than simply copy, text — as it entails communicating an overview of all available evidence.

Recent work on neural models has engendered rapid progress on abstractive summarization[15,16]; we do not aim to survey this extensively here. Illustrative of recent progress — and most relevant to this work — is the Bidirectional and Auto-Regressive Transformers (BART) model[11], which recently achieved state-of-the-art performance on abstractive summarization tasks. Because it forms the basis of our approach, we elaborate on this model in Section 2.

Despite progress in summary generation, evaluating abstractive summarization models remains challenging[17]. Automated metrics calculated with respect to reference summaries such as ROUGE[18] provide, at best, a noisy assessment of text quality. Of particular interest in the setting of evidence syntheses is the *factuality* of generated summaries: Here, as in many settings, users are likely to value accuracy more than other properties of generated text[19,20]. Unfortunately, neural models for abstractive summarization are prone to 'hallucinations' that do not accurately reflect the source document(s), and automatic metrics like ROUGE may not capture this[21].

This has motivated recent efforts to automatically evaluate factuality. Wang *et al.* proposed *QAGS*, which uses automated question-answering to measure the consistency between reference and generated summaries[22]. Elsewhere, Xu *et al.*[23] proposed evaluating text factuality independent of surface realization via Semantic Role Labeling (SRL). We extend this emerging line of work here by manually evaluating the factuality of summaries produced of clinical trial reports, and proposing a domain-specific method for automatically evaluating such narrative syntheses.

2 Methods

Data

We use 4,528 systematic reviews composed by members of the Cochrane collaboration (`https://www.cochrane.org/`). These are reviews of all trials relevant to a given clinical question. The systematic review abstracts together with the titles and abstracts of the clinical trials summarized by these reviews form our dataset. All data was downloaded via PubMed (i.e., we use only abstracts). The reviews include, on average, 10 trials each. The average abstract length of included trials is 245 words. We use the "authors' conclusions" subsection of the systematic review abstract as our target summary (75 words on average). We split this data randomly into 3,619, 455, and 454 reviews corresponding to train, development (dev), and test sets, respectively. The dataset is available at: `https://github.com/bwallace/RCT-summarization-data`.

Models

We adopt Bidirectional and Auto-Regressive Transformers (BART) as our underlying model architecture[11]. This is a generalization of the original BERT[24] Transformer[25] model and pretraining regime in which self-supervision is not restricted to the objectives of (masked) token and next sentence prediction (as in BERT). Instead, BART is defined as an encoder-decoder model with an autoregressive decoder trained to 'denoise' arbitrarily corrupted input texts. Masking tokens — the original BERT objective — is just one type of 'corruption'. This permits use of additional corruption schemes (pretraining objectives), a property that we exploit in this work (Section 2). BART achieves strong performance on abstractive summarization tasks[11], which makes it particularly appropriate for our use here.

BART defines a sequence-to-sequence network[26] in which the *encoder* is a bidirectional Transformer network and the *decoder* is autoregressive (and hence amenable to language generation tasks such as summarization). One limitation of BART (and large neural encoder models generally) is that it imposes a limit on the number of input words that can be accepted due to memory constraints; for BART this limit is 1024. We discuss this further below.

We do not modify the BART architecture, but we explore new, domain-specific pretraining strategies and methods that entail modifying inputs. For the former, we propose and evaluate additional pretraining in which the objective is to construct abstracts of RCT articles from corresponding full-texts (Section 2). For the latter, we propose and evaluate a method in which we 'decorate' input texts with annotations automatically produced by trained models, e.g., we explicitly demarcate (via special tokens) snippets in the input that seem describe interventions and key findings (Section 2). This is a simple (and as far as we aware, novel) means of incorporating prior information or instance meta-data in end-to-end neural summarization models.[3]

Initialization and pre-training strategies

We use the BART-large version of BART,[4] in which both the encoder and decoder are 12-layer Transformers. The 'vanilla' variant of BART is initialized to weights learned under a set of denoising objectives that differ in how they corrupt the input text (which is then to be reconstructed). For example, objectives include token masking (as in the original BERT), 'text infilling', and 'sentence permutation' tasks[11]. This pretraining is performed over a very large corpus comprising: `BookCorpus`[30] and `English Wikipedia`, `CC-News`[31], `OpenWebText`,[5] and `Stories`[32] (over 160GB of raw text in all). We verified via string matching that none of the target summaries (published Cochrane review abstracts) appeared in this pretraining corpora. As a natural starting point for our task, we initialize BART-large weights to those learned via fine-tuning on the XSUM abstractive summarization corpus[33].

With this as our starting point, we explored additional 'in-domain' pretraining prior to learning to summarize trials. Specifically we train BART to generate summaries from full-text articles. Specifically, we use $\sim$60k full-texts from the PubMed Central (PMC) Open-Access set that were classified as describing RCTs in humans by a previously developed model[34]. Full-texts exceed the 1024 token budget imposed by BART, and so we alternate between selecting sentences from the start and end of the article text until we reach this limit.

<T> Colistin inhalation therapy in cystic fibrosis patients with chronic Pseudomonas aeruginosa lung infection. <ABS> Significantly more patients in the colistin inhalation group completed the study as compared to the placebo group (18 versus 11). <pl> Colistin treatment was superior to placebo treatment in terms of a significantly better <out> clinical symptom score, maintenance of pulmonary function and inflammatory parameters </out>.</pl> We recommend colistin inhalation therapy for <pop> cystic fibrosis patients with chronic P. aeruginosa lung infection </pop> as a supplementary treatment to frequent courses of <inter> intravenous anti-pseudomonas chemotherapy </inter>. <s> <T> Significant microbiological effect of inhaled tobramycin in young children with cystic fibrosis. <ABS> We observed no differences between treatment groups for <out> clinical indices, markers of inflammation, or incidence of adverse events </out>. <pl> No abnormalities in <out> serum creatinine or audiometry and no episodes of significant bronchospasm </out> were observed in association with active treatment.</pl> We conclude that 28 days of <inter> tobramycin solution </inter> for inhalation of 300 mg twice daily is safe and effective for significant reduction of lower <out> airway Pa density </out> in <pop> young children with cystic fibrosis </pop>. ...

Figure 1: Input articles (here we show two for illustration) 'decorated' using special tokens to demarcate automatically extracted salient attributes: <pl> for 'punchlines' sentences (those that seem to state the main finding), and snippets of text describing study populations <pop>, interventions <inter>, and outcomes <out>, respectively.

Modifying Inputs

Another important design choice concerns the inputs that we provide to the encoder component of BART. In the most straightforward use, we simply pass along subsets of the raw titles and abstracts. We demarcate titles, abstracts, and the start of new documents with with special tokens ('<T>', '<ABS>', '<S>'). Typically, only some of the abstracts associated with a given review will fit within the aforementioned token limit. We prioritize including titles, and then sentences from the beginnings and ends of abstracts. We select the latter in a 'round-robin' (random) order from inputs, alternating between the starts and ends of abstracts, until the token budget is exhausted.

Decoration Prior work has investigated methods for automatically extracting key trial attributes from reports of RCTs, including descriptions of the study Populations, Interventions/Comparators, and Outcomes (the 'PICO' elements)[35] and identifying 'punchline' snippets that communicate the main study findings[36]. These key aspects of trials ought to figure prominently in summaries of the evidence. But in a standard end-to-end summarization approach, the model would have to implicitly learn to focus on these attributes, which seems inefficient.

We propose a simple 'decoration' strategy in which we explicitly demarcate snippets of text tagged by pretrained models as describing the aforementioned attributes. Decoration entails enclosing snippets (automatically) identified as describing the respective attributes within special tokens that denote these. We provide an example (showing only two articles) in Figure 1. This preprocessing step is a simple mechanism to directly communicate to the encoder which bits of the text seem to provide information for the aspects of interest. To identify PICO elements, we use RobotReviewer[37,38]. To identify punchline sentences, we fine-tuned a BioMed-RoBERTa model[39] on the Evidence Inference (2.0) dataset[40], which includes annotations of evidence-bearing ('punchline') sentences in RCT articles.

Sorting Rather than treating all inputs equally, we might prefer to prioritize inclusion of evidence from large, high-quality studies. To operationalize this intuition, we consider a variant in which we greedily include tokens from abstracts ordered by sample size (N) scaled by an estimate of overall risk of bias (RoB) (a proxy for study quality). We infer both of these quantities automatically using RobotReviewer[37,38]. Here RoB is the predicted probability of a study being at overall low risk of bias, based on the abstract text.

Design

We analyze the performance of five model variants that use elements of the above strategies (see Table 2). All are fine-tuned on the training set of Cochrane reviews. For 'XSUM' we initialize BART to weights estimated on the XSUM abstractive summarization task[41]. For 'Pretrain (PMC)' we continue pretraining over the PMC set as described above; all other models start from this checkpoint. 'Decorate' marks up the inputs as described above before passing them to the encoder (at train and test time). 'Sort by N·RoB' greedily picks 1024 tokens by selecting for inclusion words from abstracts with the lowest (inferred) risk of bias, scaled by (extracted) sample size (N).

[3]Though the general idea of demarcating parts of inputs with special tokens for Transformers has been used for other tasks[27,28].
[4]Provided via the `huggingface Transformers` library[29].
[5]`http://web.archive.org/save/http://Skylion007.github.io/OpenWebTextCorpus`

Name	Initialization	(Additional) Pretraining	System inputs	ROUGE-L (dev)	ROUGE-L (test)
XSUM	*XSUM*	None	Titles and abstracts	0.264	0.265
Pretrain (PMC)	*XSUM*	PMC RCTs	Titles and abstracts	0.263	0.269
Decorate	*XSUM*	PMC RCTs	Decorated	0.268	0.266
Sort by N·RoB	*XSUM*	PMC RCTs	Sorted by $N\cdot$ RoB	0.267	0.267
Decorate and sort	*XSUM*	PMC RCTs	Decorated and sorted	0.265	0.265

Table 2: Model variants and ROUGE-L measures over the dev and test sets. (Results for ROUGE-1 and ROUGE-2 are qualitatively similar.) 'PMC RCTs' is shorthand for our proposed strategy of (continued) pretraining to generate abstracts from full-texts for all RCT reports in PMC. All model variants aside from 'XSUM' start from the Pretrain (PMC) checkpoint.

Hyperparameters During fine-tuning we used learning rate of $3e{-}5$. During decoding we used beam size of 4, a minimum target length of 65, we enabled early stopping, and we prevent three consecutive n-grams from repeating. This largely follows the original BART paper[11]; we did not systematically tune these hyperparameters.

Main outcome measurements

We measure summarization system performance using both automated and manual approaches. For the former we use Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[18], which relies on word overlaps between generated and reference summaries. For the latter we enlisted medical doctors to annotate generated summaries with respect to relevance, plausibility (including fluency), and factuality (i.e., agreement with reference target summaries). For this we built a custom interface; task instructions (with screenshots) are available here: http://shorturl.at/csJPS.

As we later confirm empirically, ROUGE scores do not necessarily capture the factual accuracy of generated summaries, which is critical when generating evidence syntheses. Manual evaluation of summaries can capture this, but is expensive, hindering rapid development of new models. We propose a new approach to automatically evaluating generated narrative evidence syntheses with respect to the factuality of the findings they present. Specifically, we infer the reported directionality of the main finding in the generated and reference summaries, respectively, and evaluate the resultant level of (dis)agreement.

To derive this automated metric we use the Evidence Inference dataset[36], which comprises full-text RCT reports in which evidence-bearing snippets have been annotated, along with whether these report that the finding is a *significant decrease*, *significant increase*, or *no significant difference* with respect to specific interventions, comparators, and outcomes. We simplify this by collapsing the first two categories, yielding a binary classification problem with categories *significant difference* and *no significant difference*. Following DeYoung *et al.*[40], we train a 'pipeline' model in which one component is trained to identify 'punchline' sentences within summaries, and a second is trained to infer the directionality of these findings. Both models are composed of a linear layer on top of BioMed-RoBERTa[39].

Using these models we can infer whether reference and generated summaries appear to agree. Specifically, we use the Jensen-Shannon Divergence (JSD) — a measure of similarity between probability distributions — between the predicted probabilities for *sig. difference* and *no sig. difference* from our inference model for the generated and reference summary texts, respectively. A low divergence should then suggest that the findings presented in these summaries is in agreement. We will call this measure *findings-JSD*.

3 Results

Automated Evaluation

We report ROUGE-L scores with respect to the target (manually composed) Cochrane summaries, for both the development and test sets in Table 2. The methods perform about comparably with respect to this automatic metric. But ROUGE measures are based on (exact) n-gram overlap, and are insufficient for measuring the *factuality* of generated texts[21,42]. Indeed, we find that the summaries generated by all variants considered enjoy strong fluency, but the key question for this application is whether generated summaries are *factually correct*. Below we confirm via manual evaluation that despite achieving comparable ROUGE scores, these systems vary significantly with respect to the factuality of the summaries that they produce.

System variant	Relevance >2	Fluency >3	Factuality >3
XSUM	96	90	40
Pretrain (PMC)	98	97	34
Decorate	98	96	54
Sort by N · RoB	96	93	46
Decorate and sort	93	88	47

Table 3: Counts of generated summaries out of 100 assessed by MD 1 as exhibiting high relevance (3/3); good to very good fluency (>3/5); and moderate to strong factual agreement with reference summaries (>3/5).

Manual Evaluation

Manual annotation was performed for 100 reference Cochrane reviews and the 5 systems described above. Annotators were shown summaries generated by these systems in turn, *in random order*. Randomization was performed independently for each review (i.e., for each reference summary). Annotators did not know which system produced which summaries during assessment. We asked four questions across two pages about generated summaries.

The first page displayed only the generated summary, and asked annotators to appraise its *relevance* to the topic (the title of the corresponding systematic review) on a 3-point ordinal scale ranging from mostly off-topic to strongly on-topic. The second question on the first page concerned 'semantic plausibility', intended to measure whether the generated text is understandable, coherent, and free from self-contradictory statements. This assessment was on a five-point (Likert) scale.

Following these initial evaluations, annotators were asked two additional questions to assess the factuality of generated summaries with respect to the reference. The first concerned the direction of the reported finding in the reference summary (e.g., did the authors conclude the intervention being investigated beneficial compared with the control?). The second question then asked the annotator to assess the degree to which the generated summary agreed with the reference summary in terms of these key conclusions. Both of these judgments were collected on Likert-scales.

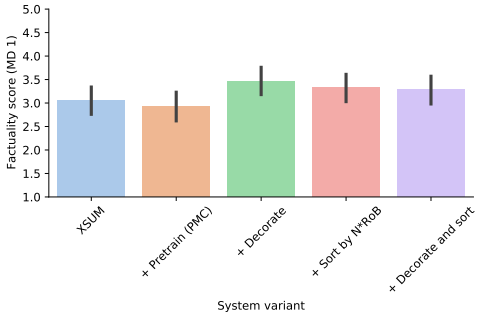
One author (IJM; a medical doctor with extensive experience in evidence-based medicine) performed the above evaluation on a 'pilot' set of 10 reference reviews, yielding 50 system judgements in all. He did not know which systems produced which outputs. This pilot set was used to assess agreement with additional prospective annotators, who we recruited via the Upwork platform[6].

We hired three candidate annotators (all with medical training) to assess the system summaries for pilot set of reviews appraised by IJM. Only one of these candidates provided reliable annotations, as determined by agreement with the reference set of assessments.[7] Scores provided by the successful annotator (who we will refer to as 'MD 1') achieved 0.535 linearly weighted κ with reference annotations concerning 'factuality', the hardest task, indicating reasonable agreement. IJM also subsequently evaluated all cases in this set where the label he had provided disagreed with MD 1's assessment (still blinded to which systems produced the corresponding summaries). These were determined reasonable disagreements, given that the task is somewhat subjective as currently framed.

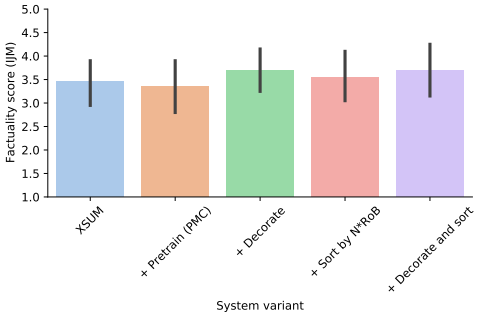
In total we collected assessments across the five system variants considered with respect to 100 unique corresponding Cochrane reference summaries from MD 1; for this we paid about $1,500 USD. As a second (post-hoc) quality check, IJM evaluated an additional randomly selected subset comprising 10 reference reviews. Agreement concerning relevance (80% exact agreement) and fluency (68% accuracy) remained high, as in the pilot round. However, in contrast to the pilot set, agreement concerning factuality on this subset was ostensibly poor (linearly weighted κ=0.04); on average IJM scored systems higher in factuality by 1.62 points. As above, IJM therefore evaluated all instances for which disagreement was ≥ 2 (again keeping blinding intact). This process again revealed predominantly reasonable subjective differences on this small set in assessing the level of agreement between the findings communicated in the generated and reference summaries, respectively. MD 1 consistently rated lower factuality scores than IJM — assigning lower numbers across the board — but relative rankings seem to broadly agree (Figure 2).

[6] http://www.upwork.com

[7] We assessed this both quantitatively and qualitatively. Rejected candidates provided uniformly high scores for all generated summaries, even in cases where, upon inspection, these were blatantly in disagreement with the reference summary.



(a) Scores from the MD hired via Upwork (MD 1) over 100 unique reference summaries (500 summary annotations).



(b) Scores from co-author (and MD) IJM over a subset of 20 unique reviews (100 summary annotations).

Figure 2: Factuality assessments performed by an individual with medical training for five systems over 100 unique reference summaries from the `dev` set (a), and by co-author and MD IJM over a small subset of twenty of these (b). All strategies except 'XSUM' start from the model checkpoint after PMC pretraining. We first evaluate the 'decoration' and sorting strategies (Section 2) independently, and then in combination; system names are consistent with Table 2.

This disagreement suggests that in future work we should work to improve the annotation task framing and guidance. The most common disagreement occurred in cases where the reference summary described a lack of reliable evidence on a topic, but *hinted* cautiously that there was some small, or low quality evidence in favor of an intervention. If an automated summary only described a lack of reliable evidence on the topic, it was ambiguous whether the overall poor state of evidence should be scored (in this instance, showing perfect agreement), or by how much the automated summary should be penalized for missing the tentative conclusion of possible benefit.

Nonetheless, in light of strong agreement on *other* rated aspects and our manual assessments of all substantial disagreements, we feel reasonably confident that MD 1 provided meaningful scores, despite the low quantitative measure of agreement on the second randomly selected set. And regardless, the broad trends across systems agree when the annotations from the two annotators are analyzed independently (Figure 2).

All systems received consistently high relevance scores from MD 1 (mean scores for system summaries produced by different systems over the 100 reviews range from 2.73 to 2.79, out of 3), and 'semantic plausibility' scores (range: 4.47 to 4.64 across systems, out of 5). Table 3 reports the counts of 'good quality' summaries with respect to the aforementioned aspects, as judged by MD 1. We can see that systems struggle to produce factual summaries.

Figure 2 (a) reports the mean factuality scores provided by MD 1 for the respective model variants. The proposed 'decorating' strategy yields a statistically significant improvement over the baseline PMC pretraining strategy (2.92 vs 3.46; $p \approx 0.001$ under a paired t-test). Note that this is the appropriate comparison because the '+ Decorate' model starts from the PMC pretrained checkpoint. Sorting inputs such that the encoder prioritizes including abstracts that describe large, high-quality studies (given the 1024 token budget imposed by BART) also increases factuality, albeit less so (2.92 vs 3.33; $p \approx 0.01$). Figure 2 (b) presents the factuality scores provided by IJM over a small subset of the data (20 unique reviews in all). The pattern is consistent with what we observed in MD 1's scores in that 'decoration' yields increased factuality (mean score of 3.35 vs 3.70).[8]

Automatically Assessing the Factuality of Evidence Synopses

ROUGE scores do not vary much across model variants, but this probably mostly reflects the fluency of summaries — which was also manually assessed to be uniformly good across systems. ROUGE (which is based on word overlap statistics) does not, however, seem to capture *factuality*, which is naturally of central importance for evidence synthesis. We tested this formally using annotations from MD 1: We regressed factuality judgements (ranging 1-5) on ROUGE-L scores (including an intercept term), over all annotated summaries. The trend is as we might expect: larger ROUGE-L scores are associated with better factuality ratings, but the correlation is not significant ($p \approx 0.18$).

We are therefore reliant on manual factuality assessments as we work to improve models. Performing such evaluations

[8]Though given the small sample of 20 reviews that IJM annotated neither difference is statistical significant when considering only these labels.

is expensive and time-consuming: Collecting annotations over 100 instances for this work cost nearly $2,000 USD (including payments to collect 'pilot' round annotations) and investing considerable time in documentation and training. Relying on manual assessments will therefore substantially slow progress on summarization models for evidence synthesis, motivating a need for automated factuality evaluation such as the *findings-JSD* measure proposed above.

Repeating the regression we performed for ROUGE-L, we can measure whether findings-JSD correlates with manual factuality assessments. We define a regression in which we predict factuality scores on the basis of the JSD scores. We find a statistically significant correlation between these with an estimated coefficient of -1.30 (95% CI: -1.79 to -0.81; $p < 0.01$), implying that the larger the disagreement concerning whether the summaries report a significant effect or not (measured using JSD), the lower the factuality score, as we might expect.

This result is promising. But despite the significant correlation this automated metric has with manual assessments, it is not strong enough to pick up on the differences between strategies. In particular, repeating the t-test on findings-JSD scores for the pretaining and decorating strategies yields a p-value of 0.40, i.e., the measure fails to meaningfully distinguish the latter from the former with respect to factuality. We conjecture that this is because while the measure significantly correlates with human assessments, it does so only modestly ($R^2 = 0.05$). We therefore conclude that this strategy constitutes a promising avenue for automatically assessing the factuality of generated summaries, but additional work is needed to define a measure that enjoys a stronger correlation with manual assessments.

4 Discussion

Above we proposed variants of modern neural summarizaiton models in which we: Perform additional in-domain pretraining (over the RCTs in PMC); 'decorate' inputs with automatically extracted information (e.g., population descriptions and evidence-bearing sentences); and sort inputs to prioritize passing along large and high-quality trials (given the limit on the length of the model input imposed by the transformer model we use).

We evaluated these models across key aspects, including relevance, 'semantic plausibility', and factuality. All systems we considered yielded highly fluent and relevant summaries. But manual analysis of generated and corresponding reference summaries revealed that the *factuality* of these systems remains an issue. The proposed decoration and sorting strategies both yielded modest but statistically significant improvements in assessed factuality.

Annotators exhibited disagreement when evaluating factuality. We believe this in part reflects the inherent difficulty of the task, but in future work we hope to improve the annotation protocol to reduce subjectivity and improve agreement. For example, being more explicit in the levels of disagreement that should map onto specific numerical scores and providing more detailed instructions regarding this may improve inter-rater agreement, as might explicitly differentiating between the factuality of *strength* of evidence and the reported directionality of the finding.

ROUGE scores — commonly used to automatically evaluate summarization systems — did not significantly correlate with factuality assessments here. We proposed a method for automatically evaluating the factuality of narrative evidence syntheses, findings-JSD, using models to infer the reported directionality of findings in generated and reference summaries. This measure significantly (though weakly) correlates with manual assessments of factuality. We view this as a promising direction to pursue going forward to facilitate automatic evaluation of evidence synopses, which in turn would support continued development of automated summarization systems for evidence synthesis.

5 Conclusions

We have demonstrated that modern neural abstractive summarization systems can generate relevant and fluent narrative summaries of RCT evidence, but struggle to produce summaries that accurately reflect the underlying evidence, i.e., that are *factual*. We proposed new approaches that modestly improve the factuality of system outputs, and described a metric that attempts (with some success) to automatically measure factuality, suggesting directions for future work. The multi-document summarization dataset is available: `https://github.com/bwallace/RCT-summarization-data`.

Acknowledgements

This work is funded by the National Institutes of Health (NIH) under the National Library of Medicine, grant R01-LM012086. We thank Ani Nenkova for helpful comments concerning evaluation.

References

[1] Sackett DL. Evidence-based medicine. In: Seminars in perinatology. vol. 21. Elsevier; 1997. p. 3–5.

[2] Gough D, Oliver S, Thomas J. An introduction to systematic reviews. Sage; 2017.

[3] Ryan G, Singh M, Dwan K. Inhaled antibiotics for long-term therapy in cystic fibrosis. Cochrane Database of Systematic Reviews. 2011;(3).

[4] Demner-Fushman D, Lin J. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics; 2006. p. 841–848.

[5] Mollá D. A corpus for evidence based medicine summarisation. In: Proceedings of the Australasian Language Technology Association Workshop; 2010. p. 76–80.

[6] Sarker A, Molla D, Paris C. Automated text summarisation and evidence-based medicine: A survey of two domains; 2017. .

[7] Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, et al. AskHERMES: An online question answering system for complex clinical questions. Journal of biomedical informatics. 2011;44(2):277–288.

[8] Nentidis A, Bougiatiotis K, Krithara A, Paliouras G. Results of the Seventh Edition of the BioASQ Challenge. In: ECML/KDD. Springer; 2019. p. 553–568.

[9] Mollá D, Santiago-Martínez ME, Sarker A, Paris C. A corpus for research in text processing for evidence based medicine. Language Resources and Evaluation. 2016;50(4):705–727.

[10] See A, Liu PJ, Manning CD. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:170404368. 2017;.

[11] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. preprint arXiv:191013461. 2019;.

[12] Zhang J, Zhao Y, Saleh M, Liu PJ. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. arXiv preprint arXiv:191208777. 2019;.

[13] Maybury M. Advances in automatic text summarization. MIT press; 1999.

[14] Nenkova A, McKeown K. Automatic summarization. Now Publishers Inc; 2011.

[15] Rush AM, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:150900685. 2015;.

[16] Lin H, Ng V. Abstractive summarization: A survey of the state of the art. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33; 2019. p. 9815–9822.

[17] Van Der Lee C, Gatt A, Van Miltenburg E, Wubben S, Krahmer E. Best practices for the human evaluation of automatically generated text. In: International Conference on Natural Language Generation; 2019. p. 355–368.

[18] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Text summarization branches out; 2004. .

[19] Reiter E. Accuracy Errors Go Beyond Getting Facts Wrong; 2020. Available from: `https://ehudreiter.com/2020/04/27/accuracy-errors-go-beyond-getting-facts-wrong/`.

[20] Reiter E, Belz A. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. Computational Linguistics. 2009;35(4):529–558.

[21] Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. arXiv preprint arXiv:200500661. 2020;.

[22] Wang A, Cho K, Lewis M. Asking and answering questions to evaluate the factual consistency of summaries. arXiv preprint arXiv:200404228. 2020;.

[23] Xu X, Dušek O, Li J, Rieser V, Konstas I. Fact-based Content Weighting for Evaluating Abstractive Summarisation. In: Association for Computational Linguistics; 2020. p. 5071–5081.

[24] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018;.

[25] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008.

[26] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Advances in neural information processing systems; 2014. p. 3104–3112.

[27] Zhang M, Tan L, Tu Z, Fu Z, Xiong K, Li M, et al. To Paraphrase or Not To Paraphrase: User-Controllable Selective Paraphrase Generation. arXiv preprint arXiv:200809290. 2020;.

[28] Rosset C, Xiong C, Phan M, Song X, Bennett P, Tiwary S. Knowledge-Aware Language Model Pretraining. arXiv preprint arXiv:200700655. 2020;.

[29] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv. 2019;p. arXiv–1910.

[30] Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: ICCV; 2015. p. 19–27.

[31] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019;.

[32] Trinh TH, Le QV. A simple method for commonsense reasoning. arXiv preprint arXiv:180602847. 2018;.

[33] Narayan S, Cohen SB, Lapata M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. arXiv preprint arXiv:180808745. 2018;.

[34] Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. Research synthesis methods. 2018;9(4):602–614.

[35] Nye B, Li JJ, Patel R, Yang Y, Marshall IJ, Nenkova A, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In: Association for Computational Linguistics (ACL). vol. 2018; 2018. p. 197.

[36] Lehman E, DeYoung J, Barzilay R, Wallace BC. Inferring which medical treatments work from reports of clinical trials. arXiv preprint arXiv:190401606. 2019;.

[37] Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. Journal of the American Medical Informatics Association. 2016;23(1):193–201.

[38] Marshall IJ, Kuiper J, Banner E, Wallace BC. Automating biomedical evidence synthesis: RobotReviewer. In: Association for Computational Linguistics (ACL). vol. 2017; 2017. p. 7.

[39] Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proceedings of ACL; 2020. .

[40] DeYoung J, Lehman E, Nye B, Marshall IJ, Wallace BC. Evidence Inference 2.0: More Data, Better Models. arXiv preprint arXiv:200504177. 2020;.

[41] Narayan S, Cohen SB, Lapata M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In: EMNLP; 2018. p. 1797–1807.

[42] Kryściński W, Keskar NS, McCann B, Xiong C, Socher R. Neural text summarization: A critical evaluation. arXiv preprint arXiv:190808960. 2019;.

ResultsMyWay: combining Fast Healthcare Interoperability Resources (FHIR), Clinical Quality Language (CQL), and informational resources to create a newborn screening application

Michael Watkins[1], Alex Au, MD[1], Truc Vuong[1], Heidi Wallis[2], Kim Hart, MS[2], Andy Rohrwasser, Ph.D., MBA[3], Karen Eilbeck, Ph.D.[1]
[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah;
[2]Utah Department of Health, Salt Lake City, Utah;
[3]Utah Public Health Laboratory, Taylorsville, Utah

Abstract

Newborn screening (NBS) can be life-changing for the families of infants who test positive for a rare condition. While resources exist to support these families, there can be delays in sharing these resources due to communication lag between the laboratory, result interpreting clinician, family of the newborn, and additional care providers. This delay can also be exacerbated when additional health history is required from the mother and infant. ResultsMyWay is a proof-of-concept application that uses Clinical Quality Language (CQL) to automate the search for this additional health history. It also translates the NBS results into Fast Healthcare Interoperability Resources (FHIR), increasing both the ease of exchange and the future utility of these data points. After the families are given the NBS results, ResultsMyWay then acts as a hub for several types of informational resources about the recently diagnosed condition.

Introduction

Newborn screening (NBS) is a public health program that screens infants shortly after birth for diseases and disorders that can severely harm the infant if early detection and care management are not provided. Of the approximately four million infants born per year in the United States, about 12,900 of them are diagnosed with a disorder via the NBS process[1].

In the event of a positive rare condition result, the guardians of the infant are contacted to inform them of the result and to provide the assurance that qualified professionals will be in contact to help establish an appropriate care plan for the child[2]. Difficulties in the delivery of information to the care providers and the guardians can lead to problems such as loss to follow-up[3], distress[4], confusion and strong emotional reactions[2]. Parental knowledge of the condition varies[4] and the knowledge gap is another pain point that must be addressed. The internet provides an opportunity for parental education about NBS screening with health department web pages[5], and there are resources such as Baby's First Test[6] that aim to collate information. But a Google search of 'nbs cystic fibrosis test result' brings back 142,000 pages which is obviously overwhelming.

Although NBS is a routine clinical service, there are still improvements that can be made in the typical NBS workflow. For example, there is the need, in many cases, to manually collect additional health information about the mother and the infant for cases where the interpretation of the results could change depending on the presence or absence of these additional factors. This collection is a burden to the interpreting clinician who now must wait for what can be a slow health information exchange process, to the guardian who now knows that something isn't right but must continue to wait until verified results are returned, and for the care provider of the mother who must search through that mother's health records as well as the infant's health history (which may include several previous clinical care providers who need to be contacted). When the results are finalized, they are typically compiled into a static portable document format (PDF) report. Representing clinical data within a PDF locks the data out of the health system and renders it only human-readable. Storing the data as computable data points however, would increase the likelihood of the data being used to enhance the care of the patient in future encounters[7].

Background

FHIR

Fast Healthcare Interoperability Resources (FHIR) is a rapidly-growing Health Level Seven International (HL7) standard that represents specific data artifacts found in typical clinical scenarios (Patient, Condition, Encounter, Observation, etc.) as standardized "resources". These resources are linked together via identifiers to provide a standardized and robust representation of clinical events[8]. The resources themselves can be thought of as both the

content model (how the thing is represented) and the artifact (the thing that is exchanged). They are exchanged via representational state transfer and can be tailored to specific needs using a strict system of published profiles and extensions. There are currently 145 FHIR resources, each at one of seven different levels of maturity ranging from 0 (draft), 1-5 (multiple aspects of "completion" considered), to N (normative)[9]. Resources and profiles are developed, balloted, and published by ~40 HL7 working groups[10]. These working groups are chaired by and made up of top health professionals from a variety of medical backgrounds worldwide.

CQL

Clinical Quality Language (CQL) is an HL7 specification for defining standardized and shareable clinical logic. CQL authoring typically involves the creation of several distinct expressions. These expressions can query clinical artifacts and employ a wide variety of operators and filters before returning a result. A collection of CQL expressions is called a Library. CQL was designed for use within both the clinical decision support and clinical quality measurement domains[11]. It has been predominantly developed and used by stakeholders in the HL7 Da Vinci Project, whose goal is to help payers and providers to positively impact clinical, quality, cost, and care management outcomes[12]. They have piloted solutions that combine CQL with the HL7 CDS Hooks[13] specification and HL7 FHIR Questionnaire resource in reducing burden within the specific contexts of Coverage Requirements Discovery[14], Documentation Templates and Rules[15], and Prior-Authorization Support[16]. These implementations use CQL to search the patient record and determine if that individual has the required pre-existing conditions to qualify for coverage of some type. NBS has a similar need in that results can occasionally be interpreted differently if certain pre-existing conditions apply to the infant or mother. ResultsMyWay pioneers the use of CQL in aiding clinical laboratory test interpretation.

Methods

While ResultsMyWay is a proof-of-concept application and not yet implemented live in the Utah NBS process, it was designed with interoperability and future implementation in mind. Therefore, several real-world scenarios were considered in the architecture design.

Account Creation

The first step was to ensure that a user can be uniquely matched to an infant in all foreseeable cases[17] (e.g., adoptive parents or other forms of non-maternal guardianship). It is anticipated that a ResultsMyWay activation code will be given to the guardian at the time of specimen collection. This will allow that guardian to match to the infant (who will have that code added as an identifier to their Patient resource) and create an account to steward the NBS process for that infant. Additional data collected during account activation include name, relationship to the newborn, email address, phone number, an option for designating a call or text when results are ready to be viewed in the application, and the option to add a secondary user on behalf of the infant. These data are used to create a FHIR RelatedPerson resource which ResultsMyWay uses to manage users.

Automated Retrieval of Factors that Impact NBS Interpretation

The Utah NBS data model includes three LOINC panels that can be drawn upon in the interpretation of certain NBS results: Feeding type (67704-7), Infant factors that affect newborn screening interpretation (57713-0), and Maternal factors that affect newborn screening interpretation (67707-0). The health factors covered by these panels range include procedures (blood transfusion, thoracic surgery involving thymectomy, etc.), conditions (liver disease, biliary atresia, septicemia, etc.), and medications (parenteral steroids, systemic antibiotics, dopamine, etc.). As part of this application, the answer lists for each of these panels were converted to CQL expressions. These expressions rely on custom FHIR ValueSets that contain synonym codes from several terminology systems (e.g., LOINC, SnomedCT, ICD-10, RxNorm) for each Condition, Observation, Procedure, or MedicationRequest required to retrieve these additional factors. The first CQL expression in Figure 1 shows a simple retrieval of all MedicationRequest resources with an RxNorm code contained in the custom ValueSet called "Dopamine Set". This expression will return the ID of any existing Dopamine medication orders (stored as FHIR MedicationRequest resources) within the infant's health record. Single direct reference codes can also be declared and used as shown by the second expression in Figure 1 which retrieves all Condition resources for the infant and filters for any whose "severity" is acute. These expressions can also return structured objects as shown in the third expression in Figure 1 where the "performed" date of the Procedure resource (representing a blood transfusion) is returned along with the ID of the resource. CQL Libraries can be stored, exchanged, and used via the FHIR Library resource[18]. A FHIR Library resource contains a base-64 encoded string representing the CQL Library, references to terminology requirements, and various other fields. The CQL Library can then be executed within the data server by sending a request including the $evaluate operation and several context parameters. The server then returns the results of the CQL Library evaluation as shown in Figure 2. In

following the example of the implementation guides published by the Da Vinci project, ResultsMyWay uses FHIR Questionnaire resources as companions to the Library resources. These questionnaires display the CQL retrieval results and capture additional needed data that were not found in the patient's medical record. These Questionnaire resources are constrained by profiles that have been created by the HL7 Structured Data Capture workgroup[19] to facilitate their use alongside CQL Libraries with each question having an optional extension field containing the name of the matching CQL expression. The application then extracts the CQL response to auto-fill that corresponding question for the user.

```
codesystem "SNOMED":'http://snomed.info/sct'
valueset "Dopamine Set":'http://test-url.com/fhir/ValueSet/dopamine'
valueset "Blood Transfusion Set":'http://test-url.com/fhir/ValueSet/blood-transfusion'
code "Acute Condition Code":'24484000' from "SNOMED" display 'Severe'

define "Dopamine":
    [MedicationRequest: "Dopamine Set"] M
    return M.id

define "Acute illness":
    [Condition] C
    where exists(
        C.severity.coding Coding
        where FHIRHelpers.ToCode(Coding) ~ "Acute Condition Code"
    )
    return C.id

define "The date of last blood product transfusion":
    [Procedure : "Blood Transfusion Set"] P
    return {
        "ID": P.id,
        "Date": P.performed
    }
```

Figure 1. Selections from a CQL Library used by ResultsMyWay.

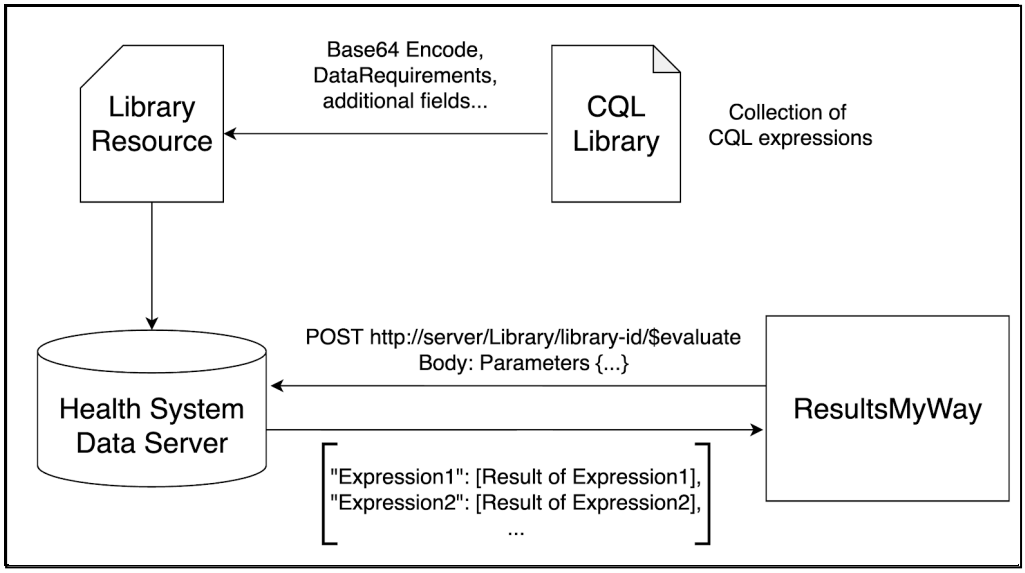


Figure 2. A visual representation of how a CQL Library can be stored and executed on a FHIR server.

NBS Results Message Mapping

In the Utah NBS system, results are formatted in the fields of an HL7 Version 2.4 ORU-R01 message before being rendered as PDF files. One aim of this work was to provide the Utah NBS Program with mappings from HL7 v2.4 to FHIR v4.0.0. This effort to map from HL7 Version 2.4 to FHIR has a dedicated HL7 working group whose findings informed the mappings used for ResultsMyWay[20]. A selection from the final mappings used for ResultsMyWay is shown in Table 1.

Table 1. A selection from the final mappings used for ResultsMyWay.

V2 Field	V2 Field Description	Utah NBS Result Segment	FHIR Resource.Field
ORC-3	Filler Order Number	F1590009202099	DiagnosticReport.identifier
ORC-5	Order Status	CM	ServiceRequest.status
ORC-12	Ordering Provider	LW3166^DUFFY^ TIMOTHY^^^^^^HC^^^^^	Practitioner.name
OBR-4	Universal Service Identifier	54090-6^TSH Panel^LN	DiagnosticReport.code
OBR-14	Specimen Received Date/Time	202006040028	Specimen.receivedTime
OBR-15	Specimen Source	DBS&Dried blood spot card [79566-6]^^^&	Specimen.type
OBR-22	Results Rpt/Status Chng - Date/Time	202006041722	Specimen.processing[x].time DateTime
OBR-25	Result Status	F	DiagnosticReport.conclusion
OBR-27	Quantity/Timing	^^^^^U	DiagnosticReport.conclusion
OBR-32	Principal Result Interpreter	39&Wallis& Heidi	DiagnosticReport.resultsInterpreter
NTE-2	Source of Comment	L	DiagnosticReport.extension
NTE-3	Comment	This is a TSH Panel Rpt comment.	DiagnosticReport.extension
NTE-4	Comment Type	RE	DiagnosticReport.extension
OBX-2	Value Type	NM	(determines valueCoding OR valueQuantity)
OBX-3	Observation Identifier	29575-8^TSH^LN	Observation.code
OBX-5	Observation Value	62	Observation.valueCoding OR valueQuantity.value
OBX-6	Units	uIU/mL	Observation.valueQuantity.units
OBX-7	Reference Range	0-40	Observation.referenceRange
OBX-8	Interpretation Codes	H	Observation.interpretation
OBX-11	Observation Result Status	F	Observation.note
OBX-14	Date/Time of the Observation	20200601	Observation.effective

Information Resource Content Management

One of the primary goals of ResultsMyWay is to act as a hub for educational, community outreach, and clinical trial resources that relate to an infant's condition. These resources are constantly changing and it is expected that a content manager within each implementing system would be assigned to monitor current content associated with each condition as well as approving new user-suggested resources via a submission system in the application. To ease this management, a simple schema was designed to represent each condition and its associated resources. A sample from this schema is shown in Figure 3.


```
"condition":"pku",
"library":{
    "basic":[…],
    "nutrition":[…],
    "research":[…],
    "media": [
        {
            "title":"Utah Newborn Screening",
            "type":"Feed",
            "image":"ut-nbs.png",
            "preview":"true",
            "url":"https://newbornscreening.health.utah.gov/news/"
        },
        {…},{…},{…},{…},{…}
    ]
},
"community":[
    {
        "title":"National PKU Alliance",
        "image":"npkua.png",
        "facebook":"https://www.facebook.com/NationalPKUAlliance",
        "instagram":"https://www.instagram.com/national_pku_alliance/",
        "twitter":"https://twitter.com/NPKUA_Info",
        "website":"https://www.npkua.org/"
    },{…},{…},{…},{…}
]
```


Figure 3. The resource content for the Phenylketonuria (PKU) condition.

Results

A demo of ResultsMyWay has been created to demonstrate the utility of the implementation (http://hematite.genetics.utah.edu/ResultsMyWay/).

Results Page

ResultsMyWay is meant to be used both before and after an infant's results are returned. Users are able to create accounts as soon as the original specimen is taken and the official process begins. Results can take days to finalize and, in the interim, ResultsMyWay provides useful general information about NBS for users and acts as a valuable data collection tool for providers. Consider the following use case: An expecting 37-year-old mother at 33 weeks

gestation with known hypothyroidism is being observed in the hospital with sudden onset nausea, vomiting, and severe headache. She is shortly diagnosed with HELLP syndrome and started on fluid resuscitation, 1 unit of packed red blood cell transfusion, and intravenous steroids. She is stabilized and prepared for immediate delivery. She gives birth to a male infant who is found to have low birth weight, significant hypoxia, and lethargic at delivery. He responds to supplemental oxygen and is transferred to the neonatal intensive care unit (NICU) where a blood test is collected. Lab results showed low thyroxine in the newborn as well as anemia. The newborn is blood typed and started on a blood transfusion while in the NICU. A sample is also collected for NBS. While this can be considered an atypical birth experience, it highlights the value of the automated retrieval of factors that could influence the interpretation of the NBS results for this infant and is shown in Figure 4.

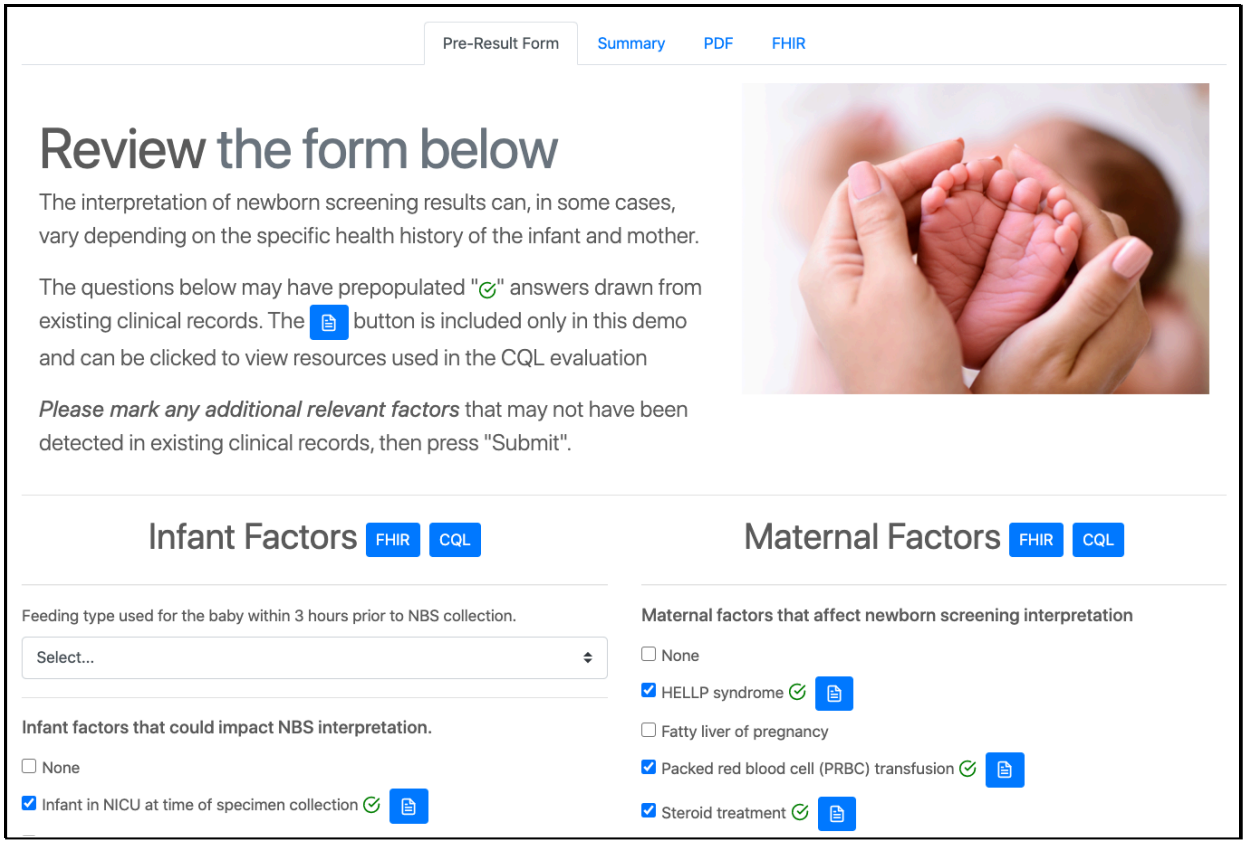


Figure 4. The initial view of the results page--showing the CQL results for the use case given above.

These questionnaire forms can be added to for clinical data that perhaps weren't yet recorded or that were missed by the CQL library. It is anticipated that a FHIR QuestionnaireResponse resource, which contains all of these responses and is generated when the user completes the form and clicks "Submit," would be sent to the appropriate result-interpreting clinician or lab technician to assist them in the interpretation of the infant's NBS results.

Once the results are finalized and ready to be communicated back to the patient, ResultsMyWay contains three options for viewing and exploring those results. The first is the "Summary" view which gives the results for each panel while visualizing relevant data ranges and providing links to explanatory resources for panels, terms, codes, and comment sources and types. This represents a significant amount of metadata that can now provide added understanding and perspective for concerned guardians. This first view is shown in Figure 5. The second view is a "PDF" view which displays the classic PDF report which is stored as a base64 string in each FHIR DiagnosticReport resource. This allows for a lossless transition from the current process to this updated process. This is intended to improve the change management for care providers who may have a wide spectrum of attitudes toward new clinical innovations. The third view allows the relevant FHIR resources to be reviewed and has been initially included in the application for the benefit of clinicians who may want to review the actual raw data artifacts as they appear in the health system.

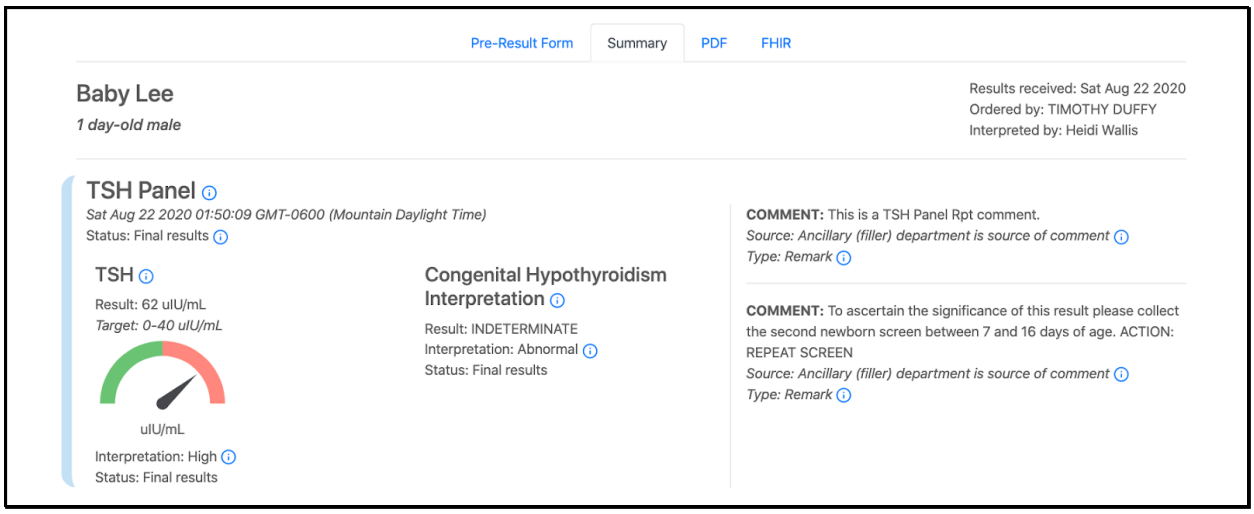


Figure 5. The summary view of the ResultsMyWay results page--showing the results of one NBS panel.

Information Resource Pages

After the results have been communicated and the users are given access to four pages which contain several different kinds of resources specific to that infant and condition.

The first is a knowledge library. This is a collection of websites, factsheets, feeds, videos, etc. that have been separated into four categories: Basic Knowledge (resources designed for all levels of background education and experience), Nutrition (resources specifically involving potential food and lifestyle adjustments), Research (a custom PubMed RSS feed), and Media (spotlights or other news media attention for the condition or others dealing with that condition). Users may suggest additional resources to be added. This knowledge library is shown in Figure 6.

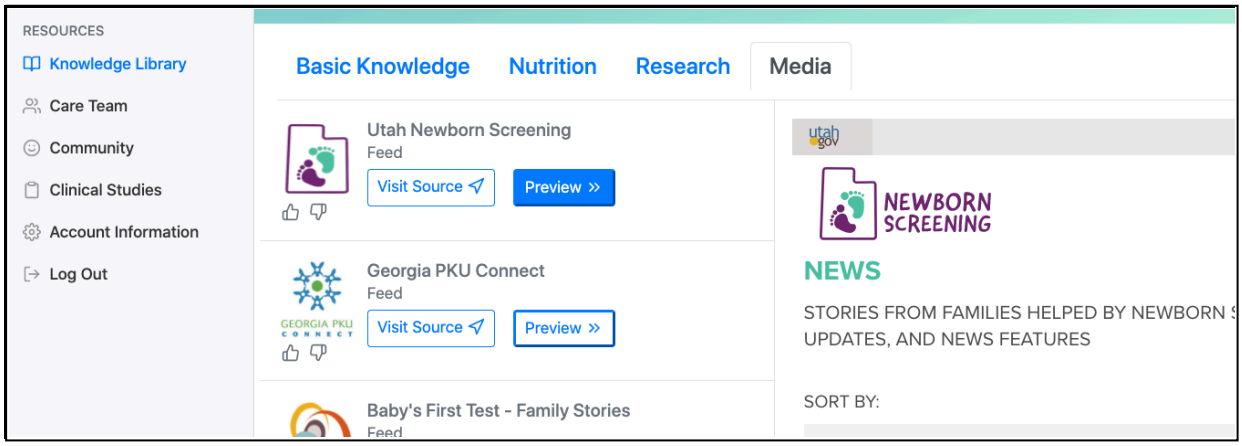


Figure 6. The knowledge library page of ResultsMyWay--showing media results for the PKU condition.

The second page uses FHIR to assemble profiles for each of the significant members of that infant's care team. With a picture (when available) and phone number, this page provides a convenient reference to the user who may be managing communication between several clinicians.

The third page provides a list of various social media or other community groups that have been organized to support those living with that certain condition and is shown in Figure 7. The fourth page shows a custom ClinicalTrials.gov RSS feed that users can scroll through if they are interested in treatment options of that kind. This would likely increase participation in these trials and be a valuable resource to the research community.

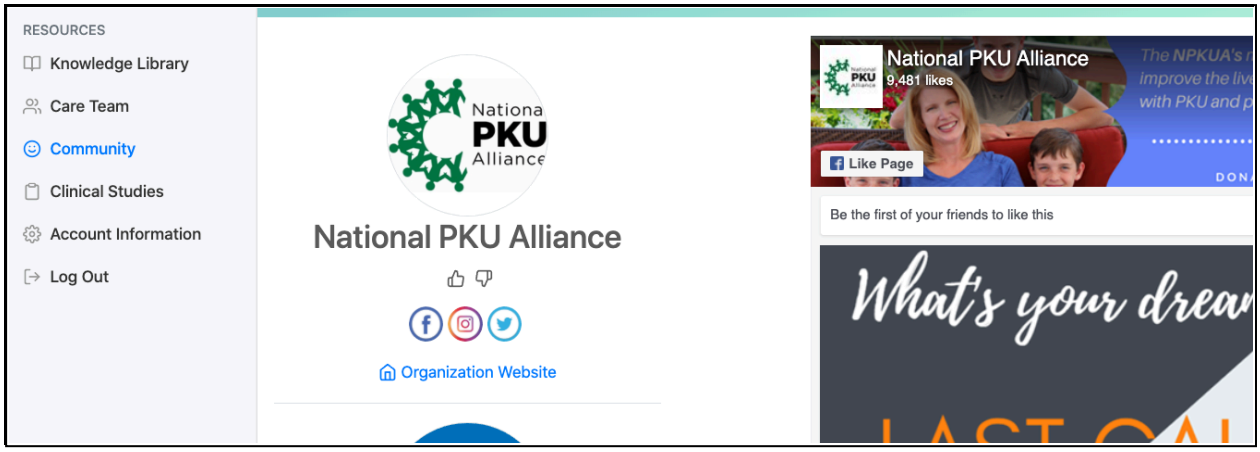


Figure 7. The community page of ResultsMyWay—showing links to the social media pages of NPKU.

Discussion

In this demonstration, we have used CQL to automate the retrieval of health factors which could influence the interpretation of NBS results. Automation would reduce the time and effort involved in the necessary health information exchange and could lead to earlier diagnoses of potential diseases and disorders. This could be critical for cases where early care management significantly decreases the likelihood of adverse effects. While CQL is an innovative and burden-reducing technology, manual input of clinical data points by users can be a dangerous option[21]. The health literacy of the user and the reliability of data entered would be unknown and could have harmful effects. Another option would be to restrict changing or adding answers to the ResultsMyWay pre-results form to clinicians only. This would increase the burden on the clinician who retrieves the NBS blood sample because they would need to take time to review the medical history of the patient in addition to educating the guardian about the app itself.

While FHIR is an interoperable and effective way to represent NBS results, this application also requires a significant amount of FHIR data server transactions. This could become an expensive burden to the implementing health system in situations where these transactions are meant to be employed with more thrift. However, these considerations were out of scope for this proof-of-concept application. Another critical consideration is the security of protected health information. While the SMART specification[22] can ensure the secure launching of the app and the security of the data for authorized users, the user onboarding process will likely need additional measures to verify identity and relation to the newborn.

Conclusion

NBS is a routine clinical process that can result in complete lifestyle readjustment for the families of infants who test positive for a rare condition. ResultsMyWay has been designed to empower the guardians of these infants by providing informational resources and to assist the clinicians involved by streamlining the interpretation and exchange of NBS results.

Acknowledgments

Funding for this work has been provided by the Centers for Medicare & Medicaid Services as part of a Health Information Technology project and by the National Library of Medicine through Grant Number T15LM007124 to MW.

References

1. Sontag MK, Yusuf C, Grosse SD, et al. Infants with Congenital Disorders Identified Through Newborn Screening — United States, 2015–2017. MMWR Morb Mortal Wkly Rep 2020;69:1265–1268. Available from: http://dx.doi.org/10.15585/mmwr.mm6936a6
2. Salm N, Yetter E, Tluczek A. Informing parents about positive newborn screen results: parents' recommendations. J Child Health Care. 2012;16(4):367-381. DOI:10.1177/1367493512443906

3. Hunter LL, Meinzen-Derr J, Wiley S, Horvath CL, Kothari R, Wexelblatt S. Influence of the WIC Program on Loss to Follow-up for Newborn Hearing Screening. Pediatrics. 2016;138(1):e20154301. DOI:10.1542/peds.2015-4301
4. Chudleigh J, Buckingham S, Dignan J, et al. Parents' Experiences of Receiving the Initial Positive Newborn Screening (NBS) Result for Cystic Fibrosis and Sickle Cell Disease. J Genet Couns. 2016;25(6):1215-1226. DOI:10.1007/s10897-016-9959-4
5. Araia MH, Potter BK. Newborn screening education on the internet: a content analysis of North American newborn screening program websites. J Community Genet. 2011;2(3):127-134. DOI:10.1007/s12687-011-0046-0
6. Baby's First Test. What is Newborn Screening? [Internet] Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services (HHS); 2020. Available from: https://www.babysfirsttest.org/
7. Watkins M, Eilbeck K. FHIR Lab Reports: using SMART on FHIR and CDS Hooks to increase the clinical utility of pharmacogenomic laboratory test results. AMIA Jt Summits Transl Sci Proc. 2020;2020:683-692. Available from https://pubmed.ncbi.nlm.nih.gov/32477691/
8. HL7 International. Overview - FHIR v4.0.1 [Internet]. 2019 Nov 1 [cited 2020 Aug 27]. Available from: https://www.hl7.org/fhir/overview.html
9. HL7 International. Change Management and Versioning [Internet]. 2019 Nov 1 [cited 2020 Aug 27]. Available from: https://www.hl7.org/fhir/versions.html#maturity
10. Confluence. HL7 Work Groups & Projects [Internet]. 2020 Dec 1 [cited 2020 Aug 27]. Available from: https://confluence.hl7.org/pages/viewpage.action?pageId=4489802
11. HL7 International. Clinical Quality Language Release 1 STU 4 (1.4) [Internet]. 2019 Jun 23 [cited 2020 Aug 27]. Available from: https://cql.hl7.org/
12. HL7 International. Da Vinci Project [Internet]. 2020 [cited 2020 Aug 27]. Available from: https://www.hl7.org/about/davinci/
13. HL7 International. CDS Hooks [Internet]. 2019 [cited 2020 Aug 27]. Available from: https://cds-hooks.hl7.org/
14. HL7 International. Da Vinci Coverage Requirements Discovery (CRD) FHIR IG (v0.3.0: STU 1 Ballot 2) [Internet]. 2019 Mar 27 [cited 2020 Aug 27]. Available from: http://hl7.org/fhir/us/davinci-crd/2019May/
15. HL7 International. Da Vinci Documentation Templates and Rules (DTR) FHIR IG (v0.1.0: STU 1 Ballot 1) [Internet]. 2019 Mar 27 [cited 2020 Aug 27]. Available from: http://hl7.org/fhir/us/davinci-dtr/2019May/
16. HL7 International. Da Vinci Prior Authorization Support (PAS) FHIR IG (v0.1.0: STU 1 Ballot 1) [Internet]. 2019 Aug 5 [cited 2020 Aug 27]. Available from: http://hl7.org/fhir/us/davinci-pas/2019Sep/
17. Duncan J, Narus SP, Clyde S, Eilbeck K, Thornton S, Staes C. Birth of identity: understanding changes to birth certificates and their value for identity resolution. J Am Med Inform Assoc. 2015;22(e1):e120-e129. DOI:10.1136/amiajnl-2014-002774
18. HL7 International. Library - FHIR v4.0.1 [Internet]. 2019 Nov 1 [cited 2020 Aug 27]. Available from: https://www.hl7.org/fhir/library.html
19. HL7 International. SDC Home Page [Internet]. 2019 Mar 29 [updated 2020 Nov 13; cited 2020 Aug 27]. Available from: http://build.fhir.org/ig/HL7/sdc/
20. Confluence. 2-To-FHIR Project [Internet]. 2019 [updated 2020 Oct 26; cited 2020 Aug 27]. Available from: https://confluence.hl7.org/display/OO/2-To-FHIR+Project
21. Wuerdeman L, Volk L, Pizziferri L, et al. How accurate is information that patients contribute to their Electronic Health Record?. AMIA Annu Symp Proc. 2005;2005:834-838. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560697/
22. Computational Health Informatics Program, Boston Children's Hospital. SMART Health IT [Internet]. 2019 [cited 2020 Aug 27]. Available from: https://docs.smarthealthit.org/

CQL4NLP: Development and Integration of FHIR NLP Extensions in Clinical Quality Language for EHR-driven Phenotyping

Andrew Wen, MS[1], Luke V. Rasmussen, MS[2], Daniel Stone, BS[1], Sijia Liu, PhD[1], Rick Kiefer[1], Prakash Adekkanattu[3], Pascal S. Brandt[4], Jennifer A. Pacheco, MS[2], Yuan Luo, PhD[2], Fei Wang, PhD[3], Jyotishman Pathak, PhD[3], Hongfang Liu, PhD[1], Guoqian Jiang, MD, PhD[1]

[1]Mayo Clinic, Rochester, MN; [2]Northwestern University, Chicago, IL; [3]Weill Cornell Medicine, New York, NY; [4]University of Washington, Seattle, WA.

Abstract

Lack of standardized representation of natural language processing (NLP) components in phenotyping algorithms hinders portability of the phenotyping algorithms and their execution in a high-throughput and reproducible manner. The objective of the study is to develop and evaluate a standard-driven approach - CQL4NLP - that integrates a collection of NLP extensions represented in the HL7 Fast Healthcare Interoperability Resources (FHIR) standard into the clinical quality language (CQL). A minimal NLP data model with 11 NLP-specific data elements was created, including six FHIR NLP extensions. All 11 data elements were identified from their usage in real-world phenotyping algorithms. An NLP ruleset generation mechanism was integrated into the NLP2FHIR pipeline and the NLP rulesets enabled comparable performance for a case study with the identification of obesity comorbidities. The NLP ruleset generation mechanism created a reproducible process for defining the NLP components of a phenotyping algorithm and its execution.

Introduction

Unstructured data in electronic health records (EHRs) has been increasingly recognized as an important source for enabling accurate phenotyping[1]. Natural language processing (NLP) tools have been widely used for the purpose of phenotype identification and extraction from clinical narratives. For examples, Gundlapalli, et al.[2] developed algorithms to improve efficiency of psychosocial phenotyping using NLP with a large corpus of clinical text. Y, et al.[3] used NLP to reveal the occurrence patterns of medical concepts in EHR narrative notes and developed algorithms to identify patients with rheumatoid arthritis and coronary artery disease cases. Zhang et al.[4] developed a PheCAP system that leverages both structured and unstructured EHR data for high-throughput phenotyping. Liu et al.[5] compared four NLP systems for extracting generic phenotypic concepts, and demonstrated that ensembles of NLP can improve both generic phenotypic concept recognition and patient specific phenotypic concept identification over individual systems. However, lack of standardized representation of the NLP components in phenotyping algorithms hinders the portability of the phenotyping algorithms and their execution in a high-throughput and reproducible manner.

In recent years, there has been increasing calls to enable standardization and portability of NLP algorithms and approaches for clinical phenotyping. Rasmussen, et al[6] proposed a generalizable framework for phenotype algorithms portability and discussed the common data model (CDM)-based approach and their interaction with NLP-related tasks. Sharma et al.[7] developed a portable NLP-based phenotyping system that leverages the OMOP CDM and demonstrated that the system of standardization enables a consistent application of numerous rule-based and machine learning based classification techniques downstream across disparate datasets. Savova, et al.[8] developed a DeepPhe NLP system for extracting cancer phenotypes from clinical records and created a cancer phenotype information model[9] that leverages HL7 Fast Healthcare Interoperability Resources (FHIR) standard. Hong, et al.[10] developed a FHIR-based EHR phenotyping framework with an implementation of a FHIR-based clinical data normalization pipeline known as NLP2FHIR[11], and demonstrated that the developed approach could improve the data aspect of phenotyping portability across EHR systems and enhance interpretability of the machine learning-based phenotyping algorithms.

Clinical quality language (CQL)[12] is a HL7 standard to allow expressing logic that is both human readable and machine processable. It is part of the effort to harmonize standards used for electronic clinical quality measures (eCQMs) and clinical decision support (CDS). As of 2019, CQL has been adopted by CMS[13] to represent expression logic in the eCQMs to simplify the representation and reusability of the logic used in national quality measure programs. One of the key features is that CQL makes logic expressions independent of any specific data models.

Such data models include but are not limited to the Quality Data Model (QDM) and FHIR. As the creation of HER-driven phenotype algorithms shares many common requirements with the definition of clinical quality measures/clinical decision support rules, there are emerging interests in the clinical research informatics communities to explore use of CQL as a tool for the standard representation and execution of phenotype algorithms (i.e., structured selection criteria designed to produce research-quality phenotypes)[14].

The objective of this study is to develop and evaluate a standard-driven approach known as CQL4NLP that integrates a collection of NLP extensions represented in the HL7 FHIR standard into CQL to support next generation EHR-driven phenotyping.

Materials and Methods Materials

NLP2FHIR Pipeline. NLP2FHIR[11] is a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured EHR data. The underlying clinical NLP engines are based on cTAKES[15], MedXN[16], and MedTime[17]. NLP2FHIR currently supports extracting clinical data objects that can be rendered into the FHIR resources of Condition, Procedure, MedicationStatement (including Medications), and FamilyMemberHistory. In this study, we will explore the integration of the FHIR extensions into the NLP2FHIR pipeline.

Obesity Discharge Summary Datasets. We will leverage the following two obesity discharge summary datasets for evaluation design as we used in a previous study[10]. The first one is the i2b2 Obesity Challenge dataset which contains a total of 1237 annotated discharge summaries. The second one is the obesity discharge summaries extracted from MIMIC-III dataset, which contains 2000 randomly selected discharge summaries. For each discharge summary, the status of patients' obesity and fifteen comorbidities are annotated using a simple classification scheme: present (Y), or absent (N). The fifteen obesity comorbidities are: asthma, atherosclerotic cardiovascular disease (CAD), congestive heart failure (CHF), depression, diabetes mellitus (DM), gallstones/cholecystectomy, gastroesophageal reflux disease (GERD), gout, hypercholesterolemia, hypertension (HTN), hypertriglyceridemia, obstructive sleep apnea (OSA), osteoarthritis (OA), peripheral vascular disease (PVD), and venous insufficiency. In this study, we have selected a subset of the comorbidities for evaluation design.

PheKB Phenotype Algorithms. The Phenotype KnowledgeBase[18] (PheKB, http://phekb.org) is an online environment supporting the workflow of building, sharing, and validating electronic phenotype algorithms. Most of the phenotype algorithms in PheKB are publicly accessible and can be searched by data modalities or methods used such as ICD and CPT codes, Laboratories, Medications, Vital Signs, and NLP. In this study, we retrieved all publicly accessible PheKB phenotype algorithms as of March 9, 2020 that were annotated as "Natural language processing" (n=42). This list was accessed via the PheKB filtering URL: https://phekb.org/phenotype/methodsmodalities/natural-language-processing. For each phenotype, we downloaded all publicly accessible files linked on the PheKB page, as well as any referenced publications.

Methods

Creation of FHIR NLP Extensions. We first analyzed the NLP-specific data elements collected from different sources, including 1) OMOP NOTE_NLP table[19] that encodes all output of NLP on clinical notes; 2) cTAKES[15] and ConTEXT[20] output data models; 3) NLP2FHIR FHIR NLP extensions[11]; 4) PhEMA QDM NLP extensions, and 5) FHIR document resource Composition[21]. Table 1 shows an initial input of the NLP-specific data elements from different sources. And then we consolidated the identified data elements and created a minimal NLP data model with our project team consensus. This NLP data model includes a collection of FHIR NLP extensions with the FHIR Composition resource to enable the phenotyping algorithm authoring.

Table 1. An initial input of the NLP-specific data elements from different sources.

NLP-specific Data Elements	Reference Sources	NLP-specific Data Elements	Reference Sources
note_nlp_id	OMOP NOTE_NLP Table	Composition.category	FHIR Document Resource
note_id	OMOP NOTE_NLP Table	Composition.code	FHIR Document Resource
section_concept_id	OMOP NOTE_NLP Table	Composition.section.code	FHIR Document Resource
snippet	OMOP NOTE_NLP Table	Composition.entry.Condition	FHIR Document Resource
offset	OMOP NOTE_NLP Table	Composition.entry.Procedure	FHIR Document Resource
lexical_variant	OMOP NOTE_NLP Table	Composition.entry.MedicationStateme	FHIR Document Resource
note_nlp_concept_id	OMOP NOTE_NLP Table	Composition.entry.FamilyMemberHist	FHIR Document Resource
note_nlp_source_concept_id	OMOP NOTE_NLP Table	offset	NLP2FHIR NLP Extension, cTAKES, OMOP NOTE_NLP Table
nlp_system	OMOP NOTE_NLP Table	raw_text	NLP2FHIR NLP Extension
nlp_date	OMOP NOTE_NLP Table	context	NLP2FHIR NLP Extension, cTAKES
nlp_datetime	OMOP NOTE_NLP Table	nlp_system	NLP2FHIR NLP Extension, OMOP_NOTE_NLP Table
term_exists	OMOP NOTE_NLP Table	nlp_date	NLP2FHIR NLP Extension, OMOP_NOTE_NLP Table
term_temporal	OMOP NOTE_NLP Table	nlp_datetime	NLP2FHIR NLP Extension, OMOP_NOTE_NLP Table
term_modifiers	OMOP NOTE_NLP Table	term_temporal	NLP2FHIR NLP Extension, cTAKES, OMOP_NOTE_NLP Table
Document Type	PhEMA QDM NLP Extension	confidence_score	NLP2FHIR NLP Extension
Document Section	PhEMA QDM NLP Extension	conditional_modifier	NLP2FHIR NLP Extension, cTAKES
Negation	PhEMA QDM NLP Extension, ConTEXT	negated_modifier	NLP2FHIR NLP Extension, cTAKES
Experiencer	PhEMA QDM NLP Extension, ConTEXT	certainty_modifier	NLP2FHIR NLP Extension, cTAKES
Temporality	PhEMA QDM NLP Extension, ConTEXT	LabDeltaFlag_modifier	NLP2FHIR NLP Extension, cTAKES
Certainty	PhEMA QDM NLP Extension, ConTEXT		
Recorder	PhEMA QDM NLP Extension, ConTEXT		

Generation of NLP Rulesets for Targeted NLP Phenotyping Tasks. CQL definitions make wide use of valuesets (collections of codes from medical terminologies) to represent medical concepts. We implemented a mechanism that can transform the valuesets defined in a CQL-based phenotyping algorithm into a collection of editable NLP rulesets based on both UMLS Concept Unique Identifiers (CUIs) and display names. Specifically, value set OIDs as supplied in the source CQL are retrieved as a FHIR ValueSet resource from the NLM's Value Set Authority Center (VSAC). The individual codes (whether they be ICD9, ICD10, SNOMED, LOINC, etc.) are then cross-referenced using the UMLS and any synonym or children codes as defined by the UMLS are also retrieved. The display names of these individual synonym and children codes as then defined by the UMLS are then used to construct the NLP ruleset. We then integrated the NLP rulesets into the NLP2FHIR clinical data normalization pipeline to support targeted NLP tasks for phenotyping. The generated NLP rulesets can be reviewed by investigators for adding or removing entries if applicable and then are finalized and exported to a format compatible with the customized dictionary in the NLP engine (ie, cTAKES[15]) used in the NLP2FHIR pipeline. By replacing the default dictionary with the customized dictionary, the NLP2FHIR pipeline can be configured to perform targeted NLP tasks for information retrieval. For example, with an NLP ruleset for the diabetes mellitus generated and a corresponding customized dictionary configured, the NLP2FHIR pipeline will produce diabetes mellitus-specific NLP outputs to support NLP phenotyping tasks as defined in the CQL phenotype algorithm. Figure 1 shows an overview of the CQL4NLP pipeline's workflow

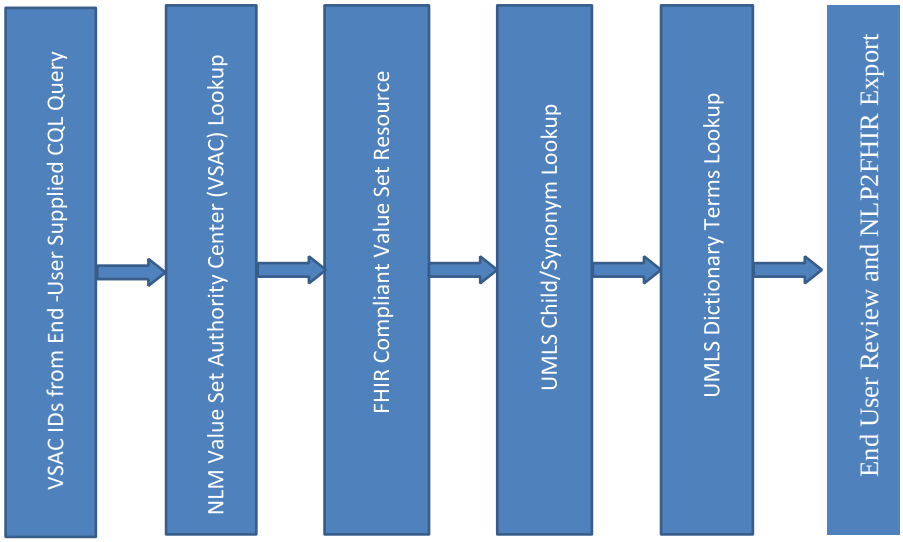


Figure 1. CQL4NLP Pipeline Workflow.

Evaluation Design. We evaluated the utility of the NLP extensions using the phenotype algorithms published in PheKB, and the performance of the NLP rulesets using an obesity phenotyping algorithm and two obesity datasets.

1) Utility of the NLP extensions. From our collection of phenotypes from PheKB, we reviewed each of the downloaded files for guidance from the phenotype author regarding the implementation of NLP. Because the phenotype definitions were predominantly represented as a narrative description of the logic, technically anyone implementing each phenotype could choose to use NLP at their institution, and could do that in any capacity. Therefore, we focused our review on the phenotype author explicitly noting where and how NLP should be used. For each phenotype definition, we identified guidance from the author that stipulated the use of various categories of NLP logic as described in Table 2.

Table 2. Guideline for the review of the PheKB phenotype algorithms with a NLP component

NLP System	String; Does the phenotype prescribe the use of a specific NLP system?
Unstructured Term List	Boolean; Does the phenotype's NLP component use a list of terms not tied to a medical vocabulary?
Structured Term List	Boolean; Does the phenotype's NLP component use a list of coded concepts from a medical vocabulary?
Document Type	String; Document type(s) that the NLP is intended to be run against. If no list is provided, no specific guidance from the author is provided. Multiple note types may be used in different parts of the phenotype, and so the list reflects all document types mentioned throughout the phenotype.
Document Section	String; Document section(s) that the NLP is intended to be run against within a document. Sections may be used in different parts of the phenotype, and may be associated with some specific document types. This is not explicitly captured, and so the list reflects all document sections mentioned throughout the phenotype.
Document Author	String; Restrictions on the person, department, etc. who authored the document. There may be some overlap conceptually between this and document type in some situations, but this was annotated separately.
Coordination of terms	String; Is there any specific coordination of terms that needs to be performed. Examples so far include terms being within the same sentence, a concept plus a specific body location term, or medication name and dose.
Negation	Boolean; Does NLP explicitly account for negated terms?
Hypothetical/Certainty	Boolean; Does NLP look at the certainty or hypothetical nature of terms?
Historical	Boolean; Does NLP consider historical context of an identified term?
Experiencer	String; Does NLP consider who experienced a specific event/condition? This relates to the patient or family members.
Use of NLP not explicitly described	Boolean; Noted if the algorithm never describes any explicit use of NLP. In these instances, I think we just exclude it, but I'm including them in the review for completeness.

2) Performance of NLP rulesets. When discussing standardization approaches, it is important to evaluate the ability for the standardization model to adequately represent the underlying data as well as the impact of the inherent granularity loss that accompanies any standardization effort. The purpose of the evaluation is therefore to demonstrate that the performance of the generated rulesets for targeted NLP tasks is comparable with that of the generic NLP2FHIR pipeline. We note that NLP2FHIR's NLP-based information extraction is done based on a UMLS dictionary definition. For the purposes of evaluating any loss in granularity resulting from standardization, we can thus compare the performance of the random forest model using the original full-form input as well as a version with input constructed from a standardized CQL definition.

To that end, we have adapted our results from a previous study done on the i2b2 obesity and comorbidity phenotyping challenge.[10] This previous study resulted in a best-performing random forest model for the identification of obesity's comorbidities utilizing NLP artifacts as input features. Specifically, we reused an obesity phenotype algorithm with a number of NLP components that define the positive mentions of 7 of these comorbidities in unstructured discharge summary: Asthma, Atherosclerotic cardiovascular disease, Congestive heart

failure, Depression, Diabetes mellitus, Hypercholesterolemia, and Hypertension. To construct a standardized CQL definition that would appropriately capture the data needs of our random forest model, we examined the trained best-performing random forest models from our previous study and identified the top 10 most-contributing features for each comorbidity. We then mapped these features into corresponding NLM Value Set Authority Center (VSAC) OIDs to embed into a corresponding CQL definition. This CQL definition was then run through CQL4NLP, and the resulting dictionary was plugged back in to NLP2FHIR to generate a new feature set that could then be used as model input. The model was then re-trained and the resulting performance compared to ensure that the performance using NLP inputs generated from the standardized definitions is comparable to that of the model using the unstandardized input definitions.

Results

A Minimal NLP Data Model: Table 3 shows the data elements of a minimal NLP data model. In total, there are 11 data elements defined in the minimal NLP data model, in which three of the data elements can be represented with the FHIR Composition elements, and two of them with the FHIR Valueset, and six of them with the FHIR extensions. All 11 data elements were identified based on their application in real-world eMERGE phenotyping algorithms (n=42) (see the Material section). Figure 2 shows a portion of a CQL definition example illustrating the use of the minimal NLP data model.

Table 3. The data elements of a minimal NLP data model

Data Element	Type	Implementation	Description
Document Type	Composition	Composition.type	type of document
Document Section	Composition	Composition.section.code	classification of section
Document Author/Recorder	Composition	Composition.author	Who and/or what authored document
Unstructured Term List	Valueset	Valueset	unstructured terms
Structured Term List	Valueset	Valueset	structured terms from a vocabulary
NLP System	FHIR Extension	nlp_system	Name and version of the NLP system that extracted the term. Useful for data provenance
Conditional Modifier/Coordination	FHIR Extension	conditional_modifier	Used to indicate that a procedure or assertion occurs under certain conditions
Negation/Negated Modifier	FHIR Extension	negation_modifier	Used to indicate that a procedure or assertion did not occur or does not exist
Certainty/Certainty Modifier	FHIR Extension	certainty_modifier	An introduction of a measure of doubt into a statement
Temporality/historical	FHIR Extension	temporality_modifier	The time modifier associated with the extracted term
Experiencer	FHIR Extension	experiencer	who experienced a specific event/condition. This relates to the patient or family members.

```
valueset "Asthma VSAC Codes": '2.16.840.1.113883.3.117.1.7.1.271'
valueset "Atherosclerotic cardiovascular disease VSAC Codes": '2.16.840.1.113883.3.464.1003.104.12.1003'
valueset "Congestive heart failure VSAC Codes": '2.16.840.1.113883.3.526.3.376'
valueset "Depression VSAC Codes": '2.16.840.1.113883.3.600.145'
valueset "Diabetes mellitus VSAC Codes": '2.16.840.1.113883.3.464.1003.103.12.1001'
valueset "Hypercholesterolemia VSAC Codes": '2.16.840.1.113762.1.4.1047.100'
valueset "Hypertension VSAC Codes": '2.16.840.1.113883.3.464.1003.104.12.1016'
valueset "Unstructured Discharge Summary Section Codes":'2.16.840.1.113883.3.464.1003.104.12.1016.1'


define "Presence of Diabete Mellitus Mentions in Unstructured Discharge Summary":
    exists (["Composition": "Discharge summary"] doc
    where doc.section.code in "Unstructured Discharge Summary Section Codes"
          and doc.section.entry.reference = '2.16.840.1.113883.3.464.1003.103.12.1001'
          and doc.section.entry.type = 'Condition'
          and exists (
            doc.section.entry.extension Extension
            where Extension.url.value = 'http://projectphema.org/fhir/extensions/certainty_modifier'
            and (Extension.value as FHIR.decimal) = 1.0
            )
    )
```

Figure 2. A portion of a CQL definition example illustrating the use of the minimal NLP data model

Utility of the NLP extensions. As shown in Figure 3, all 11 data elements were identified in the real-world eMERGE phenotyping algorithms. Out of 42 algorithms, 90.5% of algorithms had unstructured term list definition; 59.5% of

algorithms had document type definition; 35.7% of algorithms had document section definition; and 35.7% of algorithms had negation definition.

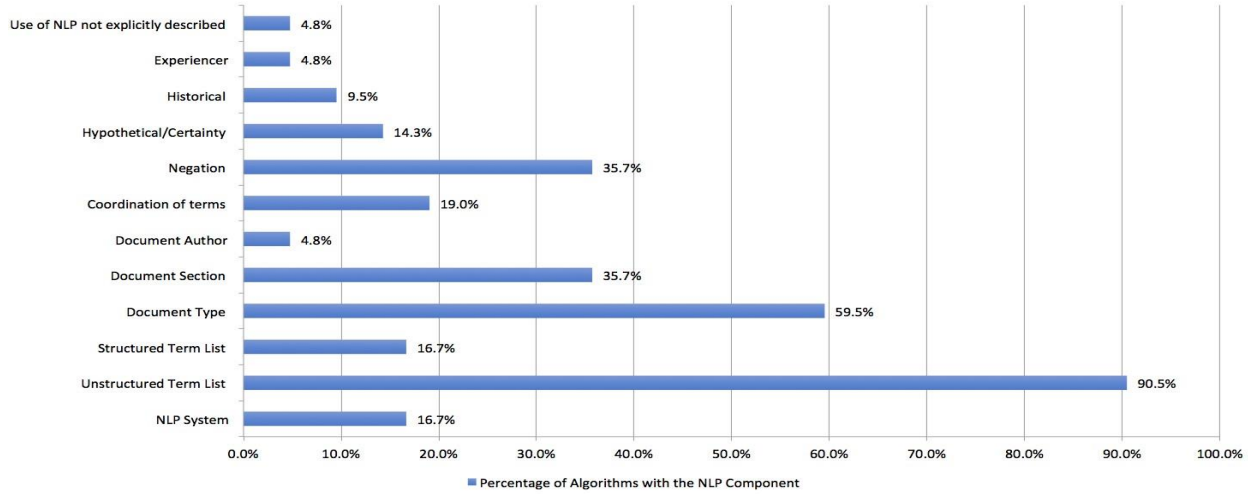


Figure 3. Percentage of algorithms with the NLP components

Performance of NLP rulesets. A total of 34 valuesets were identified from VSAC using the random forest model features (see Table 4). The NLP ruleset generation mechanism was integrated into the NLP2FHIR pipeline and the NLP rulesets generated from the 34 valuesets enabled similar performance for the identification of 7 obesity comorbidities from unstructured discharge summaries as that of generic NLP2FHIR pipeline used in the original study (baseline) (see Table 5). The results indicated that standardized valuesets, if well-constructed, can be used in the CQL to define NLP components for effectively supporting targeted NLP tasks for computable phenotyping with minimal loss in granularity.

Table 4. A list of valuesets (n=34) identified from VSAC using the random forest model features.

valueset "Anticoagulants": '2.16.840.1.113762.1.4.1206.19'
valueset "Respiratory Problems": '2.16.840.1.113883.3.666.5.1588'
valueset "Inflammatory and Autoimmune Disorders": '2.16.840.1.113883.3.3157.1834'
valueset "Depression Medications": '2.16.840.1.113762.1.4.1182.101'
valueset "Beta Agonists": '2.16.840.1.113762.1.4.1204.16'
valueset "Asthma Therapy": '2.16.840.1.113883.3.464.1003.196.12.1212'
valueset "Wheezing": '2.16.840.1.113762.1.4.1182.55'
valueset "Allergy and Intolerance Treatment": '2.16.840.1.113762.1.4.1186.7'
valueset "Asthma": '2.16.840.1.113883.3.526.3.362'
valueset "Inhaled Corticosteroids": '2.16.840.1.113762.1.4.1182.58'
valueset "Antithrombotic Therapy": '2.16.840.1.113883.3.117.1.7.1.201'
valueset "Vascular Surgery": '2.16.840.1.113883.3.666.5.713'
valueset "Heart Disease": '2.16.840.1.113762.1.4.1182.308'
valueset "Lipid-Lowering Agent": '2.16.840.1.113883.3.117.1.7.1.217'
valueset "Cardiac Surgery excl CABG": '2.16.840.1.113883.3.666.5.701'
valueset "CABG": '2.16.840.1.113883.3.666.5.694'
valueset "Chest Pain": '2.16.840.1.113762.1.4.1146.750'
valueset "Venous Thromboembolism": '2.16.840.1.113883.3.117.1.7.1.279'
valueset "Clinical Drugs": '2.16.840.1.113762.1.4.1010.4'
valueset "CHD and CHD Risk Equivalent": '2.16.840.1.113883.3.600.863'
valueset "Respiratory Disease": '2.16.840.1.113883.3.666.5.2162'
valueset "Chronic Kidney Diseases": '2.16.840.1.113762.1.4.1182.276'
valueset "Congenital Heart Disease": '2.16.840.1.113762.1.4.1146.1092'
valueset "Depressive Disorder": '2.16.840.1.113883.3.600.145'
valueset "Dyspnea": '2.16.840.1.113762.1.4.1182.47'
valueset "Psychiatric Disorder": '2.16.840.1.113883.3.117.1.7.1.299'
valueset "Sleep Disorder": '2.16.840.1.113883.3.1240.2017.3.2.2007'

valueset "Adverse Clinical Reactions": '2.16.840.1.113883.3.2074.1.1.30'
valueset "Hypertension": '2.16.840.1.113883.3.3157.4022'
valueset "Diabetes": '2.16.840.1.113883.3.464.1003.103.12.1001'
valueset "Obesity": '2.16.840.1.113762.1.4.1164.45'
valueset "Diabetic Nephropathy": '2.16.840.1.113883.3.464.1003.109.12.1004'
valueset "Hyperlipidemia": '2.16.840.1.113762.1.4.1222.73'
valueset "Anti-Hypertensive Therapy": '2.16.840.1.113762.1.4.1116.423'

Table 5. The performance comparison between NLP ruleset-based approach and baseline approach.

Study	Comorbidity	Pmicro	Pmacro	Rmicro	Rmacro	F1micro	F1macro
CQL4NLP (NLP Rulesets)	Asthma	0.9782	0.4709	0.9782	0.4867	0.9782	0.4785
	CAD	0.8753	0.6884	0.8753	0.4777	0.8753	0.488
	CHF	0.9113	0.6051	0.9113	0.6217	0.9113	0.6131
	Depression	0.9565	0.9148	0.9565	0.9051	0.9565	0.9099
	Diabetes	0.9026	0.4472	0.9026	0.4582	0.9026	0.4526
	Hypercholesterolemia	0.8964	0.4511	0.8964	0.4538	0.8964	0.4516
	Hypertension	0.9321	0.61	0.9321	0.6119	0.9321	0.6109
	Averages	**0.9218**	**0.5982**	**0.9218**	**0.5736**	**0.9218**	**0.5721**
Original (Baseline)	Asthma	0.9484	0.4656	0.9484	0.4315	0.9484	0.4462
	CAD	0.8793	0.4405	0.8793	0.4604	0.8793	0.4501
	CHF	0.9073	0.6025	0.9073	0.6185	0.9073	0.6102
	Depression	0.9625	0.9341	0.9625	0.9086	0.9625	0.9208
	Diabetes	0.9085	0.4518	0.9085	0.4604	0.9085	0.456
	Hypercholesterolemia	0.9104	0.4564	0.9104	0.4617	0.9104	0.4587
	Hypertension	0.9222	0.6073	0.9222	0.5944	0.9222	0.6
	Averages	**0.9198**	**0.5655**	**0.9198**	**0.5622**	**0.9198**	**0.5631**

Discussion

CQL4NLP and Implications on Standardized Distributions of Phenotyping Algorithms. The distribution of information contained within clinical narratives as opposed to structured data sources varies from institution to institution. This has important implications on traditional CQL queries that are designed to operate solely on structured data, as the information that can be retrieved solely from structured data may not align perfectly with that of other institutions. It follows that the viability of the CQL query to adequately perform phenotyping may come into question if, for instance, a crucial component of the query is simply not present in structured data. CQL4NLP may be able to mitigate this issue. From our case study on the obesity and comorbidity phenotyping task, we have seen that we can generate fairly comprehensive NLP dictionaries with a high degree of fidelity, which allows for a standardized representation of NLP rulesets and autonomous dictionary generation from VSAC value set OIDs. As such, the value set definitions contained within the CQL query can be translated into an equivalent, fairly comprehensive, NLP ruleset due to its nature of cross-referencing with the UMLS,[15] thus allowing for successful execution and utilization of all relevant information of the query, regardless of whether said information is in a structured or unstructured data source.

This support for converting CQL queries into equivalent NLP rulesets thus allows CQL to become a better standard for expressing phenotyping algorithms, as a limitation amongst many phenotyping algorithms, for example

cataracts[22], inflammatory bowel disease[23], and rheumatoid arthritis[24], necessitated inclusion of NLP-derived criteria. On the other hand, under this scheme, such custom definitions for NLP components may no longer be as necessary.

CQL4NLP, NLP2FHIR Integration, and Computable Phenotypes. It is important to note, however, that it is not sufficient to simply generate NLP rulesets from CQL in order to render inclusion of NLP feasible for computable phenotyping purposes. Fundamentally, CQL is a query that is run on information presented in the format of some underlying data model (e.g. FHIR, QDM), rather than a definition on how to extract said information itself, as would be the case with an NLP ruleset. There therefore exists an additional step for computable phenotyping purposes, where the NLP ruleset must be run upon some data source and said information presented in a data model that is accepted by a CQL execution engine. For our purposes, and due to its wide adoption, we chose FHIR as the target data model.

Incorporating NLP-Derived Artifacts into CQL Queries. Of course, it is important to note that NLP-derived artifacts are not directly congruent with structured data, as they possess metadata elements that may be highly clinically relevant, such as negation and hypotheticals, that would not be present in artifacts derived from structured data.

The capability to handle this metadata as part of the query is thus critical. Fortunately, when utilizing a FHIR model, CQL already provides a limited capability to handle queries against arbitrary metadata fields, using a system termed FHIRpath where fields within FHIR resources are traversed and can be referenced using a dot-delimited, path-like structure. We can thus query these metadata elements, which if utilizing NLP2FHIR would be stored in FHIR extension metadata fields, and could then simply incorporate these elements into any CQL query by attending to several Boolean clauses.

Nevertheless, this requirement does somewhat contradict our desired target of having CQL queries be operable on both structured and NLP-derived data without need for any modification, although said modifications are relatively minor. There are two possible approaches to address this, via pre-filtering of the metadata to sensible defaults, or by direct modification of the CQL execution engine to incorporate those defaults into a query. In this study, we opted for the former due to ease of implementation, although the latter may offer a more robust solution.

With respect to the defaults used, we note that a resource being present in structured data is naturally semantically equivalent to being a positive, present, non-hypothetical, and with the patient as the subject. It stands to reason that this would also be the criteria we would filter NLP artifacts by default, and as such in our pre-filtering approach we only retained NLP artifacts that matched these criteria. While a reasonable default, this is somewhat limiting – prefiltering does mean that we would not be able to utilize other metadata values than the default, even if so desired, as the requisite information is filtered out even prior to query computation. This is particularly pertinent in the case of family history mentions, which are commonly exclusively captured in unstructured data and may serve great clinical relevance. The more elegant solution then would be a direct modification of the CQL execution engine to assume these defaults if not otherwise specified as part of the query – this is feasible so long as the NLP-based metadata is stored along some consistent path, as would be the case with the adoption of FHIR NLP extensions. We have left such an implementation as part of future work.

Future Work. A key concern of any NLP system is the question of robustness and reproducibility across a variety of different platforms. To an extent, we have addressed the question of reproducibility by pairing NLP dictionary generation against the UMLS metathesaurus and VSAC managed value sets, ensuring a consistent, versioned, and curated definition for the reconstructed NLP rulesets.

The primary issue with this approach, however, is the issue of robustness. Fundamentally, it has been found that to achieve optimal performance with rule-based NLP systems, some customization of rulesets is necessary. Such is likely to be the case here, and while we have included limited provisions for such customizations on the part of the user as part of the post-dictionary generation step for any users, such customizations are not re-encoded into a standardized, distributable format. These customizations may be useful as, writ large, they may be crowd-sourced to further enhance dictionaries.[25, 26] We have left examination of how to distribute such customizations as part of a standard format, possibly through extensions to FHIR ValueSets, as future work.

Conclusion

We demonstrated that it is feasible to use the FHIR NLP extensions together with the FHIR Composition resource for standardized representation of NLP components in phenotyping algorithms. The NLP ruleset generation mechanism created a reproducible process for defining standardized NLP components of a phenotyping algorithm

and its execution against common data models. The source code of the CQL4NLP ruleset generation application is publicly available at: https://github.com/BD2KOnFHIR/CQL4NLP.

Acknowledgement

Research reported in this publication was supported by National Institutes of Health under the awards FHIRCat (R56EB028101), BD2K (U01 HG009450), and PhEMA (R01 GM105688). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1):117-21. doi: 10.1136/amiajnl-2012-001145. PubMed PMID: 22955496; PMCID: PMC3555337.
2. Gundlapalli AV, Redd A, Carter M, Divita G, Shen S, Palmer M, Samore MH. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. J Am Med Inform Assoc. 2013;20(e2):e355-64. doi: 10.1136/amiajnl-2013-001946. PubMed PMID: 24169276; PMCID: PMC3861921.
3. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Cai T. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc. 2015;22(5):993-1000. doi: 10.1093/jamia/ocv034. PubMed PMID: 25929596; PMCID: PMC4986664.
4. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, Huang J, Ho YL, Ananthakrishnan AN, Xia Z, Shaw SY, Gainer V, Castro V, Link N, Honerlaw J, Huang S, Gagnon D, Karlson EW, Plenge RM, Szolovits P, Savova G, Churchill S, O'Donnell C, Murphy SN, Gaziano JM, Kohane I, Cai T, Liao KP. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). Nat Protoc. 2019;14(12):3426-44. doi: 10.1038/s41596-019-0227-6. PubMed PMID: 31748751; PMCID: PMC7323894.
5. Liu C, Ta CN, Rogers JR, Li Z, Lee J, Butler AM, Shang N, Kury FSP, Wang L, Shen F, Liu H, Ena L, Friedman C, Weng C. Ensembles of natural language processing systems for portable phenotyping solutions. J Biomed Inform. 2019;100:103318. doi: 10.1016/j.jbi.2019.103318. PubMed PMID: 31655273; PMCID: PMC6899194.
6. Rasmussen LV, Brandt PS, Jiang G, Kiefer RC, Pacheco JA, Adekkanattu P, Ancker JS, Wang F, Xu Z, Pathak J, Luo Y. Considerations for Improving the Portability of Electronic Health Record-Based Phenotype Algorithms. AMIA Annu Symp Proc. 2019;2019:755-64. PubMed PMID: 32308871; PMCID: PMC7153055.
7. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, Rasmussen L, Jiang G, Pathak J, Luo Y. Developing a portable natural language processing based phenotyping system. BMC Med Inform Decis Mak. 2019;19(Suppl 3):78. doi: 10.1186/s12911-019-0786-z. PubMed PMID: 30943974; PMCID: PMC6448187.
8. Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, Harris D, Hochheiser H, Lin C, Chavan G, Jacobson RS. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records. Cancer Res. 2017;77(21):e115-e8. doi: 10.1158/0008-5472.CAN-17-0615.
9. Hochheiser H, Castine M, Harris D, Savova G, Jacobson RS. An information model for computable cancer phenotypes. BMC Med Inform Decis Mak. 2016;16(1):121. doi: 10.1186/s12911-016-0358-4.
10. Hong N, Wen A, Stone DJ, Tsuji S, Kingsbury PR, Rasmussen LV, Pacheco JA, Adekkanattu P, Wang F, Luo Y, Pathak J, Liu H, Jiang G. Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. J Biomed Inform. 2019;99:103310. doi: 10.1016/j.jbi.2019.103310. PubMed PMID: 31622801; PMCID: PMC6990976.
11. Hong N, Wen A, Shen F, Sohn S, Wang C, Liu H, Jiang G. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open. 2019;2(4):570-9. doi: 10.1093/jamiaopen/ooz056. PubMed PMID: 32025655; PMCID: PMC6993992.
12. Clinical Quality Language Specification 2020 [August 24, 2020]. Available from: https://cql.hl7.org/.
13. CMS CQL 2020 [August 24, 2020]. Available from: https://ecqi.healthit.gov/cql.

14. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, Zhu Q, Xu J, Montague E, Carrell DS, Lingren T, Mentch FD, Ni Y, Wehbe FH, Peissig PL, Tromp G, Larson EB, Chute CG, Pathak J, Denny JC, Speltz P, Kho AN, Jarvik GP, Bejan CA, Williams MS, Borthwick K, Kitchner TE, Roden DM, Harris PA.
Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc. 2015;22(6):1220-30. doi: 10.1093/jamia/ocv112. PubMed PMID: 26342218; PMCID: PMC4639716.
15. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507-13. doi: 10.1136/jamia.2009.001560.
16. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H. MedXN: an open source medication extraction and normalization tool for clinical text. J Am Med Inform Assoc. 2014;21(5):858-65. doi: 10.1136/amiajnl-2013-002190. PubMed PMID: 24637954; PMCID: PMC4147619.
17. Sohn S, Wagholikar KB, Li D, Jonnalagadda SR, Tao C, Komandur Elayavilli R, Liu H. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. J Am Med Inform Assoc. 2013;20(5):836-42. doi: 10.1136/amiajnl-2013-001622.
18. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, Pacheco JA, Tromp G, Pathak J, Carrell DS, Ellis SB, Lingren T, Thompson WK, Savova G, Haines J, Roden DM, Harris PA, Denny JC. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016;23(6):1046-52. doi: 10.1093/jamia/ocv202.
19. OMOP NOTE_NLP 2020 [August 24, 2020]. Available from: https://github.com/OHDSI/CommonDataModel/wiki/NOTE_NLP.
20. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform. 2009;42(5):839-51. doi: 10.1016/j.jbi.2009.05.002. PubMed PMID: 19435614; PMCID: PMC2757457.
21. FHIR Composition 2020 [August 24, 2020]. Available from: https://www.hl7.org/fhir/composition.html.
22. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, Chen L, Denny JC, Wilke RA, Pathak J, Carrell D, Kho AN, Starren JB. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc. 2012;19(2):225-34. doi: 10.1136/amiajnl-2011000456. PubMed PMID: 22319176; PMCID: PMC3277618.
23. Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, Gainer VS, Murphy SN, Szolovits P, Xia Z, Shaw S, Churchill S, Karlson EW, Kohane I, Plenge RM, Liao KP. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflamm Bowel Dis. 2013;19(7):1411-20. doi: 10.1097/MIB.0b013e31828133fd. PubMed PMID: 23567779; PMCID: PMC3665760.
24. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010;62(8):1120-7. doi: 10.1002/acr.20184. PubMed PMID: 20235204; PMCID: PMC3121049.
25. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, Liu S, Sohn S, Liu H, Fan J. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med. 2019;2:130. doi: 10.1038/s41746-019-0208-8. PubMed PMID: 31872069; PMCID: PMC6917754.
26. Peterson KJ, Jiang G, Brue SM, Liu H. Leveraging Terminology Services for Extract-Transform-Load Processes: A User-Centered Approach. AMIA Annu Symp Proc. 2016;2016:1010-9. PubMed PMID: 28269898; PMCID: PMC5333225.

Impact of Data Entry Interface Design on Cognitive Workload, Documentation Correctness, and Documentation Efficiency

Bryan A. Wilbanks, PhD, DNP, CRNA[1], Jacqueline A. Moss, PhD, RN, FAAN[1]
[1]University of Alabama at Birmingham, Birmingham, AL

Abstract

Clinical documentation serves as the legal record of patient care and used to guide clinical decision making. Inadequately designed data entry user-interfaces may result in unintended consequences that negatively impact patient safety and outcomes because inaccurate information is used to guide clinical decision making. This study utilized an electronic simulated documentation interface (i.e., artificial electronic health record) combined with eye-tracking hardware to analyze documentation correctness, documentation efficiency, and cognitive workload of anesthesia providers (N = 20) generating documentation using different computer-assisted data entry types (drop-down box, radio button, check-box, and free text with autocomplete suggestions). Our study methodology incorporating eye-tracking with electronic health record user interfaces to assess documentation correctness, efficiency, and cognitive workload can be translated to other health care provider types.

Introduction

Inadequate data entry user-interface design impairs documentation correctness, efficiency, and cognitive workload associated with using electronic health records (EHR)[1,2.] Poor user-interface design may also result in unintended consequences that impair patient safety and outcomes[3, 4] because incorrect information is used to guide future clinical decision making[7]. For example, user-interfaces that are percieved as complicated by the user may result in documenting patient care information into the wrong area of the EHR which result in clinicians overlooking important information because it is not located where they expect it[9.] The lack of understanding in how to optimize data entry interface design has been identified as a contributing factor to unintended consequences that result in impaired patient outcomes[3, 8, 9]. We were unable to locate literature that identified user-interface design issues that were EHR vendor specific. One study that evaluated 11 EHR vendors in aggregate found wide variability in how vendors assessed interface design issues prior to implementation, and that many user-interface customizations during implementation were based primarily on purchaser requests without assessing impact on usability or patient safety[10]. There was relatively little evidence in the literature to guide the selection of specific data entry methods according to the type of data documented.

In order to fill this knowledge gap, this study was designed to evaluate how pairing specific computer-assisted data entry types (e.g., drop-down box, radio button, check-box, and free text with autocomplete suggestions) to specific anesthesia documentation data elements influenced documentation correctness, documentation efficiency, and cognitive workload of anesthesia providers. Understanding the impact of pairing specific types of data with specific data entry methods can improve patient outcomes by enhancing documentation correctness through improved documentation efficiency and reduced cognitive workload. Our study is EHR vendor agnostic because we are assessing the user-interface at a granular level (data entry for a single datum) and the results could be translated to various EHR vendors because the information to be documented are similar. This approach also has the advantage of focusing on each specific data entry method without the potentially unknown confounding sociotechnical issues associated with large system implementation evaluation.

Background

Computer-assisted Data Entry. Clinical documentation is the process of generating a record of patient care, to serve as the legal record , assist in reimbursement for services provided, inform clinical decision support, and to create a repository of information for secondary data analysis (e.g., clinical research or quality improvement initiatives[4, 11]. Computer-assisted data entry consists of automated electronic tools to facilitate the generation of clinical documentation and if used inappropriately can produce errors in documentation[13]. For example, using documentation fields with default values to document antibiotic administration can result in inaccurate information being documented if the clinician does not edit the default values when they deviate from actual patient care. For anesthesia documentation, computer-assisted data entry methods that do not incorporate default values have been found to have higher documentation correctness and documentation efficiency when compared to paper based documentation[13].

Cognitive Workload. Medical errors have been consistently linked to the increased cognitive workload of clinicians[3, 5, 14-16]. Increased cognitive workload may impair documentation quality and result in inaccurate information being used for clinical decision making[17]. Cognitive workload consists of all the psychological processes that occur to complete a task[18]. Characteristics of the task to be completed, the individual completing the task, and the environment in which the task occurs can alter cognitive workload[18].

The link between pupil diameter changes and cognitive workload has been well established[19]. Pupil size is determined by autonomic nervous system activity that is altered by cognitive workload, emotion, and fatigue[20, 21]. Increased cognitive workload will result in pupil dilation with a return to resting pupil size when cognitive workload decreases[20, 21]. Eye tracking is beneficial in measuring cognitive workload because it overcomes the major limitation of subjective psychometric instruments by allowing a continuous and reliable assessment of objective workload[22].

Methods

Design and Setting. This study utilized an electronic simulated documentation interface (ESDI) combined with eye-tracking hardware to analyze documentation correctness, documentation efficiency, and cognitive workload of anesthesia providers generating documentation. The ESDI was a Windows-based software program, specifically designed for this study, that presented a series of documentation tasks to the participants by first displaying a video-based clinical scenario (see Figure 1 for example), followed by the participant documenting the patient care events that were observed (see Figure 2 for example). Study participants used a standard keyboard and mouse to interact with the program graphical user interface to document simulated patient care events viewed in the video. A description of the development of the ESDI, creation of simulated patient care events, and incorporation of the eye-tracking hardware are discussed below. We chose to use an ESDI because it would allow the use of eye-tracking and the ability to assess documentation at a more granular level (i.e., documentation data element with data entry method) than is possible with an EHR that presents multiple documentation options simultaneously. The ESDI in a controlled setting also eliminated multiple confounding variables related to the clinical environment and helped to narrow the relation of the data collected to the specific data and documentation type presented.

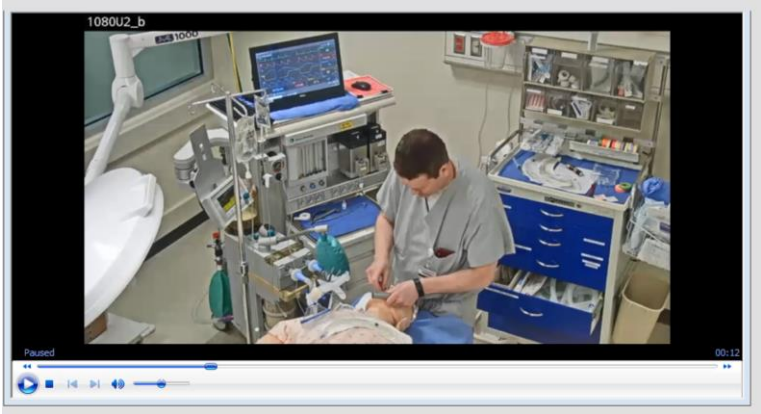


Figure 1. This example image represents a video-based clinical scenario that was presented to the study participant (this video depicts intubation of the trachea). The patient care activities presented in this video were documented in the following data entry screen.

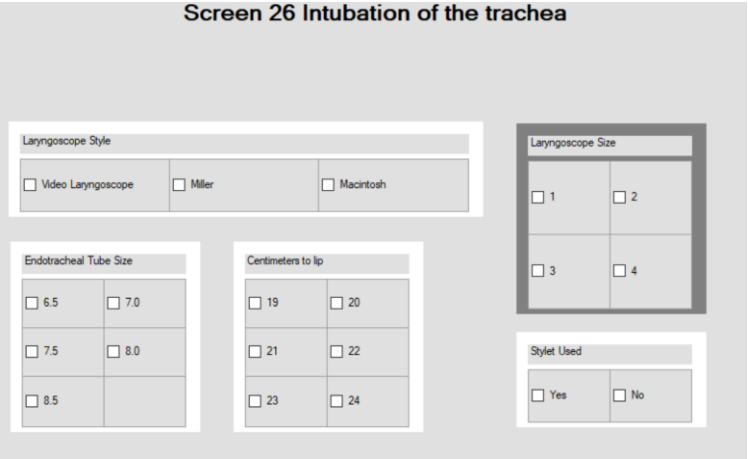


Figure 2. This example image shows the data entry screen that was used to document patient care activities that were viewed in a video-based clinical scenario. This user-interface evaluates the pairing of check-boxes (a type of computer-assisted data entry) with intubation of the trachea (a specific anesthesia documentation data element).

Study Participants. Convenience and snowball sampling were used to recruit nurse anesthetists (N =20). Inclusion criteria included nurse anesthetists who had more than one year of experience using any EHR. Exclusion criteria included participants who required the use of corrective eyeglasses, have a disease of the eyes, or on a medication that alters pupil reactivity. Individuals with eye disease were excluded because of the potential impact on eye-tracking and measuring pupil size. There was an additional study participant that was excluded from data analysis because of an eye tracking hardware malfunction that resulted in failure to collect pupil diameter sizes.

The Electronic Simulated Documentation Interface. Our study used an ESDI instead of an EHR because we were assessing how pairing specific data elements to data entry approaches performed in an EHR-agnostic environment. The use of an EHR would have introduced bias from other data elements and documentation options displayed on a single screen (i.e., our ESDI presented only one documentation pairing without other visual elements on the computer screen). The ESDI was developed in collaboration with an experienced software programmer and designed to automatically capture data for documentation correctness, documentation efficiency, and cognitive workload (via pupillary changes recorded by an eye-tracker) for every participant. A description of the operational definitions for these variables are described below. The video-based clinical scenarios were developed and recorded by the primary researcher of this study (BAW) and were based on routine anesthesia tasks. Video-based clinical scenarios were used to provide a consistent clinical scenario with standard content for documentation of clinical events. Additionally, narrative text that described the clinical scenario was not used instead of video because this might create bias from differences in our study participants literacy instead of documentation of observed clinical events.

The documentation elements of the ESDI were composed of a subset of the minimum required documentation data elements that are present in the intraoperative anesthesia documentation[23]. This sub-set of documentation data elements was selected because it consisted of the information that is most likely to be incorrectly documented in the EHR[24]. The nine minimum required data elements that the study participants were required to document included: (a) antibiotic administration, (b) inhaled gas flow rates, (c) neuromuscular function testing, (d) fluid intake/output, (e) intubation of the trachea, (f) extubation of the trachea, (g) insertion of a laryngeal mask airway, (h) removal of a laryngeal mask airway, and (i) medication administration. Since there were four different data-entry methods and nine different required documentation data elements there were a total of 36 unique pairings.

Before beginning the study each participant completed a tutorial built into the ESDI. The tutorial presented a series of five standardized video-based clinical scenarios with corresponding data entry screens to demonstrate how each specific computer assisted data entry type (drop-down box, radio button, check-box, and free text with autocomplete suggestions). The purpose of this tutorial was to familiarize the study participant with the computer assisted data entry methods to reduce any history bias caused by lack of familiarity with the ESDI functionality.

Study Variables. The study variables were documentation correctness, documentation efficiency, and cognitive workload. Documentation correctness is often defined in terms of correctness and completeness[13, 25]. Documentation correctness is high if the documentation contains the minimum required amount of information, and that information is a true description of actual patient care[13, 26]. Documentation efficiency is defined in terms of the total amount of time required to do a specific task and the total number of physical interactions[4]. Lastly, cognitive workload is defined conceptually as the amount of mental tasks that must be completed in a given time frame to complete a pre-defined objective[27]. All study variables were automatically calculated via the ESDI software to help improve the reliability of measurements.

Documentation correctness. Documentation correctness was operationalized as the percent-agreement score between the study participant's final documentation and the expected documentation elements associated with the standardized patient case scenario used to develop the ESDI.

Documentation efficiency. The ESDI presented multiple data entry screens that each contained a unique pairing of computer-assisted data entry method to a specific mandatory data element. Documentation efficiency was calculated using two approaches: (a) the amount of time between the first appearance of the data entry screen to the time the documentation was entered, and (b) the total number of mouse clicks and keystrokes for each data entry screen.

Cognitive workload. Cognitive workload was measured using eye-tracking equipment that measures real-time changes in pupil size during the study participant's use of the ESDI. The subtractive method of pupillary baseline correction was used in this study because it was one of the most common approaches used in the literature, and is the least effected by bias from inaccurate baseline pupil size measurement caused by high eye blinking rates[28]. Baseline correction is necessary to reduce the bias caused by normal variations in pupil size between individuals[28]. Subtractive baseline correction is done by subtracting the maximum pupil diameter size from the resting pupil diameter size[2]. The GP3 HD Eyetracker 150 Hz (Vancouver, British Columbia) was used for this study and has been shown to be reliable at measuring pupillary changes at intervals of 8 milliseconds with an accuracy in pupil diameter size measurement of +/- one pixel[9].

The eye-tracking equipment was mounted below the computer monitor. The ambient room lighting and desktop computer screen luminescence were kept constant to prevent artifact due to pupillary accommodation to changes in lighting. Since normal pupil diameter size in well-lit rooms ranges from 2 mm to 5 mm, any values outside of this range were identified as artifacts and treated as missing data[29].

Ethical Considerations. Prior to study participant recruitment this study received approval from the University of Alabama at Birmingham institutional review board (IRB-00001656).

Data Analysis. SPSS version 25 (Armonk, NY) was used to analyze the data with descriptive statistics, ANOVAs, effect sizes, and Pearson *r*. ANOVAs were used to detect differences in documentation correctness, documentation efficiency, and cognitive workload for each unique pairing of computer-assisted data entry method to mandatory documentation data element. Pearson *r* was used to detect associations between documentation correctness, documentation efficiency, and cognitive workload. The equivalent non-parametric tests were used for non-normally distributed data.

Results and Analysis

Twenty study participants completed the study protocol during a three-week time period. The study sample was 60% (n = 12) female with an average age of 43 (SD = 6.75) years. There were no differences in the study variables based on gender. The results and analysis of the study data are presented below.

Documentation Correctness.

Computer-assisted Data Entry Methods. There was a large effect size difference ($\eta^2 = 0.2$) in documentation correctness between the different computer-assisted data entry methods ($F(35, 716) = 4.91$, $p < .001$). A *post-hoc* analysis identified check-boxes as having the lowest documentation correctness and radio buttons as the highest. Documentation correctness for the computer-assisted data entry methods rated highest to lowest were radio buttons (M = 92%, SD = 15%), free text (M = 89.5%, SD = 25%), drop-boxes (M = 85%, SD = 24%), and check-boxes (M = 83%, SD = 23%).

There was a medium effect size association between documentation correctness and total number of mouse clicks ($r = -0.20$, $p < .001$). A higher number of mouse clicks occurred with the use of radio buttons and free text with all of the anesthesia documentation data elements and resulted in the highest documentation correctness. The anesthesia

documentation data elements that demonstrated the highest documentation correctness were associated with extubation of the trachea, medication administration, and neuromuscular function testing.

Anesthesia Documentation Data Elements. See Table 1 for a summary of documentation correctness for each specific anesthesia documentation data element. There was a medium effect size difference ($I]^2 = .07$) in documentation correctness between each type of documentation data element ($F(8, 711) = 6.39$, $p < .001$). A *post-hoc* analysis identified the anesthesia documentation data elements with the highest documentation correctness were extubation of the trachea (M = 95.7%, SD = 11.1%), medication administration (M = 94.7%, SD = 11%), and neuromuscular function testing (M = 93.8%, SD = 24.4%). Additionally, fluid intake/output (M = 78.6%, SD = 25.9%) and antibiotic administration (M = 81.6%, SD = 28.6%) had the lowest documentation correctness.

Table 1. Documentation correctness for each type of anesthesia documentation data element.

Documentation Data Elements	**Mean**	**Standard Deviation**	**Number**
Extubation of the trachea	95.7%	11.1%	80
Medication administration	94.7%	11.0%	80
Neuromuscular Function Testing	93.8%	24.4%	80
Insertion of laryngeal mask airway	87.8%	21.1%	80
Inhaled gas flow rates	85.6%	26.7%	80
Intubation of the trachea	85.1%	16.0%	80
Removal of laryngeal mask airway	84.7%	18.0%	80
Antibiotic administration	81.6%	28.6%	80
Fluid intake and output	78.6%	25.9%	80
Average for All Data Elements	87.5%	21.9%	720

Note. This table displays the documentation correctness (percentage correct) for each type of anesthesia data. The data elements are listed in order of highest to lowest.

Documentation Efficiency

Computer-assisted Data Entry Methods. There was a large effect size difference ($\eta^2 = 0.44$) in the total time generating documentation be**tween** the different computer-assisted data entry methods ($F(3, 716) = 185.96$, $p < .001$). An increase in the total time spent generating documentation reflects decreased documentation efficiency (i.e., higher time durations are worse). A *post-hoc* analysis identified that radio buttons and check-boxes had no statistically significant differences between each other, but the other computer-assisted data entry methods differed. The total time spent documenting from the most to least efficient is check-boxes (M = 10.66 seconds, SD = 4.94), radio buttons (M = 11.57 seconds, SD = 6.57), drop-boxes (M = 16.11 seconds, SD = 7.76), and free text (M = 30.65 seconds, SD = 14.26).

The Kruskal-Wallis test was used to detect large effect size differences for keystrokes (chi-square = 692.40, df = 3, $p < .001$) and mouse clicks (chi-square = 470.39, df = 3, $p < .001$) between the different computer-assisted data entry methods. Free text had the highest number of keystrokes (M = 56.09, SD = 34.74) and drop-boxes had the highest number of mouse clicks (M = 4.39, SD = 1.60). There were no statistically significant differences in mouse clicks between radio buttons and check-boxes.

Several statistically significant associations were identified. There was a large effect size association between study participant age and total time generating documentation ($r = 0.47$, $p < .001$). A large effect size association existed between the total number of keystrokes and the total time generating documentation ($r = 0.68$, $p < .001$). There was a large negative effect size association between total number of keystrokes and mouse clicks ($\rho = -0.49$, $p < .001$).

Anesthesia Documentation Data Elements. A large effect size difference ($\eta^2 = 0.15$) existed for the total time generating documentation between each specific anesthesia documentation data element ($F(8, 711) = 16.16$, $p < .001$). See Table 2 for a summary of total time spent generating documentation for each documentation data elements. A *post-hoc* analysis identified the data elements with the highest documentation efficiency were neuromuscular function testing (M = 7.2 seconds, SD = 6.57) and inhaled gas flow rates (M = 12.99 seconds, SD = 9.26). The documentation data elements with the worst documentation efficiency were fluid intake/output (M = 24.28 seconds, M = 9.71) and extubation of the trachea (M = 21.31 seconds, SD = 15.45).

Table 2. Documentation efficiency for each type of anesthesia documentation data element.

Documentation Data Elements	Mean (in seconds)	Standard Deviation	Number
Fluid intake and output	24.28	9.71	80
Extubation of the trachea	21.31	15.45	80
Intubation of the trachea	19.64	13.50	80
Removal of laryngeal mask airway	19.16	15.92	80
Antibiotic administration	18.80	8.14	80
Insertion of laryngeal mask airway	17.05	11.89	80
Medication administration	14.80	4.92	80
Inhaled gas flow rates	12.99	9.26	80
Neuromuscular function testing	7.20	6.57	80
Average for All Data Elements	17.25	12.11	720

Note. This table displays the documentation efficiency (total time used to generate documentation) for each type of anesthesia data. The data elements are listed in order of highest to lowest total time spent on data entry. Lower documentation times reflect higher documentation efficiency.

Cognitive Workload

Computer-assisted Data Entry Methods. Cognitive workload was calculated as the maximum pupil diameter size minus resting pupil diameter size (subtractive baseline correction). There was a small to medium effect size difference ($\Pi^2 = .04$) in cognitive workload between the different computer-assisted data entry methods ($F(3, 698) = 9.96$, $p < .01$). Free text (M = 0.547 mm, SD = .301) had the highest cognitive workload compared to check-boxes (M = 0.425 mm, SD = 0.298), radio buttons (M = 0.411, SD = 0.275), and drop-boxes (M = 0.401 mm, SD = 0.265). Cognitive workload was similar for check-boxes, radio buttons, and drop-boxes. There were 18 pupil diameter measurements treated as missing data because they were less than 2 mm or greater than 5 mm. This missing data was 2.5% of total pupil diameter measurements (18 of 720 measurements).

Anesthesia Documentation Data Elements. No statistically significant differences existed for cognitive workload between any of the anesthesia documentation data elements. Small effect size associations existed between cognitive workload and time spent documenting ($r = 0.16$, $p < .001$), total number of keystrokes ($\rho = 0.15$, $p < .001$), and total number of mouse clicks ($\rho = -0.17$, $p < .001$).

Discussion

Pairing computer-assisted data entry methods to anesthesia documentation data elements requires a consideration of the collective relationships between documentation correctness, efficiency, and cognitive workload. Additionally, the inherent properties of the computer-assisted data entry methods need to be considered to optimally pair data entry methods to the type of information to be documented. A discussion of these topics is presented below.

Documentation Correctness & Efficiency. This study found a large negative effect size association between the overall documentation correctness and efficiency, which is supported by the literature[13, 24, 30]. There is often a reciprocal relationship between documentation correctness and efficiency in EHRs where attempts to improve one often impairs the other[13, 30]. This may be partially explained by documentation correctness decreasing as the study participant was forced to spend more time with data entry and more physical interactions with the data entry user-interface. Anesthesia providers may often ignore generation of documentation in favor of direct patient care; consequently, data entry user-interfaces that require a lot of time to complete may be more likely to be ignored or abbreviated by the anesthesia provider[23].

While there is a negative association between documentation correctness and efficiency when evaluating an overall data-entry interface, there is a positive association when looking at some specific data entry fields[31, 32]. Improving the efficiency for specific data entry fields related to nursing patient admission histories has been shown to improve documentation correctness[31]. Furthermore, documentation of quality measures for oncology patients has also been shown to be more accurate when the documentation process is more efficient[33]. Our study found a similar positive relationship for documenting neuromuscular function testing and fluid intake/output where increasing the efficiency of data entry resulted in improved documentation correctness. This may be because neuromuscular function testing and fluid intake/output both had the least amount of information for data entry compared to the other documentation data elements. Since documentation data entry for fluid intake/output was the most inefficient anesthesia documentation data element in our study, documentation correctness may be improved by determining a more efficient means of documentation.

Older study participants were less efficient using free text for data entry because they documented more contextual information. For example, one of the older study participants documented, “Pt. gas flow changed to increase sats. Now at 2L O2 and 1L Air.” while a younger participant documented, “2L of O2 and 1L Air.” While both responses were technically the same in documentation correctness the older participant provided a richer description to justify the patient care provided. Younger study participants were more efficient in using all of the computer-assisted data entry options but documented only the minimum amount of information in free text. We could not locate literature that described the impact of age or total years of clinical anesthesia experience on documentation generation practices.

Cognitive Workload. This study found negative associations between cognitive workload and documentation efficiency. An increase in the amount of time, number of keystrokes, and number of mouse clicks required to complete documentation increased cognitive workload. Data entry user-interfaces need to be designed to avoid excessive keystrokes or mouse clicks that increase cognitive workload because it may result in medical errors[34]. The

cognitive workload related to using check-boxes, radio buttons, and drop-boxes with the anesthesia documentation data elements were similar. Individual anesthesia documentation data elements have been shown to have similar cognitive workload[2]. Free text had a medium effect size difference in cognitive workload from the other computer-assisted data entry methods and was also the most inefficient.

There were also no statistically significant differences in cognitive workload related to the anesthesia documentation data elements. Consequently, measuring cognitive workload is not useful at a granular level (i.e., single computer-assisted data entry method), but may be beneficial when evaluating an entire data entry user-interface that incorporates different types of computer-assisted data entry methods. Future research needs to explore the cognitive workload associated with multiple pairings of computer-assisted data entry methods to document anesthesia documentation data elements. Additionally, future research could incorporate high-fidelity clinical simulations that mimic real-world events when evaluating user-interfaces because the people and environment is known to alter how information technology is used.

Properties of the Computer-assisted Data Entry Methods. The inherent properties of radio buttons, check-boxes, drop-down boxes, and free text need to be considered before pairing them with specific documentation data elements. Cognitive workload is similar for each computer-assisted data entry method based on our findings (except for free text), therefore it does not need to be considered when choosing the other computer-assisted data entry methods. Radio buttons should be used with less than five data selection options and over use of radio buttons increases cognitive workload secondary to information overload[35, 36]. Check-boxes are ideally used with binary data options and may result in information overload if too many options are available[2, 35]. Our study found that the use of radio buttons or checkboxes with more than five data selection options resulted in impaired documentation efficiency. Drop-down boxes are suited for use when there are more than five possible data selection options[35, 36]. Drop-down boxes that require scrolling will increase cognitive workload and decrease documentation correctness[35, 36].

Free text is an appropriate selection when there is a virtually unlimited number of possible values that may be documented[36]. This study identified free text as the second highest in documentation correctness, the most inefficient, and the highest cognitive workload compared to other computer-assisted data entry methods. In the literature, free text has been consistently linked with improved documentation correctness but less completeness of data (i.e., clinicians fail to document important information)[30, 37, 38]. It is recommended to limit the use of free text if there is an option to use other computer-assisted data entry methods because free text is more likely to be incomplete[30]. Free text also limits data reusability (e.g., retrospective studies or quality improvement initiatives) because it requires human interpretation[30].

Limitations

This study had several limitations. Documentation generation in real-world settings combine multiple computer-assisted data entry methods simultaneously for each anesthesia documentation data element. This study evaluated unique pairings of a single computer-assisted data entry method with a single type of anesthesia documentation data element. This approach was chosen because of sample size limitations and the generalizability may have been impaired. Our eye-tracking hardware was not capable of measuring pupil diameter measurements consistently if the clinician wore corrective eyeglasses (which was an exclusion criteria in this study), so further studies on this topic should use eye-tracking hardware capable of compensating for corrective eyeglasses so the findings are more generalizable.

Conclusion

Inadequately designed data entry user-interfaces may result in impaired patient safety and outcomes because incorrect information is used to guide future clinical decision making. There is often tension between documentation correctness and efficiency in EHRs where attempts to improve one often impairs the other. We found this to be true in our study, documentation correctness was negatively associated with efficiency. However, we found that documentation data elements that contained a minimal amount of information (e.g., neuromuscular function testing or fluid intake/output) showed improved efficiency and correctness with the use of check boxes and radio buttons. Overall, free text was the least efficient, followed by drop boxes, with check boxes and radio buttons being the most efficient. Significant differences were noted in correctness between types of data entry methods with check boxes

having the lowest documentation correctness and radio buttons the highest. Increasing the number of manual keyboard operations during documentation was shown to decrease efficiency and increase cognitive workload. However, cognitive workload associated with each individual computer-assisted data entry methods were similar when evaluating documentation at a granular level (i.e., a single type of computer-assisted data entry). This study showed how pairing specific entry methods with types of specific data can effect completeness, correctness, and cognitive workload. Inadequately designed data entry user-interfaces may result in impaired patient safety and outcomes. These study findings show how user interface design can be enhanced to increase the quality of clinical documentation.

References

1. Marian AA, Bayman EO, Gillett A, Hadder B, Todd MM. The influence of the type and design of the anesthesia record on ASA physical status scores in surgical patients: paper records vs. electronic anesthesia records. BMC Med Inform Decis Mak. 2016;16:29.
2. Wanderer JP, Rao AV, Rothwell SH, Ehrenfeld JM. Comparing two anesthesia information management system user interfaces: a usability evaluation. Can J Anaesth. 2012;59(11):1023-31.
3. Sittig DF, Wright A, Ash J, Singh H. New Unintended Adverse Consequences of Electronic Health Records. Yearb Med Inform. 2016;(1):7-12.
4. Mamykina L, Vawdrey DK, Stetson PD, Zheng K, Hripcsak G. Clinical documentation: composition or synthesis? J Am Med Inform Assoc. 2012;19(6):1025-31.
5. Campbell EM, Sittig DF, Ash JS, Guappone KP, Dykstra RH. Types of unintended consequences related to computerized provider order entry. ? J Am Med Inform Assoc. 2006;13(5):547-56.
6. Sittig DF, Singh H. Defining health information technology-related errors: new developments since to err is human. Arch Intern Med. 2011;171(14):1281-4.
7. Wilbanks BA, Geisz-Everson M, Clayton BA, Boust RR. Transfer of care in perioperative settings: A descriptive qualitative study. AANA J. 2018;86(5):401-7.
8. Ellsworth MA, Dziadzko M, O'Horo JC, Farrell AM, Zhang J, Herasevich V. An appraisal of published usability evaluations of electronic health records via systematic review. J Am Med Inform Assoc. 2016; 24(1):218-226.
9. Zheng K, Abraham J, Novak LL, Reynolds TL, Gettinger A. A Survey of the Literature on Unintended Consequences Associated with Health Information Technology: 2014-2015. Yearb Med Inform. 2016(1):13-29.
10. Ratwani RM, Fairbanks RJ, Hettinger AZ, Benda NC. Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors. *Journal of the American Medical Informatics Association : JAMIA.* 2015;22(6):1179-1182.
11. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc. 2011;18(2):181-6.
12. Shoolin J, Ozeran L, Hamann C, Bria W, 2nd. Association of Medical Directors of Information Systems consensus on inpatient electronic health record documentation. Applied clinical informatics. 2013;4(2):293-303.
13. Wilbanks BA, Berner ES, Alexander GL, Azuero A, Patrician PA, Moss JA. The effect of data-entry template design and anesthesia provider workload on documentation accuracy, documentation efficiency, and user-satisfaction. Int J Med Inform. 2018;118:29-35.
14. Weber-Jahnke JH, Mason-Blakley F, editors. On the safety of electronic medical records. International Symposium on Foundations of Health Informatics Engineering and Systems; 2011: Springer.
15. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak. 2017;17(1):36.
16. Peute LW, De Keizer NF, Van Der Zwan EP, Jaspers MW. Reducing clinicians' cognitive workload by system redesign; a pre-post think aloud usability study. Stud Health Technol Inform. 2011;169:925-9.
17. Gartner D, Zhang Y, Padman R. Cognitive workload reduction in hospital information systems : Decision support for order set optimization. Health Care Manag Sci. 2017.
18. Young MS, Brookhuis KA, Wickens CD, Hancock PA. State of science: mental workload in ergonomics. Ergon. 2015;58(1):1-17.
19. Coyne J, Sibley C, editors. Investigating the Use of Two Low Cost Eye Tracking Systems for Detecting Pupillary Response to Changes in Mental Workload. Proc Hum Factors Ergon Soc Annu Meet; 2016: SAGE Publications Sage CA: Los Angeles, CA.

20. Kok EM, Jarodzka H. Before your very eyes: the value and limitations of eye tracking in medical education. Med Educ. 2017;51(1):114-22.
21. Mosaly PR, Mazur L, Marks LB, editors. Usability evaluation of electronic health record system (EHRs) using subjective and objective measures. Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval. 2016: 313-316.
22. Matthews G, Reinerman-Jones LE, Barber DJ, Abich Jt. The psychometrics of mental workload: multiple measures are sensitive but divergent. Hum Factors. 2015;57(1):125-43.
23. American Association of Nurse Anesthetists. Documenting the Standard of Care: The Anesthesia Record. Park Ridge, IL: American Association of Nurse Anesthetists; 2010: 14.
24. Wilbanks BA, Moss JA, Berner ES. An observational study of the accuracy and completeness of an anesthesia information management system: recommendations for documentation system changes. Comput Inform Nurs. 2013;31(8):359-67.
25. Wilbanks BA, Geisz-Everson M, Boust RR. The Role of Documentation Quality in Anesthesia-Related Closed Claims: A Descriptive Qualitative Study. Comput Inform Nurs. 2016;34(9):406-12.
26. Wilbanks BA. An integrative literature review on accuracy in anesthesia information management systems. Comput Inform Nurs. 2014;32(3):56-63.
27. Wilbanks BA. An integrative literature review of contextual factors in perioperative information mangagement systems. Comput Inform Nurs. 2013;31(12):622-8.
28. Mathot S, Fabius J, Van Heusden E, Van der Stigchel S. Safe and sensible preprocessing and baseline correction of pupil-size data. Behav Res Methods. 2018;50(1):94-106.
29. Watson AB, Yellott JI. A unified formula for light-adapted pupil size. J Vis. 2012;12(10):12-.
30. Wilbanks BA, Moss J. Evidence-Based Guidelines for Interface Design for Data Entry in Electronic Health Records. Comput Inform Nurs. 2018;36(1):35-44.
31. Karp EL, Freeman R, Simpson KN, Simpson AN. Changes in Efficiency and Quality of Nursing Electronic Health Record Documentation After Implementation of an Admission Patient History Essential Data Set. Comput Inform Nurs. 2019;37(5):260-5.
32. Weng CY. Data Accuracy in Electronic Medical Record Documentation. JAMA Ophthalmol. 2017;135(3):232-3.
33. Esper P, Walker S. Improving documentation of quality measures in the electronic health record. J Am Assoc Nurse Pract. 2015;27(6):308-12.
34. Zahabi M, Kaber DB, Swangnetr M. Usability and Safety in Electronic Medical Records Interface Design: A Review of Recent Literature and Guideline Formulation. Human factors. 2015;57(5):805-34.
35. Sewell JP, Thede LQ. Informatics and nursing: Opportunities and challenges: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2013.
36. Marian AA, Dexter F, Tucker P, Todd MM. Comparison of alphabetical versus categorical display format for medication order entry in a simulated touch screen anesthesia information management system: an experiment in clinician-computer interaction in anesthesia. BMC Med Inform Decis Mak. 2012;12:46.
37. Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe Practices for Copy and Paste in the EHR. Systematic Review, Recommendations, and Novel Model for Health IT Collaboration. Appl Clin Inform. 2017;8(1):12-34.
38. Markel A. Copy and paste of electronic health records: a modern medical illness. Am J Med. 2010;123(5):e9.

Data Characterization of Medicaid: Legacy and New Data Formats in the CMS Virtual Research Data Center

Nick Williams, Ph.D, Craig S. Mayer, MS, Vojtech Huser, MD, Ph.D
National Library of Medicine, Lister Hill National Center for Biomedical Communications, Bethesda, MD

Abstract

Medicaid is a significant health insurance plan providing healthcare coverage to up to a third of the population of the United Sates. We describe two different formats of Medicaid data within Center for Medicare and Medicaid Services Virtual Research Data Center. We analyze record length, age and enrollment justification among patients for both data formats. As of December 2016, the total size of Medicaid population available from CMS is 92,953,389; 45% of patients are aged 0 to 18, 26.6% are aged 19-35 and 23.2% are aged 36-64. In terms of Medicaid eligibility, 35.6% qualify due to (child) age and 26.8% qualify due to income. We also compare the volume of Medicaid to Medicare for year 2016. We conclude that Medicaid data includes patients with significant record lengths and relatively well documented enrollment justification, which are high value assets for data reuse researchers that are willing to balance known data limitations with careful analysis design and interpretation.

Introduction

The use of healthcare claims data has increased dramatically in the last decade. Thanks to Common Data Models (CDMs), more datasets are being made accessible to larger pools of researchers without requiring dataset specific analyst specialization. In 2019, Medicare data were converted to Sentinel CDM[1,2] used by the US Food and Drug Administration (FDA)[3]. Medicaid represents another significant data source which can be used by health services researchers; provided access and integration into and curation of a CDM are made available.

In the United States, constituent state governments provide healthcare to qualified state residents (low income and highly vulnerable) through Medicaid programs. Unlike the federal program Medicare, Medicaid programs differ by state in coverage, eligibility and scope. For example, a procedure may be covered by Maryland Medicaid and not covered by the Texas Medicaid program. There are also federally mandated 'minimum' coverage requirements for Medicaid Programs. Medicaid programs are complex payer facilities that compensate clinicians, hospitals, dentists and in some cases patients themselves for clinical or other maintenance care. Medicaid programs consist of fee for service clinical care but may also integrate state sponsored birth control, vaccination, outpatient pharmacy dispensation, Supplemental Nutrition Assistance Programs (SNAP; food stamps), Child Health Insurance Programs (CHIP) and 'mother with infant' services.

Recently, the Centers for Medicare and Medicaid Services (CMS) used two different systems to organize Medicaid data into a national (multi-state) data warehouse. The legacy system, Medicaid Statistical Information System (MSIS) [4] and the emergent system, Transformed Medicaid Statistical Information System (T-MSIS)[5]. For each format, CMS also defined a corresponding extract table format for research. See Table 1 for an overview of the systems and formats. Prior to 2011, CMS collected data quarterly using the legacy MSIS format. Quarterly data collection gave each state only four data refreshes per year to achieve standardized reporting on the national level. In 2011, CMS started a transition campaign from the legacy MSIS system to T-MSIS. The transition also changed the frequency of data reporting from quarterly to monthly. The transition to the new T-MSIS format also included generating state-specific reports on data quality and a dashboard called DQ-Atlas[6]. Table 1 lists the extracts for the legacy extract format (Medicaid Analytic Extract; MAX) and current format (T-MSIS Analytic File; TAF). Because data analysts eventually work with the data extract, we will use the acronyms MAX and TAF to refer to the combination of format and extract.

Table 1. Periods overview and formats and extracts

Time Period*	Data Format	Data Extract Name
1999-2014	Legacy format: MSIS	Medicare Analytic eXtract (MAX)
2014-present	Current format: T-MSIS	T-MSIS Analytic Files (TAF)

*For 2014 and 2015 data exist in both formats

The fact that data is historically represented in two formats is a problem for analysis that needs to analyze time periods spanning both formats. As of 2020, CMS has not announced any plan to convert legacy historical MAX data into the new TAF format. One solution is for an analyst to convert both formats into a well-established Common Data Model (CDM), such as Sentinel, Observational Medical Outcome Partnership (OMOP) or others.

This study compares record length, age distributions, and enrollment justification from TAF and MAX files. We also provide descriptive characteristics of CMS Medicaid data such that health services researchers can understand the strengths and weaknesses of this data for research. We present an approach (for a limited subset of data) to assess patient enrollment justification, age at stable observation point and length of observation record (enrollment episode length).

Methods

Input data

We used a complete (100%) sample of TAF and MAX data tables from CMS via the Virtual Research Data Center's Chronic Conditions Warehouse (VRDC-CCW). The data is provided through a virtual desktop that contains a SAS instance with permissions to connect to the CCW SAS server. CCW provides an encrypted beneficiary-id which is unique to the individual across data years and is consistent across Medicare and Medicaid files. We accessed MAX data for 1999-2014 and TAF data for 2014-2016. We place a greater emphasis on the new TAF data extract because future data will be added in this format. TAF and MAX data are sub-divided into enrollment and demographics, Inpatient, Medication, Long Term and Other Care tables. 'Other Care' consist in large part of outpatient visits but it also includes emergency department encounters. We use the term 'table' to refer to tabular data. CMS documentation uses the term 'file' that is synonymous in meaning. Data are using relational database paradigm with data linked by various primary keys where each record contains at least one primary key. The Beneficiary ID, MSIS ID, Social Security Number, claim id and National Provider Identifier (NPI) are examples of TAF and MAX primary keys. Data across TAF and MAX is structured such that state is identified clearly in all tables. For example, the drug dispensation table has column 'State_CD', or 'state code'. Some data tables have two state columns distinguishing state of service as well as state reporting service as two separate data elements.

Data harmonization using a common data model

In order to analyze data in both formats, for some analyses, we convert MAX and TAF into a common format. Because the OMOP format has been selected as unifying format by several recent large informatics projects (e.g., All of Us initiative or National COVID Cohort Collaborative [N3C]), we chose the OMOP model. Due to limited scope, we only target a subset of OMOP tables and within the given table, only a subset of columns as necessary to complete our analysis. In order to query and analyze data by state, we extend OMOP tables with a column for state (State_CD).

Data Characterization

Record Length: Using Medicaid enrollment data, we assessed the length of health record (enrollment duration) on individual person level. For persons with multiple enrollment periods, we used the longest period. We did not assess cumulative enrollment of multiple periods. We calculated observed enrollment episode for TAF and MAX beneficiaries for the state of Alabama for 2014-2016 (TAF) and 1999-2013 (MAX). Enrollment start and end is recorded with monthly granularity. Our goal was to provide researchers with a size of population that was observed for at least certain period of time. For example, a person with record length of 3.4 years is counted multiple times under a count of persons with at least 1 year of claims data, at least 2 years of data and at least 3 years of data.

Age Distribution: To further allow researchers to understand Medicaid population of patients, we counted patients by age category as of certain date (Dec 31, 2016 for TAF and Dec 31, 2013 for MAX). CMS allows two modes of data use agreement: identifiable data files (IDF) and limited data sets (LDS)[7]. IDF versions of TAF and MAX data include full date of birth from which age at given date can be computed. Some tables also store age (in a given calendar year) as a separate variable for claim to patient attribution.

Enrollment Justification: Medicaid requires an enrollment qualification to receive benefits. Enrollment qualification varies by state and over time. State specific enrollment codes are only informative to enrollment specialists who can identify the human readable, state specific translations. VRDC-CCW does not curate a look up table of these state-specific enrollment codes; but instead curates them as TAF or MAX enrollment codes. State specific, TAF and MAX recoded enrollment tables are available in our *result github repository* (available at github.com/lhncbc/

CRI/tree/master/VRDC/project/Medicaid). The TAF data dictionary lists 74 codes for state enrollment justification (Eligibility Group Code; ELGBLTY_GRP_CD). The MAX data dictionary uses 28 codes (Eligibility code; EL_MAX_ELGBLTY_CD). We extracted the enrollment justification at beneficiary level for a choice month (December) from TAF and MAX in 2016 and 2013, respectively. We aggregated these enrollment codes into a distinct beneficiary count and state of enrollment. For presenting the data in brief table format, we further defined four high level enrollment justification categories (Child, Disability, Income, Other) and present the data aggregated into these categories. See file Eligibility_Categories in our results github repository that shows assignment of high-level classification categories to individual enrollment codes.

Comparison to Medicare: Because the VRDC platform contains both Medicare and Medicaid data, for a single calendar year of 2016, we compared Medicaid and Medicare data using the following parameters: data lag, total size of population, and three medication related parameters (number of dispensations, number of patients with at least one dispensation and number of dispensations per patient).

Results

Data Harmonization

To facilitate identical queries that characterize claims record length and age analyses on both Medicaid formats, we converted both TAF and MAX into OMOP model. We did not make a comprehensive Extract-Transfer-Load (ETL) code, but only targeted simplified versions of two OMOP tables (OMOP OBSERVATION_PERIOD and PERSON). Because our goal is to facilitate greater use of Medicaid data using clinical informatics approaches, we made the SAS and SQL code for this conversion available at a github repository at https://github.com/lhncbc/r-snippets-bmi/tree/master/mmbox (folder 06-MedicaidOMOP). This repository also contains other code from our previous projects that use VRDC platform. The repository is not limited to Medicaid and some code snippets cover Medicare VRDC data. The name mmbox is short form for title "Medicare-Medicaid Box" and represents a collection of code snippets or supportive spreadsheet knowledge bases.

Data characterization (TAF)

Record Length (TAF): Table 2a shows length of claim-based record for the state of Alabama (the only state we evaluated). The data show that population size drops by 35.5% (from 1.28 million to 0.82 million) if a researcher requires at least two-years of follow-up compared to at least one year.

Table 2a. TAF: Record Length for single state (AL, for 2014-2016)

Record Length	Cohort Size (# of Patients)
< 1 year	1,960,477
at least 1 year	1,275,804
at least 2 years	822,817

Age Distribution (TAF): Table 3a shows age distribution for three states (AL, CA and NJ) and for US as a whole (all states aggregated). Data for all states are available in our result repository (github.com/lhncbc/CRI/tree/master/VRDC/project/Medicaid). Table 3a shows (in whole US row) that majority of patients are pediatric patients. Comparison of their states shown in the table reveals some variability. For the state of Alabama, the total population size in table 3a (when all ages are aggregated) is slightly different from population size by record length in table 2a. This difference is because table 3a is looking at enrolment at single point in time (December 31, 2016), while table 2a includes data on patients that may have been covered for <1y in a 3-year period of 2014-2016 (TAF era).

Table 3a. TAF: Age Distribution (AL, CA and NJ; as of December 2016)

State	Age category			
	0-18	**19-35**	**36-64**	**65+**
Alabama	832,911 (57.75%)	257,856 (17.88%)	226,247 (15.69%)	125,320 (8.69%)
California	5,319,013 (36.38%)	4,039,725 (27.63%)	3,893,633 (26.63%)	1,370,289 (9.37%)
New Jersey	887,302 (42.55%)	555,651 (26.65%)	483,553 (23.19%)	158,760 (7.61%)
US (whole)	**41,812,822 (45.0%)**	**21,283,820 (22.89%)**	**21,999,016 (23.66%)**	**7,857,731 (8.45%)**

Enrollment Justification (TAF): Table 4a shows enrollment justification category for 3 states (AL, CA, and NJ). Data for all states are available in our result repository. The table shows that Alabama has high proportion of patients with 'Child' enrollment justifications category (49.1%) compared to other states. California has the highest proportion of Disability enrollment category (21.5%). Income qualification is less prevalent in Alabama (16%) compared to other table 4a states.

Table 4a. TAF: Enrollment justification category (for AL, CA and NJ; December 2016)

State	Enrollment Justification Category			
	Child (age <19)	**Disability**	**Income**	**Other**
Alabama	708,161 (49.10%)	80,028 (05.55%)	235,592 (16.33%)	418,552 (29.02%)
California	4,056,266 (27.74%)	3,148,551 (21.53%)	5,240,346 (35.84%)	2,177,495 (14.89%)
New Jersey	527,312 (25.29%)	131,330 (06.30%)	667,805 (32.03%)	758,813 (36.39%)
US (whole)	**33,051,993 (35.55%)**	**12,208,059 (13.13%)**	**24,914,598 (26.80%)**	**22,778,263 (24.50%)**

TAF Data Quality Monitoring: TAF data and the transition to the new T-MSIS format has a unique advantage over legacy format in incorporating data quality checking into the new T-MSIS system. Data Quality Atlas (DQAtlas) dashboard at medicaid.gov/dq-atlas and related set of data quality tools compare data completeness and other data quality measures by state. Currently, only data for year 2016 is available in DQAtlas. The tool produces reports on quality by file type (IP, RX, LT, and OT), by data domain (Enrollment, Payments, Provider Information) and by state. See Figure 1 for an example of DQAtlas 'Inpatient Stay' view (red color indicates 'Unusable' data). Methodology for each analysis is provided and available via hyperlink. The dashboard classifies data quality of a given data domain (e.g., visit, diagnoses, procedures, medications) from 'Low concern' to 'Unusable'. DQAtlas offers high value to data reuse researchers by pointing out data quality issues. The underlying raw data quality data is also available for download in tabular format. The methodological detail provided is sufficient for users to evaluate a given dashboard report and asses how it may impact their analysis (e.g., exclude states with low data quality from the analysis). There is no counterpart of this data quality assessment for legacy MAX data.

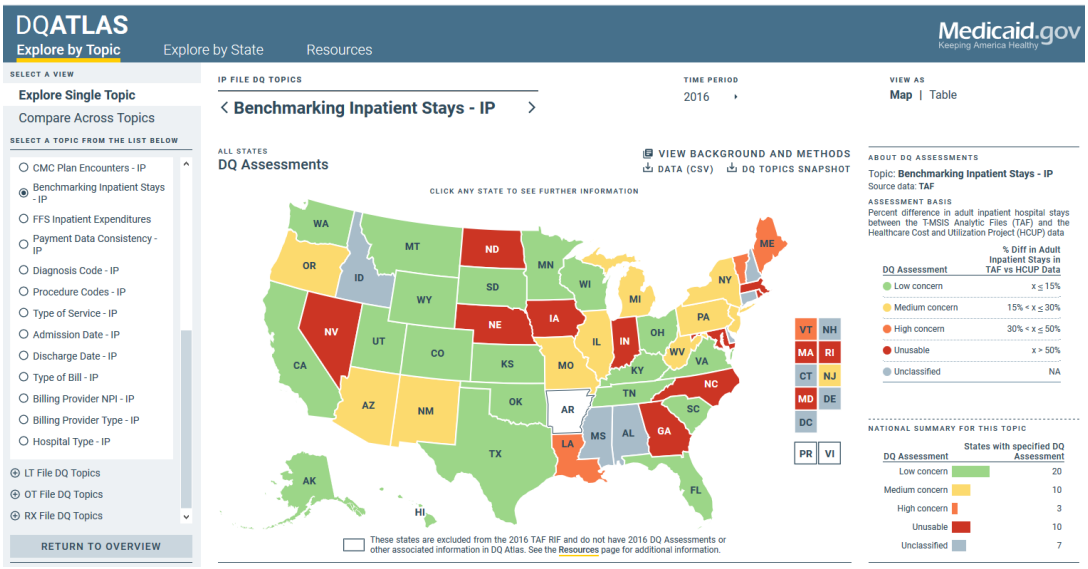


Figure 1. Example of DQAtlas view by topic (Inpatient Stays) for 2016 data (TAF era)

Data characterization (MAX)

Because the legacy MAX (and MSIS) format is no longer in active use, our analyses were simplified; we did not produce results for each individual state. Record length was evaluated in one state (AL) while age and enrollment justification category were computed on national level. We first analyzed what years are available in MAX format. Some states produced the legacy extract (MAX) from the new T-MSIS format in specific years (see result repository for years overview by state). Given 2014 inconsistencies, we chose 2013 as the end year for most of our analyses. Results and tables in this MAX section are organized in similar fashion as for TAF above. Tables use a-b suffix to show correspondence.

Record Length (MAX): In comparison with TAF, the MAX data offers a much longer temporal span of record (two years vs 10+ years). Table 2b shows record length cohort sizes for Alabama (single state we evaluated for this measure). The table shows that one year follow cohort size (of 2.31 million) drops by 56.3% (to 1 million) for cohort with at least 5 years of follow up.

Table 2b. MAX: Record Length for single state (AL; December 2013)

Record Length	Cohort size (# of Patients)
< 1 year	2,675,746
at least 1 year	2,310,698
at least 2 years	1,729,100
at least 3 years	1,405,221
at least 4 years	1,179,990
at least 5 years	1,008,900
at least 6 years	856,368
at least 7 years	717,969
at least 8 years	582,267
at least 9 years	465,068
at least 10 years	365,431
at least 11 years	273,673
at least 12 years	191,998
at least 13 years	120,059
at least 14 years	43,653

Age Distribution (MAX): Table 3b shows age distribution for whole US (all states). Most MAX patients are of age 0-18, demonstrating the strength of Medicaid data to study pediatric population. As expected, Medicaid underrepresents seniors aged 65+, and older adults (age 36-64).

Table 3b. MAX: Medicaid Age Distribution (whole US; all states, December 2013)

Age Group	Patient Count	Percentage
0-18	37,561,732	52.21%
19-35	14,088,019	19.58%
36-64	13,363,756	18.57%
65+	6,936,857	9.64%
Total	71,950,364	100.00%

Enrollment Justification (MAX): Table 4b shows enrollment justification category for all states as of December 2013. 'Child' is the most common qualification reason, followed by Disability and Other. The total Medicare population listed in table 4b (71,942,641) is different from corresponding number in table 3b (71,950,364). This is because date of birth can be missing in MSIS while patients without date of birth are included in table 3b (7,723 patients).

Table 4b. MAX: Enrollment Justification Category (All States, December 2013)

Enrollment Category	Patient Count	Percentage
Child	34,982,686	48.63%
Disability	14,454,933	20.09%
Income	9,432,647	13.11%
Other	13,072,375	18.17%
Total	**71,942,641**	**100.00%**

Comparison to Medicare

Figure 2 shows graphical comparison of age characteristics between Medicaid and Medicare (as of December 2016). The graph shows advantage of the Medicaid population for studying patients under 65 years old. The figure double counts dually enrolled patients (includes the person under both categories instead of introducing a third color). The predominant enrollment justification for Medicare is age over 65. Medicare beneficiaries under age 65 are either surviving spouse of a qualified Medicare beneficiary (who has died) or Social Security Disability Insurance (SSDI). For Medicaid, drop in patients after age 19 is most likely due to CHIP eligibility rules.

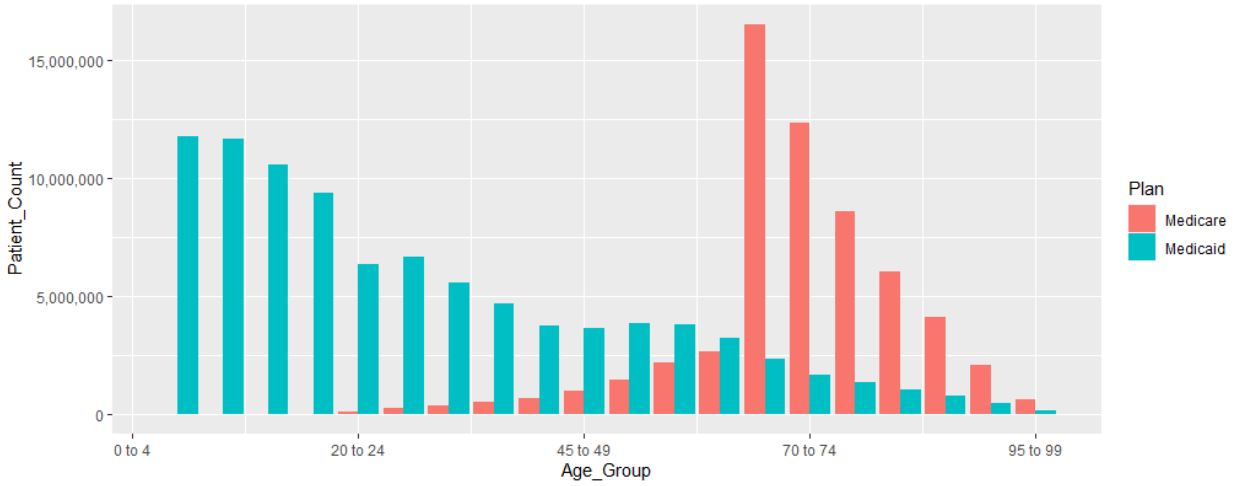


Figure 2. Population size (Y axis) comparison of Medicare and Medicaid (color) by age category (X axis; 5-year bins)

In addition to age comparison, we show, in table 5, comparison of selected measures for Medicare and Medicaid for year 2016. Both figure 2 and table 5 use TAF Medicaid data. We used 2016 because that is the latest available year with both Medicaid and Medicare data within VRDC. The first row in Table 5 clearly shows that data lag for Medicaid data is 4 years whereas for Medicare it is only 2 years. Medication-related measures show that Medicare data has more dispensations per patient and overall larger volume of dispensations (1.4 billion Medicare vs. 0.7 billion Medicaid).

Table 5. Medicare and Medicaid Comparison (as of December 2016)

Comparison measure	**Medicaid**	**Medicare**
Most recent data year (as of Aug 2020)	2016	2018
# of patients	92,953,389	59,818,296
# of dispensations	783,546,127	1,484,293,036
# of patients with a Rx dispensation	49,019,836	40,535,713
# of dispensations per patient	15.98	36.62

Discussion

VRDC users pay no extra additional project startup cost or annual cost if they include Medicaid data in their Data Use Agreement. Our data on record length show that despite more complex and time-variable Medicaid eligibly criteria, there are large cohorts of patients with multiple years of follow up (e.g., 365 thousand patients with at least 10 years of follow up in Alabama; table 3b). This is, for example, comparable to the size of All of Us cohort (356 thousand patients as of Aug 2020). We consider Medicaid data of high value to data reuse researchers, and it can answer many high level research questions, such as validation of HIV registries[8], evaluation of outcomes among special populations[9,10] or validation of complex census responses about healthcare coverage[11]. The underlying accuracy of Medicaid data in MAX is well supported for medication records[12],CHIP patients[4] and managed care patients[13]. The

Medicare data have strong advantage to be standardized on federal level by the federal nature of the program at the data origin. In stark contrast, the Medicaid TAF data quality reports (DQAtlas) that we reference point to significant differences in Medicaid data by state. Given such differences, state level analysis should always accompany any national Medicaid analysis. Several peer review studies and publications recommend state specific considerations including CMS itself[14]. CMS has given several indications that continuous quality improvement will be a key feature of T-MSIS and TAF. We expect that such efforts should only improve an already impressive Medicaid datasets.

Medicaid's primary advantage lies in presence of patients under the age of 65. While Medicare does include patients under age 65, the population mostly consists of 'total' disability patients. The Social Security Administration defines patients with 'total' disability as patients who cannot return to the workforce and who are likely to die from their disability qualifying clinical conditions (i.e., organ failure, AIDS). In addition, End Stage Renal Disease (ESRD) is also a qualifying criterion for Medicare.

In 2016, 15.56% of the Medicare population, and 91.46% of the Medicaid population was under the age of 65. Studies seeking younger patients should consider using VRDC Medicaid data from high data quality states. Medicaid has much broader enrollment criteria (74 vs 4 reasons) and while Medicaid eligibility may change over patient's lifetime, we show that there are special populations of patients that remain insured for multiple years (See tables 2a and 2b).

Limitations

Out analysis has several limitations. First, we performed some Medicaid analyses only for a single state. We chose to limit the scope for some analysis in order to reduce the report execution time. Moreover, some states did not have the expected format for some years (e.g., Arkansas does not provide TAF data for 2016). Second, our results may be different if the analysis is repeated at a later time. This is because T-MSIS formatted data is subject to updates by VRDC custodians and such updates may require rewrite of some data transformations. One such known data issue (announced by email to VRDC users) is the completeness of enrolment data. CCW announced on July 20th, 2020 a warning about TAF data quality and plans to provide an update to TAF records in late 2020 to correct the issue. Such updates are not uncommon for large data warehouses. Updates are also expected considering that the TAF format has been in full use for only three most recent data years (2014-2016). Third, due to significant data lag of 4 years, we only analyzed year 2016 as the most recent data. Researchers considering clinical analyses with Medicaid data should plan on identical limitation. Finally, in this high-level overview paper, we did not provide data stratified by both age and enrollment justification that would allow understanding of exactly which subgroups of patients have long or short claim record. We have additional pending Medicaid characterization analyses that were outside the scope of this basic overview but eventually published in future publications.

Conclusion

CMS's transition to TMSIS and TAF data extract represent years of dedicated effort to arrive at a unified federal data resource on the part of CMS and individual states. We demonstrate that Medicaid data have large volumes of age diverse, predominantly younger patients. Patients have significant record lengths and relatively well documented enrollment justification, which are high value assets for data reuse researchers that are willing to balance the known data limitations with careful analysis design and interpretation.

Acknowledgement

This work was supported by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

References

1. Ball R, Robb M, Anderson SA, Pan GD. The FDA's sentinel initiative—A comprehensive approach to medical product surveillance. Clin Pharmacol Ther. 2016;99(3):265–8.
2. The Centers for Medicare and Medicaid Services (CMS): Medicare Fee-For-Service (FFS) Claims in Sentinel Common Data Model Format [Internet]. Sentinel Initiative. 2019 [cited 2020 Aug 26]. Available from: https://www.sentinelinitiative.org/methods-surveillance-tools/software-packages-toolkits/centers-medicare-and-medicaid-services-cms

3. Cocoros NM, Pokorney SD, Haynes K, Garcia C, Al-Khalidi HR, Al-Khatib SM, et al. FDA-Catalyst—Using FDA's Sentinel Initiative for large-scale pragmatic randomized trials: Approach and lessons learned during the planning phase of the first trial. Clin Trials. 2019 Feb;16(1):90–7.
4. MacTaggart P, Foster A, Markus A. Medicaid Statistical Information System (MSIS): a data source for quality reporting for Medicaid and the Children's Health Insurance Program (CHIP). Perspect Health Inf Manag. 2011 Apr 1;8:1d.
5. Baugh D, Ires H, Irvin C, Carey A. Federal Stewardship of the Medicaid Program: Strengthening Data Systems for Effective Decision Making [Internet]. Mathematica Policy Research; [cited 2020 Aug 26]. Available from: https://www.mathematica.org/-/media/publications/pdfs/health/2017/ib-medicaid-datasystems.pdf
6. Welcome | DQ Atlas [Internet]. [cited 2020 Aug 26]. Available from: https://www.medicaid.gov/dq-atlas/welcome
7. ResDac. Differences between RIF, LDS, and PUF Data Files [Internet]. Available from: https://www.resdac.org/articles/differences-between-rif-lds-and-puf-data-files
8. Macinski SE, Gunn JKL, Goyal M, Neighbors C, Yerneni R, Anderson BJ. Validation of an Optimized Algorithm for Identifying Persons Living With Diagnosed HIV From New York State Medicaid Data, 2006-2014. Am J Epidemiol. 2020 05;189(5):470–80.
9. Brouwer ES, Napravnik S, Eron JJ, Simpson RJ, Brookhart MA, Stalzer B, et al. Validation of Medicaid Claims-based Diagnosis of Myocardial Infarction Using an HIV Clinical Cohort. Med Care. 2015 Jun;53(6):e41-48.
10. Koroukian SM, Cooper GS, Rimm AA. Ability of Medicaid claims data to identify incident cases of breast cancer in the Ohio Medicaid population. Health Serv Res. 2003 Jun;38(3):947–60.
11. Boudreaux M, Noon JM, Fried B, Pascale J. Medicaid expansion and the Medicaid undercount in the American Community Survey. Health Serv Res. 2019;54(6):1263–72.
12. Leonard CE, Brensinger CM, Nam YH, Bilker WB, Barosso GM, Mangaali MJ, et al. The quality of Medicaid and Medicare data obtained from CMS and its contractors: implications for pharmacoepidemiology. BMC Health Serv Res. 2017 26;17(1):304.
13. Li Y, Zhu Y, Chen C, Wang X, Choi Y, Henriksen C, et al. Internal validation of Medicaid Analytic eXtract (MAX) data capture for comprehensive managed care plan enrollees from 2007 to 2010. Pharmacoepidemiol Drug Saf. 2018;27(10):1067–76.
14. Centers for Medicare & Medicaid Services. TAF Technical Guidance: How to Use Illinois Claims Data [Internet]. CMS; 2020 [cited 2020 Aug 26]. Available from: https://www.resdac.org/sites/resdac.umn.edu/files/TAF_TechGuide_IL_Claims_0.pdf

The Effects and Patterns among Mobile Health, Social Determinants, and Physical Activity: A Nationally Representative Cross-Sectional Study

Jiancheng Ye[1*], Qianheng Ma[2]
[1] Feinberg School of Medicine, Northwestern University, Chicago, USA; [2] Department of Public Health Sciences, University of Chicago, Chicago, USA

Abstract

Mobile health (mHealth) technologies and applications are becoming more and more accessible. The increased prevalence of wearable and embeddable sensors has opened up new opportunities to collect health data continuously outside of the clinical environment. Meanwhile, wearable devices and smartphone health apps are useful to address the issues of health disparities and inequities. This study aims to identify different characteristics of individuals who use different mHealth technologies (wearable devices and smartphone apps) and explore the effectiveness and patterns of mHealth for impacting physical activities. We found that social determinants are significantly associated with the use of mHealth; mHealth is helping people to exercise more regularly and for a longer time. Smartphone app users are older while wearable device users are younger. Health disparities exist in mHealth use and physical activity level. Social determinants like education and income are associated with mHealth use and physical activity. The integration of passively-tracked patient-generated health data (PGHD) holds promise in increasing physical activities. Physical activity interventions that comprise wearable devices and smartphone apps may be more beneficial, since health goals, data visualization, real-time support and feedback, results interpretation, and group education could be embedded in the integrated "smart system". These findings may be useful for stakeholders like wearable device and smartphone app companies, researchers, health care workers, and public health practitioners, who should work together to design and develop "precision mobile health" products with higher personalized and participatory levels, thus improving the population health.

Introduction

Wearable device

With the development of biomedical sensors and information technologies, there is an increasing number of wearable or portable devices that are available for users to monitor various parameters, such as physiological indicators, sedentary time, energy expenditure, etc.[1,2] With these functions, wearable devices can not only collect and record health-related data but also assist individuals to maintain an active lifestyle. Studies have shown that wearable devices have the potential to increase the level of physical activity in diverse settings even as a stand-alone technique.[3] Wearable devices have also been evolved and connected with smartphones on which all the data can be stored, displayed, and sent to the health care providers. Given the data collected from real-time tracking along with the advanced artificial intelligence (AI) algorithms,[4] they can generate more specific, customized, and tailored responses back to users. These interactive means that increase physical activity levels and help maintain a structured lifestyle are increasingly promising in the information era. Also, most wearable devices are light and smart, without introducing technostress additional and burden to users.

Smartphone Apps

Smartphone apps have become an emerging technology to change people's lifestyles and behaviors, such as motivating physical activities. But there have been contrasting findings in the past studies. For instance, Flores MG et al [5] found that there was no significant association between smartphone app usage and physical activity and weight loss. Another study conducted by Coughlin SS et al[6] found that there was a modest significant association between smartphone app usage and increased physical activity. A systematic review showed some evidence that the positive effects of smartphone apps were more significant over a short period of time[7] (e.g. 3 months). That is, the intervention effect peaked at the beginning but dwindled as time went by. This phenomenon may be caused by the declined engagement with smartphone apps interventions,[8] especially for those without supportive accountability[9] or poor human-machine interaction design.

Physical activity

Physical inactivity is one of the top predictors that are related to acute or chronic disease and even mortality.[10] Without sufficient physical activities, the prevalence of non-communicable diseases is becoming more prevalent.[11] Declined

physical activity level is associated with an increased likelihood of chronic diseases. In addition, sedentary behavior such as sitting for a long time is also associated with poor health status, e.g., poor functional independence. Regular activity can improve cardiorespiratory and muscular fitness, and mental and behavioral health, etc.[12,13] The World Health Organization has proposed global recommendations on physical activity, which suggest adults participating in at least 75 minutes of vigorous physical activity or 150 minutes of moderate physical activity per week.[14] However, over 25% of adults over the world have not achieved the bottom line of the recommendations.[15] Emerging health technologies such as mobile health technologies can be effective strategies to improve the level of regular physical activity, thus achieving this essential global public health goal.

This study has two primary aims. The first is to describe the different characteristics of individuals who owned and used different mHealth technologies (wearable devices and smartphone apps). We consider social determinants and sociodemographic factors such as age, gender, educational level, income, race/ethnicity, marital status, BMI, etc. as important covariates. The second aim is to explore the effectiveness and patterns of mHealth for impacting physical activities controlling for significant factors of social determinants identified in Aim 1.

Methods

Mobile health user groups

The Health Information National Trends Survey (HINTS) is a nationally-representative survey that is administered every year by the National Cancer Institute, which provides a comprehensive assessment of the American public's current access to and use of health information. The HINTS target population is civilian, non-institutionalized adults aged 18 or older living in the United States. The most recent version of HINTS administration is HINTS 5. The 1st round of data collection for HINTS 5 (Cycle 1) was conducted from January 25 through May 5, 2017; the 2nd round (Cycle 2) was conducted from January 26 through May 2, 2018; the 3rd round (Cycle 3) was conducted from January 22 to April 30, 2019.[16] We combined data of 2017 (Cycle 1, N=3,285), 2018 (Cycle 2, N=3,504) and 2019 (Cycle 3, N=5,438) from the HINTS.[17]

Since participants who own wearable devices and smartphone apps may not use them, we just focused on usage rather than ownership. According to whether participants used health-related apps ("Tablet_AchieveGoal") or electronic device to track/monitor health ("OtherDevTrackHealth"), we defined four user groups: (1) participants didn't use apps or devices (No app and No device); (2) participants only used apps (App only); (3) participants only used health tracking devices (Device only); (4) participants used both apps and devices (App and Device).

"OtherDevTrackHealth" in 2017 and 2018 was stated as "Other than a tablet or smartphone, have you used an electronic device to monitor or track your health within the last 12 months? Examples include Fitbit, blood glucose meters, and blood pressure monitors." But in 2019, this question was assumed to be re-stated separately as two mutually exclusive questions: "WearableDevTrackHealth" (In the past 12 months, have you used an electronic wearable device to monitor or track your health or activity? For example, a Fitbit, Apple Watch, or Garmin Vivofit) and "OtherDevTrackHealth2" (In the last 12 months, have you used an electronic medical device to monitor or track your health? For example a glucometer or digital blood pressure device). Only when participants answered "No" in the former question, would they be presented with the latter. Based on the semantic analysis, we assumed "WearableDevTrackHealth" + "OtherDevTrackHealth2" was equivalent to "OtherDevTrackHealth" in Cycle 1 and Cycle 2.

Physical activities related outcomes

In this study, we studied how mhealth (health-related apps and/or health tracking devices) was associated with physical activities. We selected four physical activity outcomes for analysis as they were the common physical activity outcomes across 2017 to 2019: (1) number of days that a participant did exercises of moderate-intensity; (2) minutes for exercises of moderate-intensity per day; (3) minutes for exercises of moderate-intensity per week; (4) number of days that a participant did strength training. Outcome 3 was calculated by multiplying outcomes 1 and 2. Notably, for the Cycle 1 questionnaire in 2017, outcome 2 was listed separately as hours and minutes of exercises of moderate-intensity and thus we recalculated outcome 2 in Cycle 1 using "hour × 60 + minute" so that it could be consistent with outcome 2 in 2018 and 2019.

Sample characteristics

We included sociodemographic characteristics, such as age, self-reported gender, race and ethnicity, income level, education, marital status, and Body Mass Index (BMI) into the analyses. For race and ethnicity, we combined these two variables and defined "Non-Hispanic white", "Non-Hispanic black", "Hispanic", "others". Income level was re-

categorized as ≤ 20,000, 20,000~35,000, 35,000~50,000, 50,000~75,000, and > 75000 USD; education level was re-categorized as less than college (including the post high-school training), some college, college graduate and post-graduate degree; and marital status was re-categorized as single, in marriage or marriage-like relationship, divorced or widow or separated. For BMI, we considered the continuous scale and also the categorized BMI status as "underweight" (BMI < 18.5), "normal weight" (BMI 18.5~25), "overweight" (BMI 25~30), "obese" (BMI > 30).

Statistical analyses

As we combined data from 2017 to 2019, we first examined whether the sample characteristics were distributed differently among three years (Table 1), in which we used analysis of variance (ANOVA) for continuous variables (e.g., age and BMI) and Pearson's Chi-Square test for categorical variables. For the afterward analysis, we used inverse probability weighting (IPW)[18] to adjust for these potential issues of different distribution across years. In the IPW approach, a logistic regression model for the year with all sample characteristics involved as predictors was used to generate corresponding sample weights. The year 2019 was set as the reference level. By applying the resulting sample weights, data from 2017and 2018 could resemble data from 2019.

The following analyses aimed to explore participants' usage of mHealth technologies and thus we excluded 816 participants whose group membership couldn't be verified. We first examined whether each sociodemographic characteristic was distributed differently across four mHealth user groups (Table 2). After the sample size reduction (N=11411), for all the missing variables involved in this analysis, we applied multiple imputations (MI) for 20 iterations to guarantee sufficient imputation efficiency.[19] Especially for categorical characteristics, we specified a discriminant model to impute the categorical variables. Multiple imputations and the post-imputation analyses were implemented using PROC MI and PROC MIANALYZE in SAS. As we applied the imputation approach, we also did sensitivity analyses on the original complete data and there were no significant conflicts in the findings between the complete data and imputed data analyses.

Following imputation, the weighted versions of Poisson models (Table 3) for physical activities and logistic model (Table 4) of predicting wearable device usage were implemented separately by using imputed datasets with sample weights generated from the IPW approach above. We used Rubin's rule[20] to combine these post-imputation analysis results for valid inference because naively averaging the estimates can bring in unreasonably small standard errors that come from the non-missing replicates across imputed datasets. In addition, since outcome 2 was input by self-entry, unrealistic extreme values in outcome 2 and 3 were detected in data screening. We removed those outliers if the input values were greater than the upper 95% quantiles (120 minutes per day for outcome 2 or 630 minutes per week for outcome 3).

Data were managed and analyzed using SAS 9.4 (SAS Institute., Cary, NC). Categorical variables were compared using the chi-square test, and multivariable analysis used logistic regression. The statistical significance threshold was set at less than .05 for 2-sided tests.

Results

There are 12227 respondents in the combined HINTS sample. Table 1 presents the respondents' sociodemographic and health-related characteristics from Cycle 1 to 3. In this pooled population, most respondents were female (58.26%), with a mean age at 56.8 years old, non-Hispanic whites (64.85%), overweight (34.44%), having high school or lower degree (32.46%), married or marriage-like (53.40%), having a higher annual income (36.65%). The mean BMI is 28.50, which indicates that most respondents need a healthier lifestyle. There are no significant differences among the three cycles of HINTS across the three years except for marital status ($p<0.01$).

Table 1. Characteristics of the population by year

Demographic characteristics	**2017 (n=3285)**	**2018 (n=3504)**	**2019 (n=5438)**	**Pooled (n=12227)**	**P-value**
Age (years), mean ± SD (n)	56.34±16.14	57.02±16.73	56.93±16.89	56.80±16.65	0.19
Missing (n)	139	87	154	380	
Gender (n)					
Male, n (%)	1254(41.28)	1310(40.65)	2108(42.74)	4672(41.74)	0.14
Female, n (%)	1784(58.72)	1913(59.35)	2824(57.26)	6521(58.26)	
Missing (n)	247	281	506	1034	
Race/Ethnicity, (n)					
Non-Hispanic White, n (%)	1929(65.32)	2046 (64.93)	3129 (64.52)	7104(64.85)	0.97

Non-Hispanic Black, n (%)	427(14.46)	457(14.50)	701 (14.45)	1585(14.47)	
Hispanic, n (%)	418(14.16)	458(14.54)	730(15.05)	1606(14.66)	
Other, n (%)	179(6.06)	190(6.03)	290 (5.98)	659(6.02)	
Missing (n)	332	353	588	1273	
BMI (kg/m^2), mean ± SD (n)	28.48±6.65	28.51±6.98	28.50 ±6.81	28.50± 6.82	0.98
BMI Categories					
BMI ≤ 18.5, n (%)	43 (1.35)	70 (2.06)	97 (1.83)	210(1.77)	0.35
18.5 < BMI ≤ 25, n (%)	1007(31.57)	1027(30.15)	1614(30.53)	3648(30.70)	
25 < BMI ≤ 30, n (%)	1081(33.89)	1184(34.76)	1827(34.56)	4092(34.44)	
BMI > 30, n (%)	1059(33.20)	1125(33.03)	1749(33.08)	3933(33.10)	
Missing (n)	134	133	231	498	
Education					
High school or lower, n (%)	1061(33.35)	1135(32.87)	1672(31.66)	3868(32.46)	0.35
Some college, n (%)	714(22.45)	810(23.46)	1199(22.70)	2723(22.85)	
College graduate, n (%)	828(26.03)	910(26.35)	1402(26.55)	3140(26.35)	
Post-graduate degree, n (%)	578(18.17)	598(17.32)	1008(19.09)	2184(18.33)	
Missing (n)	104	51	257	312	
Income					
Income ≤ $ 20K, n (%)	559(18.87)	579(18.76)	904(18.84)	2042(18.83)	0.66
$ 20K< Income ≤ $ 35K, n (%)	423(14.28)	428(13.86)	614(12.80)	1465(13.51)	
$ 35K< Income ≤ $ 50K, n (%)	386(13.03)	404(13.09)	630(13.13)	1420(13.09)	
$ 50K< Income ≤ $ 75K, n (%)	530(17.89)	567(18.37)	848(17.67)	1945(17.93)	
$ 75K< Income, n (%)	1064(35.92)	1109(35.92)	1802(37.56)	3975(36.65)	
Missing (n)	323	417	640	1380	
Marital status, (n)					
Single, n (%)	512(16.17)	605(17.54)	882(16.75)	1999(16.82)	<0.01
Married/Marriage-like, n (%)	1751(55.31)	1747(50.65)	2847(54.05)	6345(53.40)	
Separated/divorced/widowed, n (%)	903(28.52)	1097(31.81)	1538(29.20)	3538(29.78)	
Missing (n)	119	55	171	345	

Abbreviations: BMI, Body mass index.

In order to explore participants' usage of mHealth technologies, Table 2 presents the characteristics of the four mHealth user groups. There were 2515 respondents (22.0%) who reported having used both wearable devices and health apps, 2224 (19.5%) only used health apps, and 1518 (13.3%) only used wearable devices. More than one-third of respondents (45.2%) neither used a wearable device or a health app. There are statistically significant differences across the four user groups ($p<0.01$).

In the "No app and No device" group, their mean age was 59.11 years old; most of them were female (56.45%), non-Hispanic whites (66.59%), and married or marriage-like (48.79%); BMI was overweight (34.01%); having a high school or lower degree (38.31%) and higher annual income (29.12%); the mean BMI was 28.10. The "App only" group was the oldest group with a mean age of 63.71 years old; most of them were female (52.16%), non-Hispanic whites (68.60%), and married or marriage-like (54.85%); BMI was obese (38.53%); having high school or lower degree (34.46%) and a higher annual income (33.61%); the mean BMI was 29.28. In the "Device only" group, their mean age was 46.26 years old; most participants were female (63.55%), non-Hispanic whites (60.35%), and married or marriage-like (59.19%); BMI was overweight (34.72%); having a college or graduate degree (31.89%) and a higher annual income (44.25%); the mean BMI was 28.04. Within the "App and Device" group, their mean age was 49.26 years old; most of them were female (63.01%), non-Hispanic whites (64.14%), and married or marriage-like (63.66%); BMI was obese (36.01%); having a college or graduate degree (35.32%) and a higher annual income (55.31%); the mean BMI was 29.00.

Among the four mHealth user groups, respondents in the "Device only" group were the youngest and healthiest, their age and BMI were the lowest. Respondents who used wearable devices were more likely to have higher education levels since most respondents in both the "Device only" group and "App and Device" group had at least a college degree. Respondents who used apps were more likely to have higher BMI. Most respondents in both the "App only" group and "App and Device" group were obese. As shown in Table 2, there were significant differences across the four user groups in terms of age, gender, race/ethnicity, BMI, education, income, marital status when controlling for all the other variables.

Table 2. Characteristics of the four mHealth user groups

Demographic characteristics	No App/Device (N=5154)	App only (N=2224)	Device only (N=1518)	App + Device (N=2515)	P-value
Age (years), mean ± SD	59.11±16.18	63.71±14.17	46.26±15.04	49.26±14.84	<0.01
Missing (n)	188	70	24	36	
Gender, n (%)					<0.01
Male	2031(43.55)	976(47.84)	522(36.45)	880(36.99)	
Female	2633(56.45)	1064(52.16)	910(63.55)	1499(63.01)	
Missing (n)	490	184	86	136	
Race/Ethnicity, n (%)					<0.01
Non-Hispanic White	3054(66.59)	1348(68.60)	863 (60.35)	1524(64.14)	
Non-Hispanic Black	615 (13.41)	262 (13.33)	209 (14.62)	357(15.03)	
Hispanic	660 (14.39)	264(13.44)	259 (18.11)	305 (12.84)	
Other	257(5.60)	91 (4.63)	99(6.92)	190(8.00)	
Missing (n)	568	259	88	139	
BMI (kg/m^2), mean ± SD	28.10±6.82	29.28±6.86	28.04±6.64	29.00±6.66	<0.01
BMI Categories, n (%)					<0.01
BMI ≤ 18.5	102 (2.04)	32(1.48)	22 (1.48)	27 (1.10)	
18.5 < BMI ≤ 25	1670 (33.39)	549(25.45)	513(34.45)	684(27.77)	
25 < BMI ≤ 30	1701(34.01)	745(34.54)	517(34.72)	865(35.12)	
BMI > 30	1528(30.55)	831(38.53)	437(29.35)	887 (36.01)	
Missing (n)	153	67	29	52	
Education, n (%)					<0.01
High school or lower,	1919 (38.31)	747(34.46)	330(22.06)	416(16.77)	
Some college	1138(22.72)	550(25.37)	354(23.66)	530(21.37)	
College graduate	1178(23.52)	510(23.52)	477(31.89)	876(35.32)	
Post-graduate degree	774(15.45)	361(16.65)	335(22.39)	658(26.53)	
Missing (n)	145	56	22	35	
Income, n (%)					<0.01
Income ≤ $ 20K	1022(22.72)	379(19.36)	191 (13.65)	190(8.06)	
$ 20K< Income ≤ $ 35K	733(16.30)	284(14.50)	147(10.51)	178(7.56)	
$ 35K< Income ≤ $ 50K	630(14.01)	278(14.20)	169(12.08)	251(10.65)	
$ 50K< Income ≤ $ 75K	803(17.85)	359(18.34)	273(19.51)	434(18.42)	
$ 75K< Income	1310(29.12)	658(33.61)	619(44.25)	1303(55.31)	
Missing (n)	656	266	119	159	
Marital status, n (%)					<0.01
Single	848(16.98)	268(12.37)	326(21.88)	429(17.34)	
Married/Marriage-like	2436(48.79)	1188(54.85)	882(59.19)	1575(63.66)	
Separated/divorced/widowed	1709(34.23)	710(32.78)	282(18.93)	470(19.00)	
Missing (n)	161	58	28	41	

Abbreviations: BMI, Body mass index.

Poisson models were employed to characterize the physical activity outcomes and compare the four user groups' performance. Except for age that was not statistically significant in the model for days per week in moderate exercises, all of the other covariates in all four models were found to be significant by F-tests. Poisson model estimates shown in Table 3 are in exponential scale, which can be interpreted as the multiple of the estimate of the reference level. For the moderate exercise outcomes, the "App and Device" group performed the best in terms of frequency and length of time spent. The days per week doing moderate exercise for "App and Device" users are 1.24 times the days for "No app and No device" participants. The length of time spent in moderate exercises per day of "App and Device" users is 1.22 times the length of "No app and No device" participants. The length of time for moderate exercises per week of "App and Device" users is 1.27 times the length of "No app and No device" participants. For the frequency of strength training, the "Device only" group performed the best. Days for strength training for "Device only" users are 1.33 times the days for "No app and No device" group users. We speculated that mHealth technologies especially wearable devices helped individuals to do physical activities more regularly and maintain sustainability.

In general, induvial using wearable devices are performing better regarding physical activity. Smartphone apps seem to be less effective in promoting physical activity. Regarding the effect of social determinants on physical activity,

there are significant differences across the four race/ethnicity groups in terms of exercise magnitude and exercise pattern. For example, white respondents had a higher frequency of doing moderate exercise but black respondents spent a longer time for it each day. We found that days of doing moderate exercises in the white population were 0.99 times the days of the black population; the minutes spent for moderate exercises in the black population were 1.01 times the length of white respondents per day and 1.01 times in a week span. Hispanic population exercised in moderate intensity with the longest time for both per day and per week. The black population had the highest frequency of strength training, and it was 1.10 times the days of the white population. The minority population performed better than the white population on physical activity. Other than race/ethnicity, we found that health disparities exist in other social determinants of health.[21] Users with higher education and higher income level were doing more exercise and strength training. Participants maintaining normal weight ($18.5 < BMI \leqslant 25$) had the highest frequency and longest time for moderate exercises. However, for underweight participants ($BMI \leq 18.5$), they did more strength training to gain weight. Therefore, participants with different social determinants had different physical activity preferences (moderate exercise vs. strength training) and patterns (multiple days for short time vs. one day for a long time). Functions of smartphone apps and wearable devices could be further improved to fit in a person's exercise style.

Table 3. Poisson models for physical activities by mHealth user groups, controlling for social determinants

Demographic characteristics	Days for Moderate Exercises in a Week		Minutes for Moderate Exercises per week		Minutes for Moderate Exercises per week		Days for Strength Training in a week	
	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
Age	1.00	(1.00, 1.00)	1.00	(1.00, 1.00)	1.00	(1.00, 1.00)	1.00	(1.00, 1.00)
Gender								
Female vs. Male	0.85	(0.84, 0.88)	0.84	(0.84, 0.86)	0.80	(0.79, 0.82)	0.75	(0.73, 0.78)
Race and Ethnicity								
Non-Hispanic White	Ref		Ref		Ref		Ref	
Non-Hispanic Black	0.99	(0.95, 1.02)	1.01	(0.99, 1.03)	1.01	(0.98, 1.03)	1.10	(1.04, 1.16)
Hispanic	0.96	(0.93, 1.00)	1.05	(1.03, 1.07)	1.02	(1.00, 1.04)	1.07	(1.02, 1.13)
Other	0.87	(0.83, 0.91)	0.99	(0.87, 0.92)	0.85	(0.84, 0.86)	0.90	(0.84, 0.97)
Education								
High school or lower	Ref		Ref		Ref		Ref	
Some college	1.08	(1.04, 1.12)	1.12	(1.08, 1.14)	1.12	(1.09, 1.14)	1.08	(1.02, 1.14)
College graduate	1.12	(1.08, 1.15)	1.15	(1.13, 1.16)	1.16	(1.14, 1.19)	1.16	(1.11, 1.22)
Post-graduate degree	1.15	(1.11, 1.20)	1.16	(1.14, 1.19)	1.20	(1.17, 1.21)	1.22	(1.15, 1.28)
Income								
Income ≤ $ 20K	Ref		Ref		Ref		Ref	
$ 20K< Income ≤ $ 35K	0.99	(0.94, 1.04)	1.04	(1.00, 1.08)	1.03	(0.99, 1.07)	0.90	(0.83, 0.97)
$ 35K< Income ≤ $ 50K	1.04	(0.99, 1.09)	1.16	(1.13, 1.21)	1.11	(1.07, 1.14)	1.02	(0.94, 1.09)
$ 50K< Income ≤ $ 75K	1.12	(1.07, 1.17)	1.20	(1.15, 1.23)	1.21	(1.16, 1.25)	1.08	(1.01, 1.16)
$ 75K<Income	1.13	(1.08, 1.19)	1.26	(1.22, 1.30)	1.30	(1.26, 1.34)	1.15	(1.06, 1.23)
Marital status								
Single	Ref		Ref		Ref		Ref	
Married/Marriage-like	0.94	(0.90, 0.98)	0.93	(0.91, 0.95)	0.89	(0.87, 0.90)	0.89	(0.84, 0.93)
Separated/divorced/widowed	0.94	(0.91, 0.98)	0.97	(0.95, 0.99)	0.92	(0.90, 0.94)	0.99	(0.93, 1.05)
BMI Status								
BMI ≤ 18.5	0.93	(0.85, 1.02)	0.85	(0.80, 0.91)	0.84	(0.77, 0.91)	1.09	(0.96, 1.26)
18.5 < BMI ≤ 25 (ref)	Ref		Ref		Ref		Ref	
25 < BMI ≤ 30	0.89	(0.87, 0.91)	0.92	(0.91, 0.93)	0.87	(0.86, 0.88)	0.85	(0.82, 0.89)
BMI > 30	0.67	(0.65, 0.69)	0.74	(0.73, 0.75)	0.64	(0.64, 0.66)	0.71	(0.68, 0.74)
Groups								
No App/Wearable Device	Ref		Ref		Ref		Ref	
App only	1.04	(1.01, 1.07)	1.03	(1.02, 1.05)	1.04	(1.02, 1.05)	1.05	(1.00, 1.11)
Wearable Device only	1.15	(1.12, 1.20)	1.15	(1.14, 1.17)	1.22	(1.20, 1.23)	1.33	(1.26, 1.39)
App and Wearable Device	1.24	(1.20, 1.27)	1.22	(1.21, 1.23)	1.27	(1.26, 1.28)	1.30	(1.23, 1.36)

Abbreviations: Est., Estimate; CI, confidence interval; BMI, Body mass index.

Since we found that people who used wearable devices had a more significantly higher level of physical activity, we further examined the patterns and characteristics of wearable device usage. We divided the sample into two groups based on whether the respondents had used wearable devices. Table 4 demonstrates the characteristics that are significantly associated with being a wearable device user. We found that females (OR 1.48, 95% CI 1.35-1.62) and black respondents (OR 1.34, 95% CI 1.18-1.65) were more likely to use wearable devices, which aligns with the finding from Table 2. Respondents with higher education and higher income were more likely to use wearable devices. Respondents who were married or near marriage (OR 1.39, 95% CI 1.23-1.58) were more likely to use wearable devices compared to those of another marital status, which was probably because the household financial burden for them was lower and the cost of using device could also be lower due to the family sharing. Interestingly, respondents who had higher BMIs were more likely to use wearable devices, however, since we don't have follow-up data, we could not specify whether wearable devices helped them to reduce the BMI level.

Table 4. Characteristics significantly associate with being a wearable device user

Demographic characteristics	Est.	Odds Ratio	95% CI
Age (years)*	-0.05	0.95	(0.95,0.96)
Gender*			
Female vs. Male	0.39	1.48	(1.35,1.62)
Race and Ethnicity*			
Non-Hispanic White	Ref		
Non-Hispanic Black	0.29	1.34	(1.18, 1.52)
Hispanic	0.05	1.05	(0.92, 1.20)
Other	0.26	1.30	(1.09, 1.55)
Education*			
High school or lower	Ref		
Some college	0.40	1.50	(1.32,1.70)
College graduate	0.54	1.71	(1.51,1.93)
Post-graduate degree	0.72	2.05	(1.78, 2.35)
Income*			
Income ≤ $ 20K			
$ 20K< Income ≤ $ 35K	0.20	1.22	(1.02,1.46)
$ 35K< Income ≤ $ 50K	0.40	1.49	(1.25,1.78)
$ 50K< Income ≤ $ 75K	0.61	1.84	(1.56,2.18)
$ 75K<Income	0.91	2.48	(2.11,2.92)
Marital status*			
Single			
Married/Marriage-like	0.33	1.39	(1.23,1.58)
Separated/divorced/widowed	0.17	1.19	(1.03,1.38)
BMI Status*			
BMI ≤ 18.5 (Underweight)	-0.39	0.68	(0.47,0.99)
18.5 < BMI ≤ 25 (Normal)	Ref		
25 < BMI ≤ 30 (Overweight)	0.29	1.33	(1.19,1.48)
BMI > 30, (Obese)	0.36	1.43	(1.28,1.60)

Abbreviations: Est, estimate; CI, confidence interval; BMI, Body mass index. *: P < 0.01

Discussion

Participants who did not use mHealth technologies and applications made up nearly half of the population. There were substantial differences among mHealth users in terms of social determinants. These distinctive characteristics may be useful to tailor health interventions accordingly. Since participants who own wearable devices and smartphone apps may not use them, we just focused on usage rather than ownership. The perceived utility of wearable devices and smartphone apps may facilitate the design and development of personalized features and products. Our findings demonstrate that smartphone app users are older while wearable device users are younger. In terms of social

determinants, health disparities exist in mHealth usage and physical activity level. Social determinants like education and income are associated with mHealth usage and physical activity since poor social determinants result in less likely owning such digital tools. Comprising wearable devices and smartphone apps may be more beneficial if health goals, data visualization, real-time support and feedback, results interpretation, and group education could be embedded into an integrated "smart system".

The increased prevalence of wearable devices and smartphone apps has opened up new opportunities to collect health data continuously outside of the clinical environment.[22] Although people have long been journaling and logging their activities as a means of managing various aspects of health, the influx of low-cost wearables make it possible to track multiple measures (e.g. heart rate, step count) passively in real-time and at high-frequency intervals so that self-report bias and recall bias can be alleviated. The health-related data that is created, recorded, or collected by people with the intention of introducing the information to the clinic to help address their health concerns, has been termed patient-generated health data (PGHD).[23] This data can be captured with little burden on the users, enabling nearly continuous data streams over extended periods of observation. The use of PGHD holds promise in increasing physical activities. Studies have shown that the effects of most behavioral changes are short-term.[24] As trackers, wearable devices can be an effective tool to generate PGHD in real time and in a long time course.[25] The integration of passively-tracked PGHD also affords clinicians more evidence to diagnose and treat illnesses and supports communication with patients for improved treatment adherence and health outcomes. Integrating these data into EHR will assist healthcare professionals in monitoring individuals' health status and providing timely support without introducing large resource expenditures. Accompanied with smartphone apps, wearable devices can better motivate and manage individual health.[26] A systematic review found that when combined with other intervention techniques, such as education or counseling, the improvement of physical activity participation is greater than wearable devices alone.[27] In addition, incentives may be useful to motivate physical activity adherence.[28] These features could be added through connecting the smartphone apps. Furthermore, wearable devices have a greater effect on decreasing the time of sedentary behaviors. Human-machine design and interactions should be utilized to develop strategies which can boost user engagement and aid sustainable effectiveness.[29]

Future research may be directed to understanding the association between the intervention effect size and time course, and studying which features of mhealth can lead to sustainable engagement. Additional efforts should be made to expand the time course of the positive effects. Specifically, software engineers should work with researchers to design and implement features that can maintain user engagement using reimbursement, rewards, competition among friends, etc. Personalized and customized design, tailored information, and regular assessment are also useful strategies to engage users. That being said, smartphone apps should not incorporate too many functions that target multiple health behavior changes, because this kind of design will introduce technostress and pressure to users. Focusing on specific health behavior is more likely to lead to longer-term effects.

We recommend that future iterations of HINTS may consider categorizing the types and functions of wearable devices, medical devices, and smartphone apps in the survey at a more granular level. Because motivations for using these technologies and tools may vary by the specific functions and services, this information would be useful to further investigate the patterns among mobile health, social determinants, and physical activity.

Limitations

For the analyses, we have identified and controlled for the social determinants for physical activities but we didn't explore the interaction between user groups and these covariates and time, which together with stratification analyses could be possible steps in future analyses. We used Poisson models to handle the skewness of the, which was less intuitive for interpretation. In addition, this sample was confronted with missing data regarding physical activity outcomes and covariates. The missing data may not be missing completely at random. Those who didn't respond to questions regarding physical activities may be less active and thus our estimates may be subject to bias. Imputing categorical variables was risky. Therefore, more advanced methods for imputation or weighting are needed to handle the missing data.

This study is the largest and most recent nationally representative study of mobile health usage across 3 cycles (2017-2019) of HINTS. However, because the survey was cross-sectional, we couldn't examine causal inferences among variables. The survey did not specify the types and main functions of wearable devices and apps; some devices were just used for physiological indicators monitoring,[1] and some apps may just focus on mental health, which may not be associated with physical activity. Meanwhile, given the limitations of the dataset, we did not have the information about mHealth usage frequency and duration. Despite these limitations, this study contributes to a better understanding

of the effects and patterns among mHealth, social determinants, and physical activities. These findings may be useful for stakeholders like wearable device and smartphone apps companies, researchers, health care workers, and public health practitioners to work together to design and develop "precision mobile health" products with higher personalized and participatory levels, thus improving the population health.

Conclusion

This nationally representative cross-sectional study incorporates a wide range of participants including different age strata, race and ethnicity, and other important social determinants. We find that social determinants are significantly associated with the use of mHealth and individuals using mHealth have more regular physical activity habits. Physical activity interventions comprising wearable devices and smartphone apps can probably be promising in promoting regular physical activity. This study is a good start for further causal inference analyses between mHealth technologies and physical activity level. All the findings may have clinical and public health relevance as improved physical activity levels can contribute to the population health. Combining the advantages of wearable devices and smartphone apps would be useful to integrate PGHD to EHR and make mobile health generate more benefits to a broader population.

References

1. Ye J, Li N, Lu Y, Cheng J, Xu Y. A portable urine analyzer based on colorimetric detection. *Analytical Methods.* 2017;9(16):2464-2471.
2. Zhang J, Fu R, Xie L, et al. A smart device for label-free and real-time detection of gene point mutations based on the high dark phase contrast of vapor condensation. *Lab on a Chip.* 2015;15(19):3891-3896.
3. de Vries HJ, Kooiman TJ, van Ittersum MW, van Brussel M, de Groot M. Do activity monitors increase physical activity in adults with overweight or obesity? A systematic review and meta‐analysis. *Obesity.* 2016;24(10):2078-2091.
4. Ye J, Yao L, Shen J, Janarthanam R, Luo Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Medical Informatics and Decision Making.* 2020;20(11):1-7.
5. Mateo GF, Granado-Font E, Ferré-Grau C, Montaña-Carreras X. Mobile phone apps to promote weight loss and increase physical activity: a systematic review and meta-analysis. *Journal of medical Internet research.* 2015;17(11):e253.
6. Coughlin SS, Whitehead M, Sheats JQ, Mastromonico J, Smith S. A review of smartphone applications for promoting physical activity. *Jacobs journal of community medicine.* 2016;2(1).
7. Wang Q, Egelandsdal B, Amdam GV, Almli VL, Oostindjer M. Diet and physical activity apps: perceived effectiveness by app users. *JMIR mHealth and uHealth.* 2016;4(2):e33.
8. Schoeppe S, Alley S, Van Lippevelde W, et al. Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity.* 2016;13(1):127.
9. Yardley L, Spring BJ, Riper H, et al. Understanding and promoting effective engagement with digital behavior change interventions. *American journal of preventive medicine.* 2016;51(5):833-842.
10. Brown WJ, Bauman AE, Bull F, Burton NW. Development of Evidence-based Physical Activity Recommendations for Adults (18-64 years). Report prepared for the Australian Government Department of Health, August 2012. 2013.
11. Reiner M, Niermann C, Jekauc D, Woll A. Long-term health benefits of physical activity–a systematic review of longitudinal studies. *BMC public health.* 2013;13(1):1-9.
12. Warburton DE, Nicol CW, Bredin SS. Health benefits of physical activity: the evidence. *Cmaj.* 2006;174(6):801-809.
13. Ye J. Pediatric Mental and Behavioral Health in the Period of Quarantine and Social Distancing With COVID-19. *JMIR pediatrics and parenting.* 2020;3(2):e19867.
14. World Health Organization. *Global recommendations on physical activity for health.* World Health Organization; 2010.
15. World Health Organization. *Global action plan on physical activity 2018-2030: more active people for a healthier world.* World Health Organization; 2019.
16. Finney Rutten LJ, Blake KD, Skolnick VG, Davis T, Moser RP, Hesse BW. Data resource profile: The national cancer institute's health information national trends survey (HINTS). *International journal of epidemiology.* 2020;49(1):17-17j.

17. Nelson D, Kreps G, Hesse B, et al. The health information national trends survey (HINTS): development, design, and dissemination. *Journal of health communication.* 2004;9(5):443-460.
18. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine.* 2015;34(28):3661-3679.
19. Rubin DB. Multiple imputation after 18+ years. *Journal of the American statistical Association.* 1996;91(434):473-489.
20. Royston P. Multiple imputation of missing values. *The Stata Journal.* 2004;4(3):227-241.
21. McHugh M, Ye J, Maechling CR, Holl JL. Anchor Businesses in the United States. 2020.
22. Ye J. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. *JMIR Medical Informatics.* 2020;8(7):e19866.
23. Shapiro M, Johnston D, Wald J, Mon D. Patient-generated health data. *RTI International, April.* 2012.
24. Cradock KA, ÓLaighin G, Finucane FM, Gainforth HL, Quinlan LR, Ginis KAM. Behaviour change techniques targeting both diet and physical activity in type 2 diabetes: A systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity.* 2017;14(1):1-17.
25. Bradley SM. Use of Mobile Health and Patient-Generated Data—Making Health Care Better by Making Health Care Different. *JAMA Network Open.* 2020;3(4):e202971-e202971.
26. Lyons EJ, Lewis ZH, Mayrsohn BG, Rowland JL. Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis. *Journal of medical Internet research.* 2014;16(8):e192.
27. Gal R, May AM, van Overmeeren EJ, Simons M, Monninkhof EM. The effect of physical activity interventions comprising wearables and smartphone applications on physical activity: a systematic review and meta-analysis. *Sports medicine-open.* 2018;4(1):42.
28. Finkelstein EA, Haaland BA, Bilger M, et al. Effectiveness of activity trackers with and without incentives to increase physical activity (TRIPPA): a randomised controlled trial. *The lancet Diabetes & endocrinology.* 2016;4(12):983-995.
29. Ye J, Zhang R, Bannon JE, et al. Identifying Practice Facilitation Delays and Barriers in Primary Care Quality Improvement. *The Journal of the American Board of Family Medicine.* 2020;33(5):655-664.

Brain Atlas Guided Attention U-Net for White Matter Hyperintensity Segmentation

Zicong Zhang, MS[1], Kimerly Powell, PhD[2,3], Changchang Yin, MS[1], Shilei Cao, MS[4], Dani Gonzalez[5], Yousef Hannawi, MD[6,*], Ping Zhang, PhD, FAMIA[1,2,*]
[1]Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA
[2]Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA
[3]Department of Radiology, The Ohio State University, Columbus, Ohio, USA
[4]Tencent Jarvis Lab, Tencent, Shenzhen, China
[5]Biomedical Engineering, The Ohio State University, Columbus, Ohio, USA
[6]Department of Neurology, The Ohio State University, Columbus, Ohio, USA
[*]Corresponding authors: yousef.hannawi@osumc.edu; zhang.10631@osu.edu

Abstract

White Matter Hyperintensities (WMH) are the most common manifestation of cerebral small vessel disease (cSVD) on the brain MRI. Accurate WMH segmentation algorithms are important to determine cSVD burden and its clinical consequences. Most of existing WMH segmentation algorithms require both fluid attenuated inversion recovery (FLAIR) images and T1-weighted images as inputs. However, T1-weighted images are typically not part of standard clinical scans which are acquired for patients with acute stroke. In this paper, we propose a novel brain atlas guided attention U-Net (BAGAU-Net) that leverages only FLAIR images with a spatially-registered white matter (WM) brain atlas to yield competitive WMH segmentation performance. Specifically, we designed a dual-path segmentation model with two novel connecting mechanisms, namely multi-input attention module (MAM) and attention fusion module (AFM) to fuse the information from two paths for accurate results. Experiments on two publicly available datasets show the effectiveness of the proposed BAGAU-Net. With only FLAIR images and WM brain atlas, BAGAU- Net outperforms the state-of-the-art method with T1-weighted images, paving the way for effective development of WMH segmentation. Availability: `https://github.com/Ericzhang1/BAGAU-Net`

Introduction

Cerebral Small Vessel Disease (cSVD) is a major public health burden leading to vascular cognitive impairment, intracerebral hemorrhage, and acute ischemic stroke[27]. Histologically, the white matter in patients with cSVD exhibits areas of demyelination and pallor that are seen as white matter hyperintensities (WMH) on the brain magnetic resonance imaging (MRI)[4,5]. A higher volume of WMH is thought to represent a higher burden of cSVD. Indeed, WMH volume load has been suggested as a potential biomarker for cSVD burden. Clinical studies have shown that a larger WMH volume are associated with cognitive impairment, worse stroke functional outcome, and worse response to acute stroke therapy[23–26]. Hence, practical methods for accurate segmentation of WMH to determine its volume are currently needed. Manual tracing WMH in MRI brain images is currently the accepted method for segmenting WMH on brain MRI images. However, it is time-consuming and may be associated inter-observer variability which limits its applicability in daily clinical practice and real time decision making in patients with acute stroke.

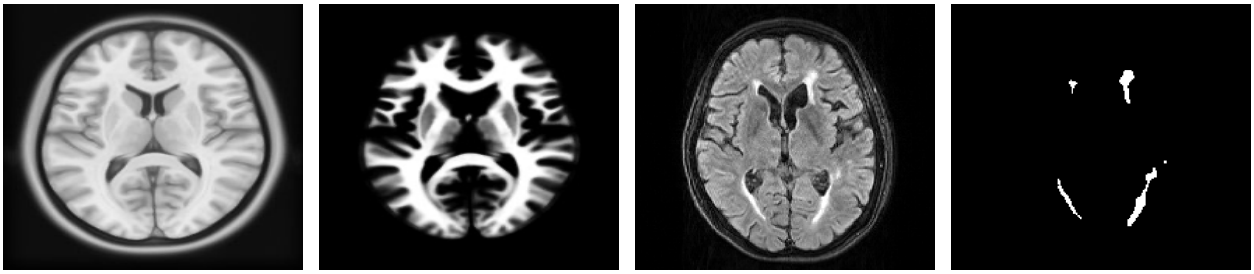
Deep neural networks have shown robust performance in WMH segmentation tasks compared to manual segmentation[9–11,19]. Most of these methods make use of T1-weighted and FLAIR imaging sequences acquired for each patient without considering their clinical usage. In general, the T1-weighted image provides the spatial location information of white matter (WM) that helps guide the WMH segmentation process[20]. Additionally, research MRI scans are often acquired using standardized protocols in a homogenous group of patients that limit their applicability to routine clinical MRI scans. However, in the real-world clinical setting, rapid MRI acquisition in the setting of acute ischemic stroke without T1 sequences have become the normal protocol to allow for critical time sensitive clinical decision making in this setting[21,22]. Hence, there is a critical need to develop a WMH segmentation method based on FLAIR sequences only that can be employed in the treatment of acute stroke patients. Brain atlases have been developed in the past to assist in imaging segmentation by improving preprocessing and image registration and they have been utilized in WMH segmentation as well[28,29]. However, these approaches heavily rely on the presence of T1 sequences for ade-

quate registration and segmentation[30,31]. Recently, atlas-based segmentation was implemented in deep learning, which is either (1) employed in a neural network to learn the correspondences between target image and atlas images[18] or (2) used to provide guidance as prior knowledge during model training[17]. Motivated by the latter case, we explored the use of publicly available WM brain atlas to guide our WMH segmentation in the absence of corresponding T1-imaging sequence.

In this paper, we propose a novel brain atlas guided attention U-Net (BAGAU-Net) for WMH segmentation to address the primary limitations mentioned above. BAGAU-Net consists of two segmentation paths that take the FLAIR image and the spatially registered WM brain atlas separately to provide accurate segmentation results. The two segmentation paths are combined using two novel attention-based connecting mechanisms. Our contribution in this work is four fold: (1) We propose to use only FLAIR and publicly available WM brain atlas to achieve robust performance on WMH segmentation compared to using T1-weighted image; (2) we propose an end-to-end dual path model called brain atlas guided attention U-Net (BAGAU-Net) that leverage an additional path, namely atlas encoding path, to effectively capture the prior knowledge from WM brain atlas to help improve segmentation performance; (3) we introduce multi-input attention module (MAM) and attention fusion module (AFM) to combine the information from two paths to further improve the segmentation performance; (4) we evaluate our model on two publicly available datasets, the 2017 MICCAI WHM segmentation challenge* and the Aging Brain: Vasculature, Ischemia, and Behavior Study (ABVIB) dataset†. The results show that BAGAU-Net has out-performed previously proposed state-out-the-art method on both datasets.

Method

As aforementioned, T1-weighed images are not acquired frequently for WMH segmentation due to time constraints in most the clinical settings. In contrast, FLAIR images are expected in the treatment of acute stroke patients, but does not possess detailed information as T1-weighted image. To relieve such a dilemma, we propose to explore extra prior knowledge to improve the performance with only FLAIR images to match the performance with T1-weighted images. Inspired by the classical concepts of atlas-based segmentation, we propose to exploit prior knowledge hidden in a WM brain to guide the segmentation of the FLAIR images. To this end, we introduce the BAGAU-Net architecture for WMH segmentation. We first describe the generation process of standard WM brain atlas. Then we go into details about the dual-path architecture, including (1) the segmentation path, (2) the atlas encoding path, (3) the multi-input attention module (MAM), and (4) the attention fusion module (AFM).



(a) **(b)** **(c)** **(d)**

Figure 1: Examples of MR images, from left to the right are: (a) T1-weighted ICBM152 atlas, (b) corresponding WM probability atlas, (c) FLAIR image, (d) manual WMH segmentation

Atlas Generation

Brain atlases are often used to identify neuroanatomical structures of the brain. We used the ICBM152 (2009c Nonlinear Symmetric, $1 \times 1 \times 1$ in mm)‡ and the corresponding WM brain atlas[5,6] for our model. *Elastix*[3,4] was used to spatially register the T1-weighted ICBM152 atlas to each of our target images with the following parameters: A multi-resolution pyramid with three levels using advanced mattes mutual information, standard gradient descent optimizer,

*https://wmh.isi.uu.nl
†https://ida.loni.usc.edu/
‡http://www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009

maximum number of iterations 2,000, and a third order B-spline interpolator. The corresponding ICBM152 WM brain atlas was then transformed to match the target image. An example of spatially registered ICBM152 T1-weighted and WM brain atlas, target FLAIR and manually segmented WMH are shown in Fig. 1.

Model Architecture

Convolution neural networks have shown robust performance in many segmentation tasks. As inspired by Li *et al.*[9], we proposed brain atlas guided attention U-Net (BAGAU-Net) that consists of two separate encoding-decoding paths. As shown in Fig 2, the upper path is a U-Net like architecture designed to extract semantic information from the image itself. The lower path is the atlas encoding path where the spatially registered atlas image is input to help guide the decoding process in the segmentation path. Moreover, we designed a multi-input attention module (MAM) and attention fusion module (AFM) to effectively combine the information between the two paths during the decoding process of segmentation path based on the attention gate (AG) as introduced by Oktay *et al.*[12]. The overview of the proposed architecture can be viewed in Fig. 2.

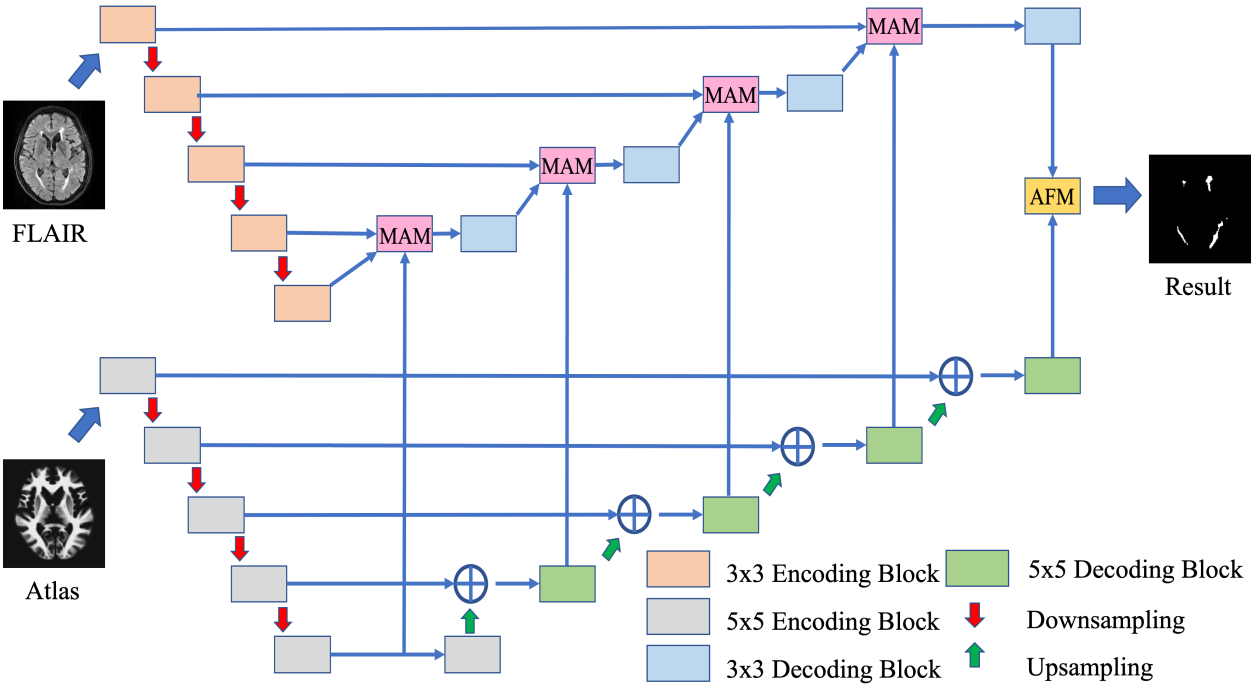
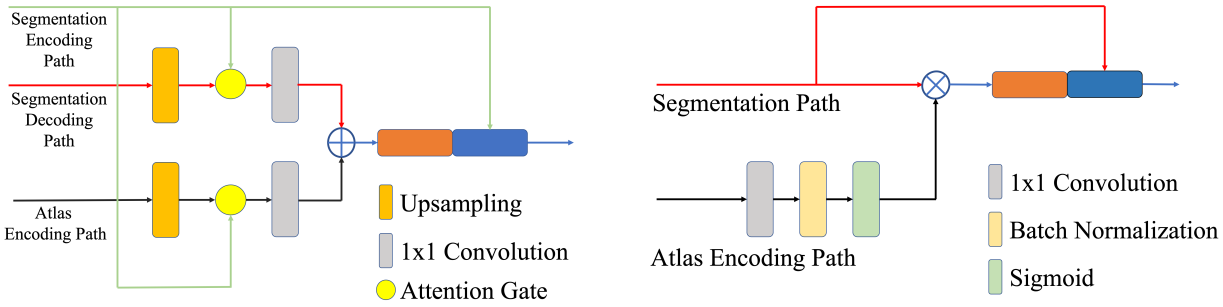


Figure 2: Overview of the BAGAU-Net for WMH segmentation. The model consists of 3x3, 5x5 encoding and decoding block, multi-input attention module (MAM), attention fusion module (AFM), upsampling, and pooling operation.

Segmentation Path

The segmentation path is designed to extract semantic features from the MRI scans. We adopt a U-Net like architecture that consists of four consecutive encoding blocks with feature channel equaling to $64, 96, 128, 256, 512$ followed by four decoding blocks connected by skip connections. Each encoding block consists of two consecutive 3×3 convolutions followed by batch normalization and a rectified linear unit (ReLU). The max-pooling operation is applied for the down-sampling process to extract high-level features. Each decoding block are constructed in a similar fashion as the encoding architecture, except that an up-sampling operation is applied at the end of the block.



(a) The structure of Multi-input Attention Module (MAM) **(b)** The structure of attention fusion module (AFM)

Figure 3: The structure of components in BAGAU-Net

Atlas Encoding Path

The atlas encoding path is used to encode spatial information from atlas images and serves as supplementary features during the decoding process of the segmentation path. We have adopted a similar architecture as our segmentation path with two modifications: (1) each 3×3 convolution in substitute with 5×5 convolution as inspired by Peng *et al.*[13] to capture multi-scale features. (2) we use addition instead of concatenation as inspired by Long *et al.*[14] to combine lower and higher level features.

Multi-input Attention Module

We implemented an attention mechanism that takes multi-input to compute target-aware features. Attention Gate, as proposed by Oktay *et al.*[12], is an efficient way of extracting salient features and contextual information, which allows the model to learn to focus on a subset of target features. $g_i \in \mathbb{R}^{Fg}$ a gating vector and $x_i \in \mathbb{R}^{Fx}$ be the input to the attention gate. The attention gate computes the attention signal $\alpha_i \in [0, 1]$ using addition attention, which is as:

$$q_{att} = W^{att}(\sigma(W_x^T x_i + W_g^T g_i + b_g)) + b_{att},$$
$$\alpha_i = \sigma(q_{att}(x_i, g_i; \theta_{att})).$$

As shown in Fig. 3 (a), MAM takes input from both segmentation path and atlas encoding path at each level and computes the combined feature based on the re-weighted summation of results computed by the attention gate.

Attention Fusion Module

Given the output from the last layer of both the segmentation path and atlas encoding path, we implemented a channel-attention mechanism to refine the feature of the last layer. The output from the atlas encoding path is used to compute the attention mask, where, as shown in Fig. 3 (b), the result is then fused with the output from the segmentation path using element-wise multiplication. A final convolution layer is applied to produce the segmentation result based the concatenation of the fused features.

Experiment

Datasets

The 2017 MICCAI WMH segmentation challenge dataset is publicly available. The dataset consists of 60 training subjects collection from 3 different types of scanners. A more detailed description of the challenge dataset can be found in Table. 1. For each subject, a FLAIR, T1, and the ground truth (manual segmentation of WMH) are provided. All images were bias-corrected using SPM12. We performed image registration through *Elastix*. The ABVIB dataset is a publicly available dataset originally established to study aging brains and cognitive consequences secondary to cardiovascular risk factors. 30 FLAIR subject images were selected and bias-corrected without associated

Table 1: Overview of the 2017 MICCAI WMH segmentation challenge dataset

Institute	Scanner	Size
UMC Utrecht	3 T Philips Achieva	20
NUHS Singapore	3 T Siemens TrioTim	20
VU Amsterdam	3 T GE Signa HDxt	20

T1 sequences. Manual segmentation of WMH for ground truth was performed by a trained student and independently verified by two investigators including a board certified neurointensivist and neurologist.

Model Training and Implementation Details

We split both WMH challenge and ABVIB dataset into training set, validation set and testing set with a ratio of 8:1:1. Instead of using DSC loss for segmentation, we adopt the Tversky loss metric[15] for all model training to provide more balanced weighting between false positives (FPs) and false negatives (FNs). Let P and G be the predict and true label correspondingly. The Tversky loss can expressed as:

$$T(P, G; \alpha) = \frac{|PG|}{|PG| + \alpha|P \setminus G| + (1-\alpha)|G \setminus P|},$$

where α in the above equation is a hyper-parameter that controls the trade off between FP and FN. We set the value of α to 0.7 based on the results of extensive search by Salehi *et al.*[15]. We use Adam optimizer with learning rate of 0.0002 and batch size of 32 for all models; the number of epochs is set to 200. To demonstrate the effectiveness of the atlas encoding path, we compared our proposed model to the winning team's solution[9] as well as the same architecture but simply adding the atlas knowledge as a third channel input. We implemented our model using Pytorch platform[7] and trained the model on single Nvidia Volta V100 GPU with 16GB memory. We adopt gradient accumulation when dealing with out of memory problems. We employed data augmentation to both datasets, including mirroring, rotation, shearing and scaling. Gaussian normalization is further applied in addition to bias-correction to reduce variation across images. A threshold of 0.5 was applied to transform the output of the model into a binary segmentation map.

Model Evaluation

The results of the BAGAU-Net segmentation were evaluated both by visual comparison of the automated segmentation and the manually segmentations and by using quantitative metrics. Four quantitative metrics were used for the quantitative assessment: Dice coefficient (DSC), average volume difference (AVD), Recall, and F1 score were calculated for each segmentation. Dice coefficient (DSC) is a measure of the amount of overlap between two segmented structures. The average volume difference measures the total volume difference between ground truth and the predicted segmentations. Recall and F1 were calculated and based on individual number of lesions.

Results and Discussion

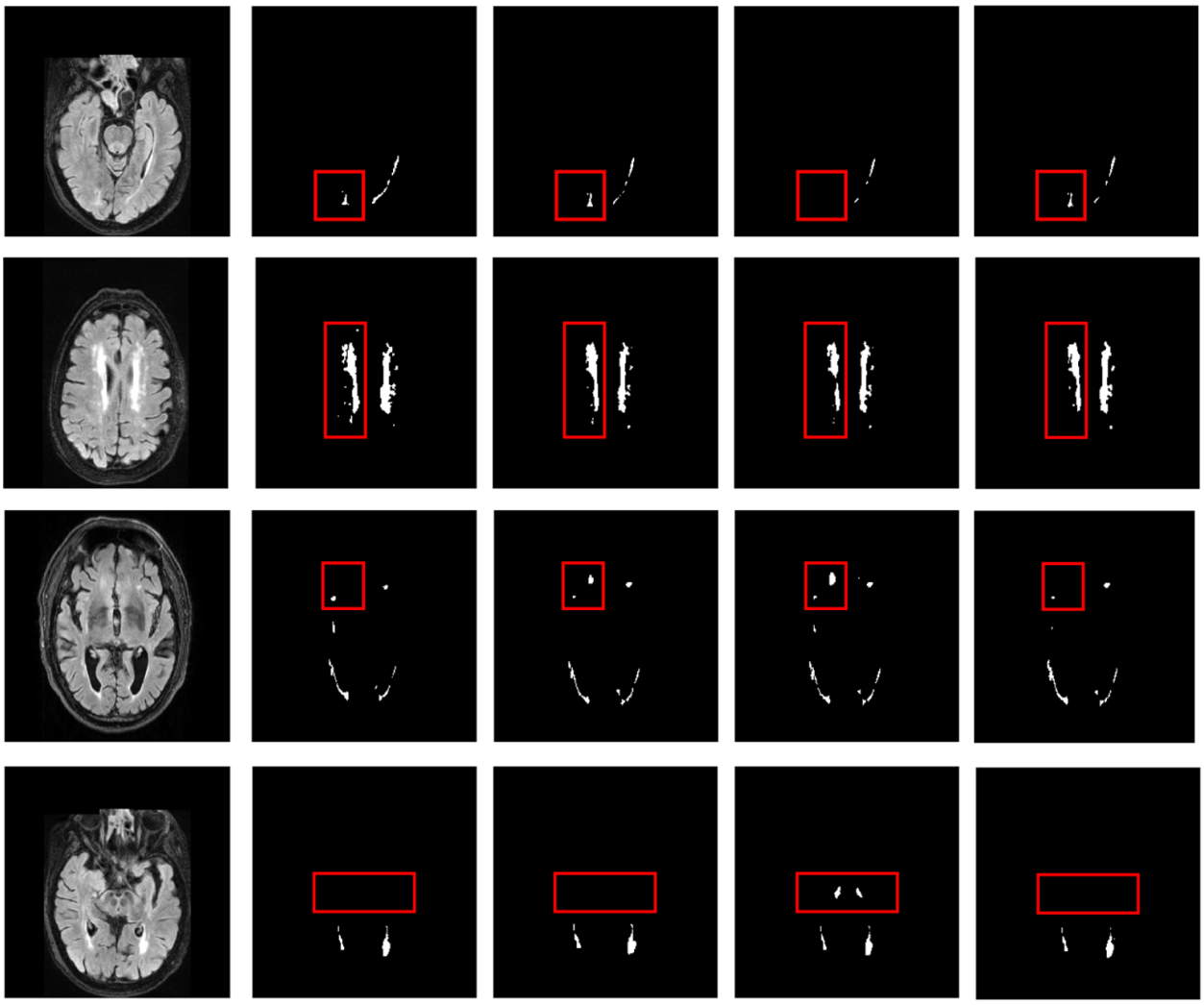
Here we show our evaluation of the proposed BAGAU-Net on the 2017 MICCAI WMH segmentation challenge and the ABVIB datasets. A quantitative evaluation of the results is summarized in Table 2 and Table 3 respectively. Results showed that incorporate the knowledge of atlas can improve segmentation performance, and our proposed BAGAU-Net yields competitive results by achieving the best performance across all metric comparisons. More importantly, our proposed method has achieved better DSC and AVD scores compared to segmentation performance using standard FLAIR and T1-weighted images, which are important metrics for evaluating WMH segmentation in clinical images. The recall and F1 statistics were lower in the ABVIB data set than in the WMH Challenge data set. This was primarily due to the presence of a number of small WMH lesions in the ABVIB dataset that were not as prevalent in the WMH Challenge dataset. However the DSC and AVD metrics were similar for both data sets, suggesting that the same WMH lesion load for the whole brain was segmented correctly using the BAGAU-Net model. This was observed to be true in the visual review.

Table 2: Results of BAGAU-Net and other methods on WMH challenge datasets. ↓ indicates that the smaller the better (0=best, and 100=worst). ↑ indicates that the greater the better (0=worst, and 100=best).

Model	DSC (%) ↑	AVD (%) ↓	Recall (%) ↑	F1 (%) ↑
U-Net[9] (FLAIR + T1)	81.39	16.63	81.41	78.12
U-Net[9] (FLAIR)	80.58	17.13	81.72	76.75
U-Net[9] (FLAIR + Atlas)	80.61	17.99	80.27	76.07
BAGAU-Net	**82.02**	**14.19**	**82.58**	**78.42**

Table 3: Results of BAGAU-Net and other methods on ABVIB datasets.

Model	DSC (%) ↑	AVD (%) ↓	Recall (%) ↑	F1 (%) ↑
U-Net[9] (FLAIR)	77.48	21.18	51.50	56.13
U-Net[9] (FLAIR + Atlas)	76.56	23.87	52.12	56.14
BAGAU-Net	**80.27**	**16.17**	**58.51**	**60.48**



(a) Original **(b)** Groud Truth **(c)** Flair + T1 **(d)** Flair **(e)** BAGAU-Net

Figure 4: Test images from the 2017 WMH challenge dataset (1st column), ground truth (2nd column), the segmentation results obtained by U-Net with FLAIR and T1 (3rd column), FLAIR only (4th) column), our proposed BAGAU-Net (5th column) with FLAIR and WM atlas.

Table 4: Results of ablation study on MAM and AFM on WMH challenge dataset

Model	DSC (%) ↑	AVD (%) ↓	Recall (%) ↑	F1 (%) ↑
BAGAU-Net (without MAM and AFM)	80.52	17.33	81.64	74.19
BAGAU-Net (without MAM)	81.29	15.72	82.13	77.31
BAGAU-Net (without AFM)	80.93	16.19	81.87	76.63
BAGAU-Net	**82.02**	**14.19**	**82.58**	**78.42**

Segmentation results are shown in Fig. 4. It is clear that the absence of T1-weighted images can result in less accurate segmentation results as shown in column 4 (FLAIR only) versus columns 3 (FLAIR + T1). It can be observed that BAGAU-Net, using only FLAIR and WM brain atlas, yield similar segmentation results compared to segmentation results using FLAIR and T1-weighted image as shown in Fig. 4 column 5. Both experimental results and visual comparison showed that BAGAU-Net can effectively capture useful information from WM brain atlas and use it as prior guidance to yield more accurate segmentation results. Compared to the state-of-the-art method developed for WMH segmentation that requires the presence of T1-weighted image, our proposed method can provide even better results with only FLAIR and publicly available WM brain atlas. Furthermore, our model performs better compared to a recent review of the automated methods for WMH segmentation in which the calculated average DSC across studies was 0.73[20].

To illustrate the effectiveness of different components in BAGAU-Net, we perform ablation studies on the WMH challenge datasets as shown in Table 4. For our baseline method, we took out both MAM and AFM from the model and replace it with simple concatenation. Compared to our baseline, we observed that adding either attention modules improves segmentation performance across all metrics. Moreover, including both MAM and AFM further improves the results.

Model Limitation

The main limitation of our model is its application in a relatively small sample size of patients with cSVD (total sample size is 90 subjects). However, more importantly, to address this limitation and increase the external validity of our model, our study subjects were derived from two different datasets (MICCAI and ABVIB). Furthermore, we have only utilized FLAIR sequences in this study to simulate real life scans in this study. The second limitation of our model is that we did not segment separately other lesion types of cSVD such as lacunes since lacunes may have different clinical and prognostic significance than WMH. Our aim in this study, however, was to first develop a successful paradigm for WMH since it is the most prevalent type of cSVD. Current paradigms are under development to subsequently segment lacunes within WMH regions. Finally, our current model was developed in cohorts of patients without other acute or chronic brain etiologies such as acute stroke or brain tumors which can introduce further heterogeneities in brain MRI scans in different clinical settings. The aim of this study, however, was to focus on the method development for combining brain atlas knowledge with deep learning for WMH alone. We will subsequently validate this approach in patients with other brain pathologies.

Conclusion and Future Work

In this paper, we proposed brain atlas guided attention U-Net to improve performance on WMH segmentation. Our model combined domain knowledge of WM brain atlas with deep convolution neural networks by introducing the segmentation path and the atlas encoding path. The proposed MAM and AFM mechanisms were applied at each decoding steps and final prediction step accordingly to help to capture more comprehensive features as well as increase model interpretability by incorporating prior knowledge. We showed that WMH segmentation with T1-weighted image could be replaced and even improved by using publicly available WM brain atlas with our proposed BAGAU-Net. Results showed that our proposed model has out-performed state-out-of-the-art for WMH segmentation on both datasets. Datasets used in this study were collected from elderly group (>65), but the brain atlas we used were generated from a younger group (44±7) (currently, there is no public available brain atlas that focus on aging groups). Ventricles and sulci spaces in aging brains are generally larger than that of young brains[16]. Hence, our model can be further improved by the implementation of an aging white matter brain atlas.

Acknowledgements

This work was funded in part by the National Science Foundation (grant number CBET-2037398) and the National Center for Advancing Translational Research (grant number UL1TR002733). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

1. Kuijf HJ, Biesbroek JM, De Bresser J, Heinen R, Andermatt S, Bento M, Berseth M, Belyaev M, Cardoso MJ, Casamitjana A, Collins DL. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. IEEE transactions on medical imaging. 2019 Mar 19;38(11):2556-68.
2. Iglesias JE, Sabuncu MR. Multi-atlas segmentation of biomedical images: a survey. Medical image analysis. 2015 Aug 1;24(1):205-19.
3. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging. 2009 Nov 17;29(1):196-205.
4. Shamonin DP, Bron EE, Lelieveldt BP, Smits M, Klein S, Staring M. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Frontiers in neuroinformatics. 2014 Jan 16;7:50.
5. Collins DL, Zijdenbos AP, Baaré WF, Evans AC. ANIMAL+ INSECT: improved cortical structure segmentation. InBiennial International Conference on Information Processing in Medical Imaging 1999 Jun 28 (pp. 210-223). Springer, Berlin, Heidelberg.
6. Fonov VS, Evans AC, McKinstry RC, Almli CR, Collins DL. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage. 2009(47):S102.
7. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. Pytorch: An imperative style, high-performance deep learning library. InAdvances in neural information processing systems 2019 (pp. 8026-8037).
8. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Springer, Cham.
9. Li H, Jiang G, Zhang J, Wang R, Wang Z, Zheng WS, Menze B. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. NeuroImage. 2018 Dec 1;183:650-65.
10. Li H, Zhang J, Muehlau M, Kirschke J, Menze B. Multi-scale convolutional-stack aggregation for robust white matter hyperintensities segmentation. InInternational MICCAI Brainlesion Workshop 2018 Sep 16 (pp. 199-207). Springer, Cham.
11. Ghafoorian M, Karssemeijer N, Heskes T, van Uden IW, Sanchez CI, Litjens G, de Leeuw FE, van Ginneken B, Marchiori E, Platel B. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Scientific Reports. 2017 Jul 11;7(1):1-2.
12. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. 2018 Apr 11.
13. Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters–improve semantic segmentation by global convolutional network. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 4353-4361).
14. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 3431-3440).
15. Salehi SS, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. InInternational Workshop on Machine Learning in Medical Imaging 2017 Sep 10 (pp. 379-387). Springer, Cham.

16. Dickie DA, Shenkin SD, Anblagan D, Lee J, Blesa Cabez M, Rodriguez D, Boardman JP, Waldman A, Job DE, Wardlaw JM. Whole brain magnetic resonance image atlases: a systematic review of existing atlases and caveats for use in population imaging. Frontiers in neuroinformatics. 2017 Jan 19;11:1.

17. Xu Z, Niethammer M. DeepAtlas: Joint semi-supervised learning of image registration and segmentation. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2019 Oct 13 (pp. 420-429). Springer, Cham.

18. Wickramasinghe U, Knott G, Fua P. Probabilistic Atlases to Enforce Topological Constraints. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2019 Oct 13 (pp. 218-226). Springer, Cham.

19. Wu J, Zhang Y, Wang K, Tang X. Skip Connection U-Net for White Matter Hyperintensities Segmentation From MRI. IEEE Access. 2019 Oct 21;7:155194-202.

20. Caligiuri ME, Perrotta P, Augimeri A, Rocca F, Quattrone A, Cherubini A. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. Neuroinformatics. 2015 Jul 1;13(3):261-76.

21. Thomalla G, Simonsen CZ, Boutitie F, Andersen G, Berthezene Y, Cheng B, Cheripelli B, Cho TH, Fazekas F, Fiehler J, Ford I. MRI-guided thrombolysis for stroke with unknown time of onset. New England Journal of Medicine. 2018 Aug 16;379(7):611-22.

22. Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, McTaggart RA, Torbey MT, Kim-Tenser M, Leslie-Mazwi T, Sarraj A. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. New England Journal of Medicine. 2018 Feb 22;378(8):708-18.

23. Etherton MR, Wu O, Giese AK, Lauer A, Boulouis G, Mills B, Cloonan L, Donahue KL, Copen W, Schaefer P, Rost NS. White matter integrity and early outcomes after acute ischemic stroke. Translational stroke research. 2019 Dec 1;10(6):630-8.

24. Etherton MR, Wu O, Rost NS. Recent advances in leukoaraiosis: white matter structural integrity and functional outcomes after acute ischemic stroke. Current cardiology reports. 2016 Dec 1;18(12):123.

25. van Norden AG, de Laat KF, Gons RA, van Uden IW, van Dijk EJ, van Oudheusden LJ, Esselink RA, Bloem BR, van Engelen BG, Zwarts MJ, Tendolkar I. Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol. BMC neurology. 2011 Dec 1;11(1):29.

26. Arsava EM, Rahman R, Rosand J, Lu J, Smith EE, Rost NS, Singhal AB, Lev MH, Furie KL, Koroshetz WJ, Sorensen AG. Severity of leukoaraiosis correlates with clinical outcome after ischemic stroke. Neurology. 2009 Apr 21;72(16):1403-10.

27. Au R, Massaro JM, Wolf PA, Young ME, Beiser A, Seshadri S, D'Agostino RB, DeCarli C. Association of white matter hyperintensity volume with decreased cognitive functioning: the Framingham Heart Study. Archives of neurology. 2006 Feb 1;63(2):246-50.

28. Oishi K, Faria A, Jiang H, Li X, Akhter K, Zhang J, Hsu JT, Miller MI, van Zijl PC, Albert M, Lyketsos CG. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants. Neuroimage. 2009 Jun 1;46(2):486-99.

29. Wang Y, Catindig JA, Hilal S, Soon HW, Ting E, Wong TY, Venketasubramanian N, Chen C, Qiu A. Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. Neuroimage. 2012 May 1;60(4):2379-88.

30. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. Fsl. Neuroimage. 2012 Aug 15;62(2):782-90.

31. Ashburner J. SPM: a history. Neuroimage. 2012 Aug 15;62(2):791-800.

Recommender system of scholarly papers using public datasets

Jie Zhu, M.S., Braja G. Patra, Ph.D., Ashraf Yaseen, Ph.D.
University of Texas Health Science Center at Houston
Houston, TX, USA

Abstract

The exponential growth of public datasets in the era of Big Data demands new solutions for making these resources findable and reusable. Therefore, a scholarly recommender system for public datasets is an important tool in the field of information filtering. It will aid scholars in identifying prior and related literature to datasets, saving their time, as well as enhance the datasets reusability. In this work, we developed a scholarly recommendation system that recommends research-papers, from PubMed, relevant to public datasets, from Gene Expression Omnibus (GEO). Different techniques for representing textual data are employed and compared in this work. Our results show that term-frequency based methods (BM25 and TF-IDF) outperformed all others including popular Natural Language Processing embedding models such as doc2vec, ELMo and BERT.

Introduction

Recommendation systems, or recommenders, are information filtering systems that employ data mining and analytics of users' behaviors, including preferences and activities, to predict users' interests in information, products or services. There are broadly two types of recommenders: collaborative filtering and content-based. The former works by utilizing the rating activities of items or users, while the latter works by comparing descriptions of items or profiles of users' preferences.

With the ever-growing public information online, recommendation systems have proven to be an effective strategy to deal with information overload. In fact, recommenders are thriving in this era of Big Data with wide commercial applications in recommending products (e.g. Amazon), music[1], movies[2], books[3], news articles[4], and many more.

Applications of recommendation systems are currently expanding beyond the commercial to include scholarly activities. The first recommendation system for research papers was introduced in the CiteSeer project[5]. Following that, Science Concierge[6], PURE[7], pmra[8] were also developed for recommending articles. More recent experiments include Colin and Beel's[9] and A. Mohamed Hassan et al.'s[10], in which they experimented with Natural Language Processing (NLP) models.

The aforementioned systems are all paper-to-paper recommendations, i.e., they provide recommendations of papers similar to a given paper. To date, no prior research has yet been performed on recommending papers based on public datasets, to the best of our knowledge. There are many public datasets available on the internet which might be useful to researchers for further exploration. A scholarly literature recommendation system for datasets is an important and very helpful tool in the field of information filtering. It can aid in identifying prior and related literature to the dataset's topic. It can save researchers' time as well as enhance the experience of the dataset's re-usability. Further, recommending literature to datasets is also a field of research yet to be explored.

In this paper, we described the development of a content-based recommendation system that recommends articles from PubMed corresponding to datasets (referred to as data series) from Gene Expression Omnibus (GEO). GEO is a public repository for high-throughput microarray and next-generation sequence functional genomics data. As of Feb 05, 2020, there are 124,825 data series available in GEO (A series record links together a group of related samples and provides a focal point and description of the whole study[11]). Many of these series' data were collected at enormous effort only to be used just once. We believe that dataset use and reuse can be significantly improved when recommending research papers that are relevant to such dataset to researchers, an idea consistent with NIH Strategic Plan for Data Science[12]. We experimented and compared a variety of vector representations from traditional term-frequency based methods and topic-modeling to embeddings, and evaluated different recommendations using existing citations as a reference. The work described herein is part of the dataset re-usability platform (GETc Research Platform) developed at The University of Texas Health Science Center at Houston available at http://genestudy.org.

Relevant work

CiteSeer[5] is a content-based recommender based on keywords matching, Term Frequency-Inverse Document Frequency (TF-IDF) for word information and Common Citation-Inverse Document Frequency (CCIDF) for citation information. Science Concierge[6] is another content-based research article recommendation system using Latent

Semantic Analysis (LSA) and Rocchio Algorithms with large-scale approximate nearest neighbor search based on ball trees. PURE[7], another content-based PubMed article recommender developed using a finite mixture model for soft clustering with Expectation-Maximization (EM) algorithm, which achieved 78.2% precision at 10% recall with 200 training articles. Lin and Wilbur developed pmra[8], a probabilistic topic-based content similarity model for PubMed articles. Their method achieved slight but statistically significant improvement on precision@5 compared to BM25.

With the popularity of NLP models, such as Google's doc2vec, USE, and most recently BERT, there have been some efforts in incorporating these embedding methods in research papers recommenders. Colin and Beel[9] experimented with doc2vec, TF-IDF and key phrases for providing related-article recommendations to both digital library Sowiport[13] and the open-source reference manager Jabref[14]. A. Mohamed Hassan et al.[10] evaluated USE, InferSent, ELMo, BERT and SciBERT for reranking results from BM25 for research paper recommendations.

Materials

We used data series from GEO and MEDLINE articles from PubMed. For GEO series, metadata such as title, summary, date of publications and names of authors were collected using a web crawler. We also collected the PMIDs of the articles associated with each series. From these PMIDs, metadata of corresponding articles such as title, abstract, authors, affiliations, MeSH terms, publisher name were also collected. Figure 1 shows an example of GEO data series, Figure 2 shows an example of PubMed publication.

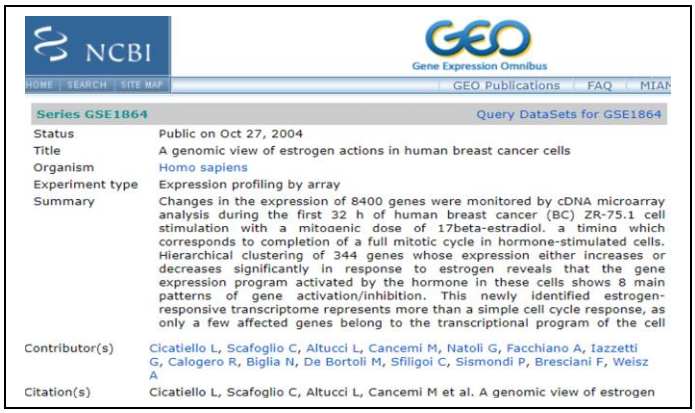


Figure 1. An example of GEO data series

Figure 2. An example of PubMed publication.

In order to automatically evaluate our recommendations, using metrics such as precision and recall, we kept only the series that have associated citations (publications). That left us with a total of 72,971 unique series and 50,159

associated unique publications. Multiple series can reference the same paper(s). 96% of the series have only 1 related publication and the rest have between 2 to 10.

Methods

We adopted an information retrieval strategy, where the data series are treated as queries and the list of recommended publications as retrieved documents. In our experiments, series were represented by their titles and summaries; while publications were represented by their titles and abstracts. Further, we removed stop words, punctuation, and URLs from summaries of series before transforming them into vectors.

We used cosine similarity as the ranking score, which is a popular measure in query-document analysis[15] for the similarity of features due to its low-complexity and intuitional definition. In our case, we only returned the top 10 recommendations based on cosine similarity, which is a realistic scenario where few people would check the end of a long recommendation list. Figure 3 shows our recommender's architecture.

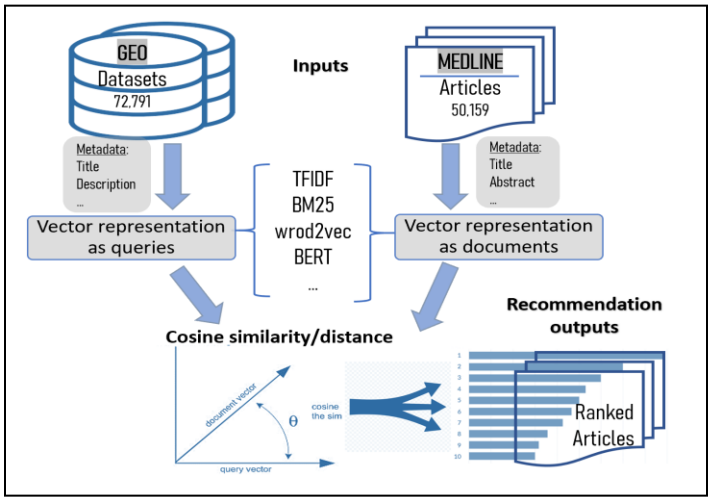


Figure 3. Literature recommendation system architecture

The recommendations were then evaluated using existing series-articles relationships from series metadata using MRR@10, recall@1, recall@10, precision@1, and MAP@10.

Vector representation

Methods of representing textual data in recommendation systems are ranging from traditional term-frequency based methods and topic-modeling to embeddings. Below is the list of methods we experimented with in this study:

TF-IDF: a numerical statistical representation of how important a word is to a document in a collection or corpus[16]. For each vocabulary V, the value increases proportionally to the number of times that V appears in the document (term frequency, TF) and is offset by the total number of documents that contain V (inverse document frequency, IDF). We used TF-IDF implementation from scikit-learn[17].

BM25: a ranking function that is based on a probabilistic retrieval framework that utilizes adjusted values of TF and IDF and document length[18]. We used BM25 implementation from genism[19].

LSA: a topic modeling technique that utilizes singular value decomposition (SVD) on a term-frequency matrix to find a low-rank approximation representation. We used TruncatedSVD from scikit-learn for LSA implementation, with a reduced dimension equals to 300.

word2vec[20,21]: a two-layer neural network that is trained to reconstruct linguistic contexts of words by mapping each unique word to a corresponding vector space. We utilized word2vec implemented in gensim, with an embedding dimension of 200.

doc2vec[21]: a neural network method that extends word2vec and learns continuous distributed vector representations for variable-length pieces of texts. We utilized doc2vec implemented in gensim, with an embedding dimension of 300.

ELMo[22]: a deep, contextualized bi-directional Long Short-Term Memory (LSTM) model that was pre-trained on 1B Word Benchmark[23]. We used the latest TensorFlow Hub implementation[24] of ELMo to obtain embeddings of 1024 dimensions.

InferSent[25]: a bi-directional LSTM encoder with max-pooling that was pre-trained on the supervised data of Stanford Natural Language Inference (SNLI)[26]. There are two versions of InferSent models, and we used one with fastText word embeddings from Facebook's github[27], with the resulting embedding dimension of 4096.

USE[28]: Universal Sentence Encoder, developed by Google, has two variations of model structures: one is transformer-based while the other one is Deep Average Network (DAN)-based, both of which were pre-trained on unsupervised data such as Wikipedia, web news and web question-answer pages, discussion forums, and further on supervised data of SNLI. We used the TensorFlow Hub implementation of transformer USE to obtain embeddings of 512 dimensions.

BERT[29]: Bidirectional Encoder Representations from Transformer developed by Google, which has previously achieved state-of-the-art performance in many classical natural language processing tasks. It was pre-trained on 800M-words BooksCorpus and 2500M-word English Wikipedia using masked language model (MLM) and next sentence prediction (NSP) as the pre-training objectives. We used the package Sentence-BERT[30] to obtain vectors optimized for Semantic Textual Similarity (STS) task, which is of 768 dimensions.

SciBERT[31]: a BERT model that was further pre-trained on 1.14M full-paper corpus from semanticscholar.org[32]. Similarly, we used Sentence-BERT to obtain vectors of 768 dimensions.

BioBERT[33]: a BERT model that was further pre-trained on large scale biomedical corpus, i.e. 4.5B-word PubMed abstracts and 13.5B-word PubMed Central full-text articles. Similar to BERT, vectors of 768 dimensions were obtained using Sentence-BERT.

RoBERTa[34]: a robust version of BERT that has been further pre-trained on CC-NEWs[35] corpus, with enhanced hyperparameters choices including batch-sizes, epochs, and dynamic masking patterns in the pre-training process. We used Sentence-BERT to obtain vectors of 768 dimensions.

DistilBERT[36]: a distilled version of BERT with a 40% reduced size, 97% of the original performance while being 60% faster. We used Sentence-BERT to obtain vectors of 768 dimensions.

For all term-frequency based methods, the experiments were performed on 8 Intel(R) Xeon(R) Gold 6140 CPUs@ 2.30GHz. For embedding based methods, the experiments were performed using 1 Tesla V100-PCIE-16GB GPU. The implementations of the experiments are at https://github.com/chocolocked/RecommendersOfScholarlyPapers

Evaluation metrics

The following metrics were used to evaluate our system:

Mean reciprocal rank (MRR)@k: The Reciprocal Rank (RR) measures the reciprocal of the rank at which the first relevant document was retrieved. RR is 1 if the relevant document was retrieved at rank 1, RR is 0.5 if document is retrieved at rank 2, and so on. When we average the top k retrieved items across queries, the measure is called the Mean Reciprocal Rank@k[37]. In our case, we chose k=10.

Recall@k: At the k-th retrieved item, this metric measures the proportion of relevant items that are retrieved. We evaluated both recall@1 and recall@10.

Precision@k: At the k-th retrieved item, this metric measures the proportion of the retrieved items that are relevant. In our case, we are interested in precision@1. Since most of our data series has only 1 corresponding publication, which means most of the data only has 1 relevant item.

Mean average precision (MAP)@k: Average Precision is the average of the precision value obtained for the set of top k items after each relevant document is retrieved. When average precision is averaged again over all retrieval, this value becomes mean average precision.

Detailed procedure-example

Below, we demonstrate the detailed procedure using BM25 and data series 'GSE11663' as an example:

- For each of the 50,159 publications, we concatenated processed titles with abstracts. We then created a BM25 object, its dictionary and corpus out of the list.
- For 'GES11663', we concatenated the title (*'human cleavage stage embryos chromosomally unstable'*) and the processed summary (*'embryonic chromosome aberrations cause birth defects reduce human fertility however neither nature incidence known develop assess genome-wide copy number variation loss heterozygosity single cells apply screen blastomeres vitro fertilized preimplantation embryos complex patterns chromosome-arm imbalances segmental deletions duplications amplifications reciprocal sister blastomeres detected large proportion embryos addition aneuploidies uniparental isodisomies frequently observed since embryos derived young fertile couples data indicate chromosomal instability common human embryogenesis comparative genomic hybridisation*') and got its vector representation using dictionary: [(27, 1), (32, 1), (44, 1), (46, 1), (80, 1), (116, 1), (141, 1), (175, 1), (182, 1), (190, 1), (360, 2), (390, 1), (407, 1), (530, 1), (649, 1), (663, 1), (725, 1), (842, 1), (844, 1), (999, 1), (1034, 1), (1186, 1), (1235, 1), (1370, 1), (1634, 1), (1635, 1), (1636, 1), (1761, 1), (1862, 1), (2023, 1), (2174, 1), (2224, 1), (2292, 1), (2675, 1), (2677, 1), (3023, 1), (3082, 1), (3113, 1), (3144, 2), (3145, 2), (3153, 1), (3697, 1), (4265, 1), (4935, 1), (5021, 1), (5105, 1), (5775, 1), (6665, 1), (6772, 1), (6828, 1), (7298, 1), (7372, 1), (7684, 1), (7808, 1), (7949, 1), (8211, 1), (8344, 1), (8569, 2), (8974, 1), (9009, 1), (9302, 1), (9705, 1), (10480, 1), (11360, 1), (17139, 1), (24769, 1), (28560, 1), (38594, 1), (54855, 1), (228500, 1), (250370, 1)]. Then we used *sklearn 'cosine-similarity'* to get similarity scores of all 50,159 publications with this series.
- Since 'GSE11663' has the citation ['19396175', '21854607'] (without order), and our top 10 recommendations were ['19396175', '23526301', '16698960', '25475586', '29040498', '23054067', '27197242', '23136300', '24035391', '18713793']. Our recommendations hit top 1. In this case, we calculated

$$MRR@10: \frac{1}{1} = 1, \qquad recall@1: \frac{1}{3} = 0.33,$$

$$recall@10: \frac{1}{3} = 0.33, \qquad precision@1: \frac{1}{1} = 1,$$

$$MAP@10: \frac{1}{2} * (1 + 0) = 0.5.$$

- We repeated the above two steps, and computed average for all 72,971 series

Results

Table 1 shows the results of our experiments with different vector representations. BM25 outperformed all other methods in terms of all evaluation metrics, with MRR@10, recall@1, recall@10, precision@1, and MAP@10 of 0.785, 0.742, 0.833, 0.756, and 0.783 respectively, followed closely by TF-IDF. None of the embedding methods alone was able to outperform BM25. Furthermore, word2vec, doc2vec, and BioBERT were among the top embedding methods outperforming ELMo, USE, and the rest.

Our findings show that traditional term-frequency based methods (BM25, TF-IDF) were more effective for recommendations compared to embedding methods. Contrasting previous beliefs that embeddings can conquer it all, given their performances in standardized general NLP tasks such as sentiment analysis, Questions & Answering (Q&A), and Named Entity Recognition (NER). They failed to show advantage in the simple scenario of capturing semantic similarity as measured by cosine similarity. Even though the context were not exactly the same, Colin and Beel[9] did find out in their studies that doc2vec failed to defeat TF-IDF or key phrases in the two experimental setups of publication recommendations for digital library Sowiport and reference manager Jabref. Moreover, A. Mohamed Hassan et al.[12] also concluded in their study that none of the sentence embeddings (USE, InferSent, ELMo, BERT and SciBERT) that they had employed were able to outperform BM25 alone for their research paper recommendations.

One possible reason could be that traditional statistical methods produce better features when the queries are relatively homogenous, Ogilvie and Callan[38] showed that single database (homogeneous) queries with TF-IDF performed unanimously better than multi-database (heterogenous) queries when no additional IR techniques, such as query expansion, were involved. Currently, we are only using GEO datasets for queries which are all related to gene expressions. But as we introduce more diverse datasets for our platform in the future, e.g. immunology and infectious

Table 1. MRR@10, Recall@1, Recall@10, Precision@1, and MAP@10 for recommenders using different vector representations.

Vector representations		**Metrics**				
		MRR@10	***Recall@1***	***Recall@10***	***Precision@1***	***MAP@10***
Term-frequency based & topic modeling	TF-IDF	0.721	0.655	0.803	0.677	0.719
	BM25	**0.785**	**0.742**	**0.833**	**0.756**	**0.783**
	LSA	0.565	0.518	0.640	0.528	0.564
Embedding based: Not pretrained on NLP tasks	word2vec	0.656	0.615	0.712	0.626	0.655
	doc2vec	0.601	0.562	0.655	0.572	0.600
Embedding based: Context dependent, pretrained on NLP tasks	ELMo	0.364	0.341	0.400	0.346	0.364
	InferSent	0.534	0.502	0.579	0.511	0.534
	USE	0.411	0.377	0.468	0.383	0.411
	BERT	0.503	0.468	0.563	0.476	0.505
	SciBERT	0.435	0.399	0.493	0.406	0.434
	BioBERT	0.540	0.498	0.605	0.507	0.539
	roBERTa	0.509	0.468	0.572	0.476	0.508
	DistilBERT	0.501	0.463	0.558	0.471	0.500

disease datasets, the heterogeneity might require more advanced embedding methods. Further, as we observe approximately 8% improvement from regular BERT to BioBERT, we think it might be of importance for NLP models to be further trained on domain-specific corpus for better feature representations for cosine similarity. Another possible reason could be that, as these embeddings were pre-trained on standardized tasks, thus the embeddings might be specialized towards those tasks instead of representing simple semantic information. This could explain the observation that general text embeddings, e.g. word2vec, doc2vec, perform better than other more specialized NLP models, e.g. ELMo and BERT, which were pre-trained to perform on tasks such as Q&A, sequence classification. Therefore, we might be able to take full advantage of their potentials when formulating our problem from a simple cosine similarity between query and documents to matching classification for example; a format closer to how these models were designed for in the first place. That is also the direction we are heading towards for future experiments.

Even though we do not currently have users' feedback for manual evaluations, we did, however, manually inspect the recommendation results for the completeness of our experiments, particular for those where the cited articles did not appear within the top 5 recommendations. We randomly sampled 20 such data series and examined recommended papers by thoroughly reading through papers' abstract, introduction, and methods. We had some interesting observations regarding those cases: For example, for 'GSE96174' data series, even though our top 5 recommendations did not include the existing related article, three of them actually cited and used the data series as relevant research materials. Another example is that of 'GSE27139', where our top recommendations were from the same author that submitted the data series, and those articles were extensions from their previous research work. Due to time limitation, we could not check all the 13,013 cases, but we found at least 10 cases ('GSE96174', 'GSE836', 'GSE92231', 'GSE78579', 'GSE96211', 'GSE27139','GSE10903', 'GSE105628', 'GSE44657', 'GSE81888') that had similar situations as we mentioned above and where the top 3 recommendations were, to the best of our judgement, associated with data series of concern, even though they did not appear in the citation as of the time we did our experiments. Therefore, we believe that our recommendation systems might do even better in the real setting than the evaluations presented here.

We want to mention that we also experimented with re-ranking. The final ranking score is defined as the previous cosine-similarity adding a re-ranking score, with the re-ranking score calculated using cosine similarity of only titles of the queried dataset and of the articles. We did not find statistically significant improvements, and therefore did not report the results in this paper.

Discussion

In this work, we developed a scholarly recommendation system to identify and recommend research-papers relevant to public datasets. The sources of papers and datasets are PubMed and Gene Expression Omnibus (GEO) series, respectively. Different techniques for representing textual data ranging from traditional term- frequency based methods and topic-modeling to embeddings are employed and compared in this work. Our results show that embedding models that perform well in their standardized NLP tasks, failed to outperform term-frequency based probabilistic methods such as BM25. General embeddings (word2vec and doc2vec) performed better than more specialized embeddings (ELMo and BERT) and domain-specific embeddings (BioBERT) performed better than non-domain specific embeddings (BERT). In future experiments, we plan to develop a hybrid method combining the strengths of the term-frequency approach and also embeddings to maximize their potentials in different (heterogeneous vs. homogeneous) problem scenarios. In addition, we plan to engage users in rating our recommendations, use interrater agreement approach to further evaluate results, and incorporate the feedback to further improve our system. We hope to utilize content-based and collaborative filtering for better recommendations.

Given their usefulness, extending the applications of recommender systems to aid scholars in finding relevant information and resources will significantly enhance research productivity and will ultimately promote data and resources reusability.

References

1. Ali M, Johnson CC, Tang AK. Parallel collaborative filtering for streaming data. University of Texas Austin, Tech. Rep. 2011 Dec 8:5-7.
2. Bell RM, Koren Y. Lessons from the Netflix prize challenge. Acm Sigkdd Explorations Newsletter. 2007 Dec 1;9(2):75-9.
3. Vaz PC, Martins de Matos D, Martins B, Calado P. Improving a hybrid literary book recommendation system through author ranking. InProceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries 2012 Jun 10 (pp. 387-388).
4. Li L, Chu W, Langford J, Schapire RE. A contextual-bandit approach to personalized news article recommendation. InProceedings of the 19th international conference on World wide web 2010 Apr 26 (pp. 661-670).
5. Bollacker KD, Lawrence S, Giles CL. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. InProceedings of the second international conference on Autonomous agents 1998 May 1 (pp. 116-123).
6. Achakulvisut T, Acuna DE, Ruangrong T, Kording K. Science Concierge: A fast content-based recommendation system for scientific publications. PloS one. 2016 Jul 6;11(7):e0158423.
7. Yoneya T, Mamitsuka H. PURE: a PubMed article recommendation system based on content-based filtering. Genome informatics. 2007;18:267-76.
8. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC bioinformatics. 2007 Dec 1;8(1):423.
9. Collins A, Beel J. Document Embeddings vs. Keyphrases vs. Terms: An Online Evaluation in Digital Library Recommender Systems. arXiv preprint arXiv:1905.11244. 2019 May 27.
10. Hassan HA, Sansonetti G, Gasparetti F, Micarelli A, Beel J. BERT, ELMo, USE and InferSent Sentence Encoders: The Panacea for Research-Paper Recommendation? InRecSys (Late-Breaking Results) 2019 (pp. 6-10).
11. About GEO datasets [Internet]. GEO. 2020 [cited 18 August 2020]. Available from: https://www.ncbi.nlm.nih.gov/geo/info/datasets.html.
12. NIH strategic plan for data science [Internet]. National Institutes of Health. 2020 [cited 18 August 2020]. Available from: https://datascience.nih.gov/strategicplan.
13. Hienert D, Sawitzki F, Mayr P. Digital library research in action–supporting information retrieval in sowiport. D-Lib Magazine. 2015 Mar 4;21(3):4.
14. Kopp O, Breitenbücher U, Müller T. CloudRef-Towards Collaborative Reference Management in the Cloud. InZEUS 2018 (pp. 63-68).
15. Han J, Kamber M, Pei J. Getting to know your data. Data mining: concepts and techniques. 2011;3(744):39-81.
16. Rajaraman A, Ullman JD. Data mining. In: mining of massive datasets. Cambridge: Cambridge University Press; 2011. p. 1–17.

17. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.
18. Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. Nist special publication Sp 109 (1995): 109
19. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. InIn Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks 2010.
20. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.
21. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. InAdvances in neural information processing systems 2013 (pp. 3111-3119).
22. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. arXiv preprint arXiv:1802.05365. 2018 Feb 15.
23. Chelba C, Mikolov T, Schuster M, et al. One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005. 2013 Dec 11.
24. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous systems.
25. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364. 2017 May 5.
26. Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326. 2015 Aug 21.
27. Facebookresearch / InferSent [Internet]. GitHub repository. 2020 [cited 18 August 2020]. Available from: https://github.com/facebookresearch/InferSent.
28. Cer D, Yang Y, Kong SY, et al. Universal sentence encoder. arXiv preprint arXiv:1803.11175. 2018 Mar 29.
29. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
30. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. 2019 Aug 27.
31. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676. 2019 Mar 26.
32. Semantic scholar | AI-Powered Research Tool [Internet]. Semanticscholar.org. 2020 [cited 18 August 2020]. Available from: https://www.semanticscholar.org/
33. Lee J, Yoon W, Kim S, et al.(2019). Biobert: a pretrained biomedical language representation model for biomedical text mining. arXiv preprint arXiv:1901.08746.
34. Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.
35. Mackenzie J, Benham R, Petri M, Trippas JR, Culpepper JS, Moffat A. CC-News-En: A Large English News Corpus.
36. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. 2019 Oct 2.
37. Craswell N. Mean reciprocal rank. Encyclopedia of database systems. 2009;1703.
38. Ogilvie P, Callan J. The effectiveness of query expansion for distributed information retrieval. In Proceedings of the tenth international conference on Information and knowledge management, pp. 183-190. 2001

Expert-level prenatal detection of complex congenital heart disease from screening ultrasound using deep learning

Rima Arnaout MD[1], Lara Curran MBBS BSc[1], Yili Zhao PhD RDCS[1], Jami Levine MD[2], Erin Chinn MS[1]and Anita Moon-Grady MD[1]
[1]University of California, San Francisco, San Francisco, California, USA; [2]Boston Children's Hospital, Boston, Massachusetts, USA

Abstract

Fetal ultrasound is recommended for every pregnancy worldwide to identify congenital heart disease (CHD). Although highly accurate in principle and despite clear guidelines, diagnosis is poor due to challenges in image acquisition and interpretation. Deep learning (DL) is a cutting-edge technique for identifying patterns in imaging. We use an ensemble of DL algorithms to distinguish normal hearts from any CHD in prenatal screening ultrasound, testing in over 4,000 prenatal ultrasounds (over 4.4 million images) and achieve an AUC of 0.99 on distinguishing normal hearts from CHD.

Introduction

Congenital heart disease (CHD), the most common birth defect, can be asymptomatic in fetal life but cause significant morbidity and mortality after birth[1]. The earlier it is diagnosed, the better the outcomes and therapeutic options[2]. Fetal screening ultrasound is recommended for every pregnant woman worldwide and provides five views of the heart that, taken and used correctly, can diagnose over 90 percent of congenital heart disease[1]. In practice, however, the fetal diagnosis rate for congenital heart disease is only 30-50 percent, even where fetal ultrasound is universal[1-3]. We hypothesized that the main reason for this startling diagnosis gap is inadequate/uneven expertise in interpreting fetal cardiac images, due to the diagnostic challenge presented by a small and fast-beating fetal heart and due to relatively low exposure to CHD among caregivers (owing to its low prevalence). Supporting this hypothesis is that small, single-center clinical quality control programs can increase detection of CHD up to 100 percent[3]. We therefore decided to test whether deep-learning image analysis could improve upon diagnosis rates commonly encountered in community practice, even when trained only on a relatively small number of clinically relevant imaging studies. Previously, we demonstrated proof-of-concept in using deep learning (DL) to distinguish normal fetal hearts from two CHD lesions in a small test set of fetal echocardiograms (specialty ultrasound of the fetal heart)[4]. We now extend the generalizability and robustness of this work to distinguishing normal fetal hearts from any of 16 CHD lesions in a test set of fetal *screening* ultrasound (the non-specialty imaging recommended for every fetus worldwide), in a large test set with a CHD prevalence similar to that of the general population[5].

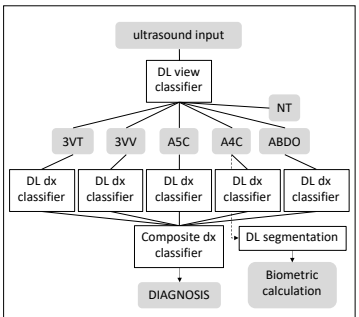


Figure 1. Diagram of ensemble model.

Methods

Using retrospectively collected echocardiograms and screening ultrasounds from fetuses 18-24 weeks of gestational age from UCSF's collection 1994-2019 (over 100,000 images from over 1300 imaging studies from multiple imaging machines, over-read by clinical experts), we trained deep learning models to identify the five screening views of the fetal heart from a screening fetal ultrasound. We then trained models to distinguish between normal hearts and any of 16 complex CHD lesions, which require referral to a fetal cardiologist and often require surgery during the neonatal period. With a supervised learning approach, we trained a ResNet-inspired CNN model to distinguish the five recommended views of the heart from among any image a fetal ultrasound (Figure 1, DL view classifier). We then developed binary classifiers using the same network architecture to distinguish, for each view, a classification of normal vs any CHD (Figure 1, DL dx classifiers). For these classification tasks, we used four different test sets, each test set was separate from the training data. FETAL-125 test set was 125 UCSF fetal echocardiograms with 30% CHD; OB-125 was 125 UCSF fetal screening ultrasounds from the corresponding fetuses as in FETAL-125; BCH-400 was 423 fetal echocardiograms from Boston Children's Hospital, with 92% CHD; and OB-4000 was 4,108 UCSF fetal screening ultrasounds with 0.9% CHD (similar to community prevalence) and represented over 4.4 million images.

Convergence plots of training and test data performance by training epoch showed an absence of overfitting. We also visualized saliency maps and gradient weighted class activation maps (GradCAMs) for test data to understand the image features that the view classification and diagnostic classification models used to make their decisions. For each heart, we arrived at a composite diagnostic decision of normal vs. CHD by applying a rules-based classifier to the five per-view diagnostic classifications above (Figure 1, composite dx classifier). (Briefly, the rules-based classifier essentially summed prediction probabilities of normal and of CHD for images within views and then across views). Separately, we trained a UNet to segment cardiothoracic structures from axial 4-chamber images and used the segmentations to calculate standard biometrics such as cardiothoracic ratio and cardiac axis from normal hearts, tetralogy of Fallot (TOF), and hypoplastic left heart syndrome (HLHS) CHD lesions.

Results

For view classification, diagnostic-quality views were found from screening fetal ultrasound with 96% sensitivity and 92% specificity (as measured in the OB-125 test set of 329,405 images). For diagnostic classification, per-view AUCs in distinguishing normal from CHD ranged from 0.72-0.88, where the abdomen view had the lowest AUC, consistent with the clinical observation that abdomen is the least useful view for CHD diagnosis. Saliency maps and GradCAMs showed that the model's decisions are based on clinically relevant image features. Per-view predictions from the remaining four views were used in the rules-based classifier to achieve AUCs in distinguishing normal vs CHD hearts of 0.98, 0.93, 0.99, and 0.89 in distinguishing normal from abnormal hearts on FETAL-125, OB-125, OB-4000, and BCH-400 test sets, respectively. Most notable was the performance on OB-4000, the largest test set that is most similar to community-based ultrasound, consisting entirely of fetal screening ultrasound and containing 0.9% CHD: an AUC of 0.99 on OB-4000 allowed a sensitivity of 95% (95%CI, 83-99%), specificity of 96% (95%CI, 95-97%), positive predictive value of 20% (95%CI, 17-23%) and a negative predictive value of 100%. For segmentation, in normal hearts predicted cardiothoracic ratio (0.52±0.03) was statistically indistinguishable from clinically measured values (Mann-Whitney U (MWU) p-value 0.2), and were the same across normal, TOF, and HLHS groups (Kruskal-Wallis p-value 0.5). As expected from clinical experience, hearts with tetralogy of Fallot had increased cardiac axis compared to normal hearts (63±16 degrees, p-value 0.007).

Discussion

To our knowledge, this is the first use of deep learning to approximately double community-reported sensitivity and specificity on a global diagnostic challenge in a test set of real fetal screening ultrasound with a CHD prevalence similar to standard population (OB-4000). Building on prior work[4], we can now distinguish normal heart from 16 complex CHD lesions with high performance, making our model a useful screening tool. We have extensively tested our model in several datasets, including outside-hospital datasets, datasets with a high prevalence of CHD, and datasets with a CHD prevalence similar to the general community (OB-4000). Our approach to both model design and testing ensured interpretability at several levels, which can help with clinical adoption. The model's performance and speed allow its integration into clinical practice as software onboard ultrasound machines to improve real-time acquisition and to facilitate telehealth approaches to prenatal care. We look forward to further multi-center testing of our model.

References Cited

1. Donofrio, M. T. et al. Diagnosis and treatment of fetal cardiac disease: a scientific statement from the American Heart Association. Circulation 129, 2183-2242, doi:10.1161/01.cir.0000437597.44550.5d (2014).

2. Oster, M. E. et al. A population-based study of the association of prenatal diagnosis with survival rate for infants with congenital heart defects. Am J Cardiol 113, 1036-1040, doi:10.1016/j.amjcard.2013.11.066 (2014).

3. Corcoran, S. et al. Prenatal detection of major congenital heart disease - optimising resources to improve outcomes. Eur J Obstet Gynecol Reprod Biol 203, 260-263, doi:10.1016/j.ejogrb.2016.06.008 (2016).

4. Arnaout, R. et al. Deep-learning models improve on community-level diagnosis for common congenital heart disease lesions. arXiv preprint server arxiv.org/abs/1809.06993 (2018).

5. Arnaout, R. et al. Expert-level prenatal detection of complex congenital heart disease from screening ultrasound using deep learning. medRxiv preprint server doi: /10.1101/2020.06.22.20137786 (2020).

A Generalizable Framework for Exhaustive Cost-Effectiveness Analysis of Drugs for a Given Clinical Indication Leveraging Real-World Evidence

D. Arneson, PhD[1], R. Vashisht, PhD[1], A. Butte, MD PhD[1]
[1]University of California, San Francisco, CA

Introduction

Hypertension is a chronic disease that affects one in three adults in the United States with an estimated annual direct cost of $47.3 billion and annual patient cost of $733[1]. Despite the widespread incidence and financial burden of hypertension, there does not exist a standardized set of prescription guidelines. What exists is a multitude of guidelines from different countries, agencies, and associations with varying recommendations spanning multiple drug classes and active ingredients leading to prescription tendencies which may not be optimized for cost-effectiveness.

To bypass the technical limitations imposed by the reanalysis of randomized controlled trials (RCTs), we capitalized on the vast and growing wealth of available prescription drug data residing in electronic health records (EHRs). These data were the basis of our novel generalizable framework to conduct cost-effectiveness analysis of prescription drugs using real-world evidence (RWE). This framework allowed us to directly compare the cost-effectiveness of all antihypertensive drugs to propose a new set of prescription guidelines as ranked lists of effectiveness per dollar based on sets predefined factors (e.g. age, race, sex, pre-existing condition, etc.).

Methods

We built a generalizable framework to conduct cost-effectiveness analysis of prescription drugs using real-world evidence. For a given drug class, we obtained the deidentified EHR records of all patients who were treated with that drug class and met a prespecified set of inclusion criteria.

The inclusion criteria for this study were: first time user of antihypertensive monotherapy, prior diagnosis of hypertension with blood pressure (BP) $\geq$ 130/80, prescription occurred during an outpatient visit, at least 1 year of medical history in the EHR prior to the prescription, at least 1 year of follow-up data in the EHR after the prescription, and at least 60 days of consecutive use of the prescription.

To evaluate the cost-effectiveness of antihypertensives on patients commonly seen in the health care system and provide updated prescription guidelines, cohorts were specified for exhaustive combinations of demographics (age, sex, race, etc.) and comorbidities (type 2 diabetes, chronic kidney disease, heart failure, etc.). The resulting cohorts were propensity score matched for all pairwise combinations of antihypertensive drug classes. We considered a number of outcomes to evaluate the relative cost-effectiveness of each drug including: quality-adjusted life year (QALY), incremental cost-effectiveness ratio (ICER), ΔBP per dollar, and incidence of secondary outcomes per dollar (e.g. heart failure, stroke, dementia, etc.). Based on the selected outcome, drugs were ranked by benefit per dollar.

We first built our framework using EHR data from the UCSF OMOP Deidentified (DEID) Clinical Data Warehouse (CDW) which uses the OMOP Common Data Model (CDM). This standardized data format allowed us to validate our findings using EHR data from four other academic medical centers which are part of the University of California Health system (UC Davis, UCLA, UC Irvine, and UCSD). The UC Health System EHR is ~5x the size of the UCSF EHR which allowed us to evaluate an additional ~35,000 additional patients who met our inclusion criteria.

Data was queried from the UCSF OMOP DEID CDW using SQL and from the UC Health Data Warehouse (UCHDW) through the UC Data Discovery Portal with Apache Spark and Databricks. All downstream analysis was conducted using R statistical software. Code for obtaining and analyzing the data will be made available to facilitate reproducibility and dissemination of the framework.

Results

We first applied our framework to 6,847 antihypertensive treated patients from the UCSF OMOP DEID CDW who met our inclusion criteria. To identify the antihypertensive drug class which produced the greatest change in blood pressure per dollar, we considered all pairwise combinations of drug classes and matched the cohorts with propensity scores to correct for potential bias. We compared the distributions of change in systolic blood pressure normalized per dollar before and after antihypertensive treatment for individuals treated with diuretics and four other major antihypertensive drug classes (Figure 1). We found that diuretics provided the largest decrease in systolic blood

pressure per dollar per day (48–66 mean decrease in systolic blood pressure per dollar spent per day) in matched cohorts compared to other major antihypertensive drug classes (Figure 1).

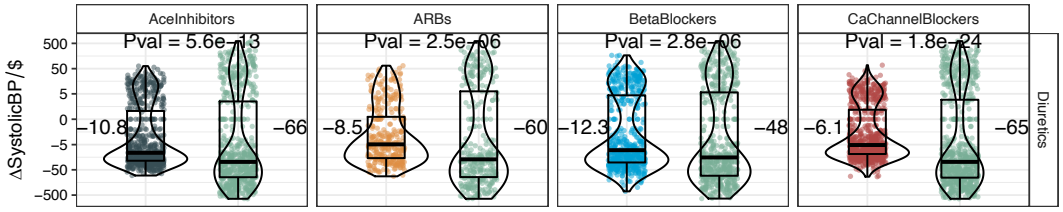


Figure 1. Results of cost-effectiveness of antihypertensives from UCSF OMOP DEID CDW. There are 4 pair-wise comparisons shown here, between one drug class (title above each plot) versus diuretics (title on the right of the row). Each point corresponds to a patient and is colored by the drug class. The outcome of change in systolic blood pressure per dollar (normalized by cost per pill per day adjusted by dose) is on the y-axis. The mean change is next to the violin plots and the p-value from a Wilcoxon rank sum test is at the top. ARBs (Angiotensin Receptor Blockers).

To evaluate the generalizability of both our framework and these findings, we extended our analysis to the UCHDW which included four additional academic medical centers (Table 1). We observed a similar effect on change in systolic blood pressure between diuretics and three of the four other antihypertensive drug classes, with few statistically significant differences (Table 1 top). However, ACE inhibitors provided a greater reduction in systolic blood pressure across all five sites with a significant difference at UCLA and close to significant at UCD, UCI, and UCSF (adjusted p-value = 0.056, 0.05, and 0.068) compared to a matched cohort of diuretics users. These changes in systolic blood pressure were then converted to change in systolic blood pressure per dollar per day using cost and frequency information (Table 1 bottom). Consistent with the findings from the UCSF OMOP DEID CDW, diuretics provided a higher average reduction in systolic blood pressure per dollar per day (range of average values: 56.7-112.2) versus the other four reported antihypertensive drug classes (range of average values: 6.13-17.8). This is largely attributable to the similar effect of diuretics on the outcome of change in systolic blood pressure versus other antihypertensive drug classes and the low average cost of diuretics versus the average cost of the other drug classes.

		UCD (n=9,718)		UCI (n=2,402)		UCLA (n=15,805)		UCSD (n=2,829)		UCSF (n=3,203)	
ΔSBP	ARBs (n=4,791)	-9.1/-9.7	0.329	-10.6/-9.6	0.797	-12.0/-12.2	0.433	-14.0/-11.2	0.336	-9.0/-11.9	0.165
	Beta Blockers (n=4,995)	-9.9/-10.3	0.599	-10.1/-9.9	0.914	-11.2/-10.6	0.396	-10.8/-11.7	0.566	-8.5/-10.4	0.276
	CCBs (7,675)	-11.0/-10.3	0.353	-11.2/-11.3	0.848	-12.2/-12.9	0.236	-13.1/-10.4	0.114	-12.4/-13	0.725
	Ace Inhibitors (n=10,745)	-10.8/-11.8	0.056	-9.7/-12.5	0.050	-12.0/-13.8	0.001	-12.0/-13.8	0.213	-11.9/-14.0	0.068
ΔSBP/$	ARBs (n=4,791)	-70.8/-9.9	1.17E-19	-74.6/-9.1	4.60E-07	-93.0/-12.2	1.46E-74	-112.2/-12.2	5.67E-06	-65.4/-12.0	4.45E-06
	Beta Blockers (n=4,995)	-79.9/-17.2	1.94E-29	-70.8/-11.2	5.73E-04	-87.8/-15.3	2.30E-45	-74.9/-17.8	6.24E-05	-56.7/-14.5	3.60E-07
	CCBs (7,675)	-90.8/-6.4	6.04E-69	-73.5/-7.0	5.02E-12	-91.8/-7.8	1.28E-121	-99.7/-6.13	5.29E-20	-92.2/-8.2	1.17E-35
	Ace Inhibitors (n=10,745)	-88.8/-12.0	9.49E-65	-70.6/-12.8	2.85E-06	-91.6/-14.1	7.64E-70	-95.4/-14.5	6.31E-12	-86.3/-14.3	3.83E-22

Table 1. Results of cost-effectiveness of antihypertensives from UCHDW. Pair-wise comparisons for drug class (each row) versus diuretics across 5 sites in the UC Health System (columns). The top half of the table reports changes in systolic blood pressure after antihypertensive treatment and the bottom half reports changes in systolic blood pressure per dollar. For each site, the left column gives the average change in systolic blood pressure (or change per dollar) with the left number corresponding to diuretics and the right number corresponding to the drug class indicated in the row. The right column for each site gives the adjusted p-value for a Wilcoxon rank sum test between matched cohorts of diuretics and the indicated drug class. ARBs (Angiotensin Receptor Blockers), CCBs (Calcium Channel Blockers).

Discussion

We developed a flexible framework that can be used to evaluate the cost-effectiveness of all drugs for a given indication. This framework was used to systematically evaluate the benefit-per-dollar of all commonly prescribed antihypertensives using RWE. Our findings were robust across multiple hospital centers in the UC Health System. The outcomes of this study can be used to inform new prescription guidelines for patient populations treated with antihypertensives. The framework we introduce here highlights a new use for RWE that can make an immediate impact on clinical prescription guidelines for any drug class it is applied to.

References

1. Park C, Wang G, Durthaler JM, Fang J. Cost-effectiveness analyses of antihypertensive medicines: a systematic review. Am J Prev Med. 2017;53(6S2):S131-S42.

Using SNOMED CT Relationships for Data Exploration and Discovery in Rare Diseases - An Interactive Data Visualization Tool

Tom Balmat[1] and Rachel L. Richesson, PhD[2]
[1]Duke University, Durham, NC; [2]University of Michigan, Ann Arbor, MI

Introduction

Electronic clinical and research data coded with standardized terminology can accelerate the discovery of detailed clinical phenotypes, which are vital to understanding the pathology and management of rare and emerging disorders, particularly multi-symptomatic disorders with varying severity and patterns of disease. The formal semantic relationships in SNOMED CT can support the exploration and analysis of data to recognize new clinical phenotypes, but tools for using these relationships in clinical analytics or research are lacking. The objective of this presentation is to demonstrate a data visualization tool that leverages the formal semantics of SNOMED CT to advance data exploration in a large research dataset on children with rare Urea Cycle Disorders (UCD).

Methods

Data source. The NIH-funded UCD natural history study includes data on over 800 participants, each with a confirmed molecular diagnosis of one of 8 different UCD subtypes followed over 12 years.[1] Data from UCD study used in our data visualization include patient identifier (anonymized), type of UCD diagnosis, visit date, structured information on hyperammonemic (HA) events (markers of disease exacerbation that are the hallmark of UCD), clinical observations from biannual physical exam and research visits. Medical history and physical exam observations were coded in SNOMED CT by research staff at each study visit, as described in (2). We imported the UCD study data into a Neo4J graph database along with the semantic relationships (i.e. the "knowledge base") of SNOMED CT, building on the work of Campbell et al.[3] We used their information model to link study participants with multiple research visits, each with multiple observations linked to SNOMED CT codes.

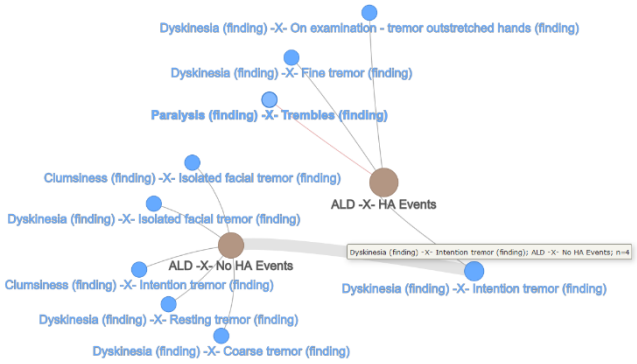
Design goals and project team. The UCD consortium[4] investigators want to explore and make full use of the rich data collected in the UCD natural history study. We collaborated with these clinical and data experts to explore the UCD and SNOMED CT data using interactive graph visualization. Because of the size of the dataset and number of SNOMED relationships, the graphs had thousands of nodes and links and were essentially unviewable. Hence, the primary design goal for the tool was to aggregate data instances into broader semantic groups based upon the relationships in SNOMED CT, which would reduce the number of nodes in a graph and allow other patterns to be detected. To meet this goal, we developed a dynamic data visualization tool to explore the prevalence of psychiatric and neurologic abnormalities across UCD diagnoses.

The tool was designed and developed by a senior data scientist (TB) who interacted with data and terminology experts for initial requirements and clinical UCD experts for iterative development. The tool was developed in R using the Shiny, RNeo4j, and visNetwork packages. Shiny provides functions for developing web pages for user interaction, while making available the entire R suite for back-end computation. RNeo4j provides an interface between R and a Neo4j database, so that queries resulting from user-specified inquiries are executed as a Cypher query against the graph database. visNetwork provides functions to generate nodes and connecting edges from queried, tabular data. In the next section we describe the architecture of the tool and some interesting challenges that we encountered.

Results

To use the visualization tool, a researcher selects SNOMED CT concepts from a hierarchical list patterned by the SNOMED CT browser (5), and additional covariates (UCD subtype, history of HA events, sex, age) to be represented as nodes in the graph. The R script queries the UCD data in the Neo4j graph database and uses it to construct node and edge descriptor tables. The descriptor tables are processed by visNetwork to produce a graph presented in a sub-window of the viewer's web page. The user reviews the graph with various roll-over labels, zooming methods, and graph reformatting operations (bipartite, tripartite, radial edge bundled). Node and edge sizes, weights, and roll-over labels indicate numbers of distinct associated participants.

A feature of visNetwork places nodes with high mass (many participants) near the "gravitational center" of other, related nodes of lesser mass. This produces sub-networks of high mass concepts surrounded by associated diagnoses and conditions. Once a graph is rendered, nodes can be repositioned using selectable methods that affect attraction (causing connected nodes to be "dragged" along). Attraction can be disabled, so that nodes can be repositioned without affecting others. Nodes can also be programmatically repositioned, and it may be useful to implement alternative algorithms, such as for collecting nodes by type (Concept nodes in one region, Participant nodes in another, Prescription nodes in a third region). Features are implemented to subset a graph by either highlighting a node's nearest neighborhood or by truncating the graph to the neighborhood of a selected node. Nodes can also be "exploded," such that instead of subnetting to the selected node, a graph is rendered using the children of the node. This drill-down method is useful in examining detailed concepts or diagnoses that produce high mass relationships observed at higher levels. Multiple node subsets are also possible, based on user specified node or edge filters.



We have used this tool with disease experts to explore the co-occurrence of multiple characteristics stratified by subgroup in the UCD natural history dataset. Our work involved multiple conversations with a team of UCD researchers and data visualization experts to evaluate and adjust evolving query parameters and results displays. In using the tool, we discovered several important graph features that must be considered during interpretation. We found that the size connected vertices must be considered when assessing significance of relative relationships. For instance, due to its heavier connecting edge, the relationship of proximal-UCD disorders to Attention Deficit Hyperactivity Disorder appeared more significant than that to Mood Disorder in our visualization. However, the heavier edge was explained by a greater number of observations in the first relationship than in the second, when in fact, the proportions of different UCD subtypes for each SNOMED concept were similar. Hence, there are unrealized opportunities to improve our interface to convey these data dynamics.

Discussion

The tool is interactive to allow iterative, drill-down queries and questions to be asked of the data, such that researchers can explore subsets of observations as new associations are revealed. This is a prototype tool but has been favorably reviewed by clinical investigators and data experts familiar with UCD and the research dataset. Preliminary experimentation has identified important patterns of association between UCD diagnoses and SNOMED CT concepts, such as for motor dysfunction and involuntary motion (tremors), that align with clinician intuition. A more formal evaluation of this tool and the value of SNOMED CT in the exploration of this dataset is forthcoming.

Our podium presentation will highlight specific benefits and challenges to using SNOMED CT relationships to display complex biomedical data in a manner that allows clinical experts to detect new and validate suspected relationships. This work demonstrates an approach using formalized and existing knowledge to mine and leverage and re-use existing datasets. We see applications for this tool in translational science in all diseases, and a particular promise for rare diseases which by definition have far fewer data resources available.

Acknowledgements This work was supported by the UCD Consortium and the Duke University Office of Research Computing. The UCD Consortium (U54HD061221) is a part of the NIH Rare Disease Clinical Research Network, supported through collaboration between the Office of Rare Diseases Research, the National Center for Advancing Translational Science (NCATS), and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). We are grateful to the following for their significant contributions to this work: Bob McCarter, Rima Izem, Marcia Bowen, W. Scott Campbell, Jay Petersen, Eric Monson, Prajwal Viendra, and Sandesh C S Nagamani, MD.

References

1. Batshaw, M. L., Tuchman, M., Summar, M., Seminara, J. A longitudinal study of urea cycle disorders. In SI: Newborn Screening, Molecular Genetics and Metabolism. 2014;113(1-2):127-130.
2. Richesson, R., et al. A web-based SNOMED CT browser: distributed and real-time use of SNOMED CT during the clinical research process. Stud Health Technol Inform 2007: 129(Pt 1): 631-635.
3. Campbell, W. S., et al. An alternative database approach for management of SNOMED CT and improved patient data queries. J Biomed Inform 2015: 57: 350-357.
4. Merritt, II, J. L., Seminara, J., Tuchman, M., Krivitzky, L., et. al. Establishing a consortium for the study of rare diseases: The Urea Cycle Disorders Consortium. Molecular Genetics and Metabolism, 2010(100):S97-S105.
5. SNOMED International. SNOMED CT Browser, 2020. URL https://browser.ihtsdotools.org

Quantifying COVID-19 In-hospital Deterioration Risk Using Acuity at Admission as Measured by the Rothman Index

Joseph Beals, PhD[1], Jaime Barnes DO[2], Daniel Durand, MD[2], Joan Rimar, DNSc, RN[3], Thomas Donohue, MD[3], Mahfuz Hoq, MD[4], Kathy Belk, BA[1], Alpesh Amin, MD[5], Michael Rothman, PhD[1]

[1]PeraHealth Inc., Charlotte, NC; [2]Sinai Hospital, Baltimore, MD; [3]Yale New Haven Health, Yale New Haven Hospital, New Haven, CT; [4]Yale New Haven Health, Bridgeport Hospital, Bridgeport, CT; [5]University of California Irvine Medical Center, Orange, CA

Introduction

Reports have noted that hospitalized COVID-19 patients have a special propensity for rapid deterioration. "*For some, the second stage starts between days five and seven when sudden rapid clinical deterioration may occur. We have not yet found any predictive symptoms of subsequent deterioration.*"[1] Hospitalized COVID-19 patients exhibit a high rate of intensive care unit (ICU) utilization, mechanical ventilation, and significant risk of in-hospital mortality.[2,3] Use of data-driven, validated models to support the triaging of COVID-19 patients has the potential to support clinicians making difficult care decisions in a more effective and objective fashion. Reports indicate that age and comorbidity correlate with COVID-19 mortality risk.[4] We postulate that both are essentially indirect estimates of patient acuity. Our work bypasses the need for a proxy-based approach by using the Rothman Index (PeraHealth, Inc. Charlotte NC) to directly measure physiologic acuity and stratify the risk of COVID-19 patients at the time of admission. The RI machine learning model has been extensively published in the literature,[5] and the RI has been validated as an indicator of patient acuity and early deterioration across all hospital care settings and within and across different patient types and diagnostic groups.[6,7]

Methods

We conducted an IRB approved analysis of 1,453 COVID-19 patients discharged between 4/1/2020-4/28/2020 and 10,093 non-COVID-19 patients discharged between 2/15/2020-4/28/2020 from Yale New Haven Health System's Yale New Haven, CT, Bridgeport, CT, and Greenwich, CT, hospitals together with 216 COVID-19 patients discharged between 3/21/2020-6/5/2020 and 4,108 non-COVID-19 patients discharged between 2/1/2020-6/7/2020 from Sinai LifeBridge Hospital in Baltimore, MD. The study population was limited to patients 18 years old and older admitted to a medical service at one of the included hospitals. The initial RI score on admission was used to predict risk of ICU utilization, mechanical ventilation, and in-hospital mortality, and corresponding AUC values were computed. We compared predictive performance of patient age, Charlson Comorbidity Index (CCI), and RI on admission using logistic regression models. Precision and recall for in-hospital mortality prediction using a range of initial RI thresholds were evaluated to determine operating points. To assess whether RI-based risk stratification could provide new insight to providers, we analyzed level-of-care assignment at admission for patients meeting the RI high risk criteria and whether patients were admitted directly to the ICU, transferred to the ICU following admission (and if so how long after admission) or never admitted to an ICU to determine potential opportunities for improved location assignment or timeliness on the basis of admission risk.

Results

While a substantially higher proportion of COVID-19 patients died and/or required ICU care and/or mechanical ventilation than non-COVID-19 patients, among patients expiring in the hospital, there was a significantly higher mean Charlson Comorbidity Index (CCI) for the non-COVID-19 population than for COVID-19 patients (Table 1). Notably, for a given acuity on admission as measured by initial RI score, COVID-19 patients had a higher risk of inpatient death than non-COVID-19 patients (Fig. 1).

Table 1. Mean CCI comparison.

Hospital	COVID-19 expired	non-COVID-19 expired
Sinai LifeBridge	3.5	4.9 **
Yale New Haven	3.8	4.9 **
Bridgeport	3.6	4.4 *
Greenwich	2.6	3.8 *

*(p < 0.06); **(p < 0.01);

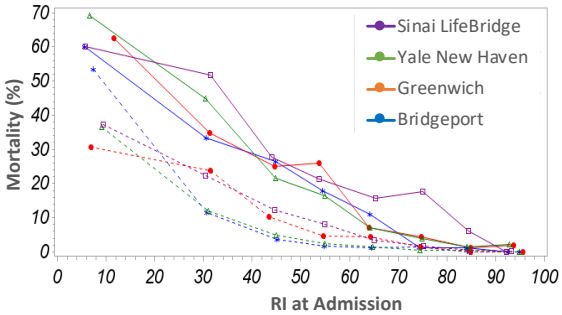


Figure 1. Mortality vs RI at admission for COVID-19 (solid lines) and non-COVID-19 (dashed lines).

As a direct measure of acuity, the RI was a substantially better predictor of COVID-19 in-hospital mortality than age or comorbidity (Table 2). At selected RI cut-points (shown in Table 3 with hospitals re-ordered and anonymized to blind hospital-level outcomes) fewer than a third of COVID-19 patients met the high risk RI criteria on admission, but had a 39-48% mortality rate, compared with around 50% of COVID-19 patients who met the low risk RI criteria and had a mortality rate of only 1-8%. Comparably large differences in ICU utilization and mechanical ventilation rates between the high and low risk groups were also found. Among COVID-19 patients who met high risk RI criteria on admission and subsequently expired, 25-46% were never admitted to an ICU. Among all patients transferred to an ICU, median time from admission to transfer ranged from 1.5-4 days, and for those patients who expired, median LOS was 4-7 days. In contrast, the majority of low risk patients did not expire, and they had approximately half the LOS of non-expiring high risk patients.

Table 2. AUC for mortality prediction.

Population	Hospital	RI	Age	CCI
COVID-19	Sinai Lifebridge	0.79	0.68	0.67
	Yale New Haven	0.87	0.78	0.70
	Bridgeport	0.87	0.75	0.72
	Greenwich	0.86	0.82	0.73
Non-COVID-19	Sinai Lifebridge	0.88	0.68	0.71
	Yale New Haven	0.90	0.73	0.74
	Bridgeport	0.93	0.81	0.75
	Greenwich	0.93	0.83	0.83

Table 3. COVID-19 patient event rates stratified by high and low risk RI criteria.

Hospital	***HIGH RISK:*** *RI < 50 (YNHHS), RI < 50 (Sinai)*				***LOW RISK:*** *RI > 70 (YNHHS), RI > 75 (Sinai)*			
	COVID-19 patients	Mortality rate	ICU utilization rate	Mechanical ventilation rate	COVID-19 patients	Mortality rate	ICU utilization rate	Mechanical ventilation rate
Hospital A	29.7%	40.9%	38.6%	25.2%	46.7%	1.0%	14.0%	3.5%
Hospital B	26.7%	45.7%	41.4%	22.0%	47.7%	2.7%	16.3%	2.7%
Hospital C	30.1%	47.7%	56.9%	41.5%	40.3%	8.0%	9.2%	8.0%
Hospital D	17.9%	39.0%	27.1%	23.7%	58.1%	2.1%	11.0%	8.9%

Discussion

COVID-19 patients exhibit a special risk profile. The risk of mortality as measured by the RI at admission is much greater than either presenting acuity or underlying comorbidities suggests. Our analysis supports the notion that clinicians are most challenged in accurately discerning risk for patients who fall somewhere in the middle of the acuity spectrum, where the apparent acuity of COVID-19 patients does not correspond to typical clinical expectations of patient risk. The RI provides a high degree of discrimination to differentiate COVID-19 populations into high and low-risk groups and to quantify the mortality risk in each group. In terms of generalizability, although alternative cut-points could have been chosen, the approach of taking a complex model and applying it in a simple manner using cut-points avoids problems, such as training set artifacts, over-fitting, or failure to account for a dynamic care and patient environment, that can arise by model training on a static data set. Use of an early, objective measure such as the RI at time of admission can support front-line clinicians in aligning level of care decisions at admission with hospital and ICU capacity constraints and help ensure that patients are placed appropriately, and resources are allocated efficiently. This applies to the high risk COVID-19 patients who may benefit from closer monitoring or more intensive therapies, and also to lower risk COVID-19 patients, some of whom may not need high levels of care, or indeed hospitalization at all.

References

1. Johnson, S. & Gottlieb, D. Breaking News: What's Working for COVID-19 Patients in the Epicenter. *Emerg. Med. News* **42**, 1 (2020).
2. Grasselli, G. *et al.* Baseline Characteristics and Outcomes of 1591 Patients Infected With SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. *JAMA* (2020) doi:10.1001/jama.2020.5394.
3. Richardson, S. *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* (2020) doi:10.1001/jama.2020.6775.
4. Jain, V. & Yuan, J.-M. Predictive symptoms and comorbidities for severe COVID-19 and intensive care unit admission: a systematic review and meta-analysis. *Int. J. Public Health* 1–14 (2020)
5. Rothman, M. J., Rothman, S. I. & Beals, J. Development and validation of a continuous measure of patient condition using the Electronic Medical Record. *J. Biomed. Inform.* **46**, 837–848 (2013).
6. Alarhayem, A. Q. *et al.* Application of electronic medical record-derived analytics in critical care: Rothman Index predicts mortality and readmissions in surgical intensive care unit patients. *J. Trauma Acute Care Surg.* **86**, 635–641 (2019).
7. Danesh, V. *et al.* Can proactive rapid response team rounding improve surveillance and reduce unplanned escalations in care? A controlled before and after study. *Int. J. Nurs. Stud.* **91**, 128–133 (2019).

Node and Edge Enrichment Analysis through Bipartite Networks: Application to Gene Mutations in Breast Cancer

Suresh K. Bhavnani PhD[1], Jeremy L. Warner MD, MS[3], Tianlong Chen PhD[1], Weibin Zhang PhD[1], Zhou Zhang PhD[5], Charles Balch MD[4], Sandra Hatch MD[2], Suzanne Klimberg MD PhD[2], Ning Liao MD PhD[6]

[1]Preventive Medicine and Population Health, [2]Cancer Center, University of Texas Medical Branch, Galveston, TX; [3]Div. of Hematology/Oncology and Dept of BMI, Vanderbilt University, Nashville, TN; [4]Div. of Surgery, MD Anderson Cancer Center, Houston, TX, USA; [5]Burning Rock Biotech, Shanghai, [6]Cancer Center, Guangdong General Hospital, Guangzhou, China

Introduction

A primary goal of precision medicine is to identify patient subgroups based on how they share key characteristics, and infer their underlying disease processes in order to design interventions that are targeted to those processes.[1] For example, breast cancer patients have been classified into five molecular subtypes (Luminal-A, Luminal-B, Triple-negative/basal like, HER2-enriched, and Normal-like).[2] However, despite the identification of these subtypes, little is understood about why subsets of patients have heterogeneous responses to current treatment paradigms.[3] For example, while most ER/PR+ HER2- patients (a clinical phenotype comprising the Luminal-A, Luminal-B, and Normal-like molecular subtypes) present at an early stage, and are cured with surgery and adjuvant therapy, a subset have poor response to current treatments, suggesting the existence of yet-to-be discovered molecular and clinical heterogeneities.

A common approach used to identify subtypes has been through the use of unipartite methods (e.g., k-means clustering, hierarchical clustering, and factor analysis), which identify *uniclusters* such as how patients cluster based on genes, or how genes co-occur across patients, with post-hoc approaches such as heatmaps to combine them. However, more recently the use of bipartite methods has shown improvements in the accuracy and interpretability of results by identifying *biclusters* such as simultaneously identifying patient subgroups and their frequently co-occurring gene mutations. Here we demonstrate how one such method called bipartite networks,[4] not only enabled the automatic identification and visualization of patient-gene biclusters, but also enabled identification of (a) which biclusters had a significantly higher proportion of patients with a specific outcome versus the rest of the data (**node enrichment**), and (b) which biclusters had genes with a significantly higher proportion of a specific mutational type within the cluster versus outside (**edge enrichment**). We demonstrate the efficacy and interpretability of this approach on a dataset of ER/PR+ HER2- Chinese breast cancer patients.

Method

***Data*.** We used a subset of data from a previous study[5] consisting of 217 Chinese breast cancer patients clinically phenotyped as ER/PR+ HER2-, and their mutational profile on 32 candidate genes. Based on their function, the mutations were categorized into four types: (1) missense variant, conservative in-frame deletion, conservative in-frame insertion, disruptive in-frame deletion, disruptive in-frame insertion, in-frame deletion, in-frame insertion, and fusion; (2) fusion and cm amp if they occur at the same time, cn amp); (3) frameshift variant, cn delete, splice acceptor variant, splice donor variant, splice region variant, splice variant, start lost, stop gained, stop lost; and (4) synonymous variant. As this subtype has low mortality rates, the outcome variables consisted of the fraction of Ki67 protein in tumor cells categorized into low (1-20), mid (21-50), and high (>50) levels, and the androgen receptor (AR) status in the tumor categorized as positive or negative. High levels of Ki67, and a negative status of AR are strong prognostic biomarkers for aggressive breast cancer in ER/PR+ HER2- patients.[6]

***Analysis*.** The analysis consisted of 4 steps: **(1) Bicluster Identification and Visualization.** (a) Represented the data as a bipartite network (Fig. 1), where nodes (circles and triangles) represented either patients or genes, the edges (lines) connecting the patient-gene pairs represented the presence of a gene mutation. Furthermore, the patient node color represented status on either Ki67 or AR (using separate networks), and the edge color represented one of four mutation types for specific patient-gene mutation pairs; (b) used bicluster modularity[4] to identify the number and boundaries of patient-symptom biclusters and the degree of biclustering (Q); (c) measured the significance of Q by comparing it to a distribution of Q generated from 1000 random permutations of the network; and (d) used the force-directed algorithm *Kamada-Kawai* to lay out the network, and *ExplodeLayout*[7] to separate the identified biclusters to improve their interpretability. **(2) Node Enrichment.** Used chi-squared with FDR correction, to measure the difference in proportion of high, mid, and low Ki67, and the proportion of positive and negative AR, in each bicluster compared to the rest of the data. **(3) Edge Enrichment.** Of those biclusters with significant node enrichment, we used chi-squared with FDR correction to measure the proportion of the mutation types for each gene-patient pair within that bicluster, versus outside that bicluster. **(4) Interpretation.** A team of oncologists interpreted the results based on: (a) the molecular mechanisms in specific biclusters that resulted in significantly high Ki67 and negative AR; and (b) potential targeted treatments.

Results

***Bicluster Identification and Visualization*.** As shown in Fig. 1, the bipartite network analysis identified 8 biclusters consisting of subgroups of breast cancer patients, and their most frequently co-occurring mutated genes, which had significant biclusteredness ($Q=0.419$, $p<.001$, $z=6.0$, two-tailed).

***Node Enrichment*. Bicluster-A** had a significantly higher proportion [$\chi^2(1, N=217)=22.81$, p<.001] of patients with high Ki67, and a significantly higher proportion [$\chi^2(1, N=217)=20.59$, p<.001] of patients with negative AR, compared to the rest of the patients. **Bicluster-B** had a significantly higher proportion of patients with low Ki67 [$\chi^2(1, N=217)=11.06$, p<.01], compared to the rest.

***Edge Enrichment*. Bicluster-A**: (1) **TP53** had a significantly higher proportion of Type-1 and Type-3 mutations [$\chi^2(1, N=217)=69.49$, p<.001]; (2) **MYC** had a significantly higher proportion of Type-2 mutations [$\chi^2(1, N=217)=48.95$, p<.001]; and (3) **FGFR1** had a significantly higher proportion of Type-2 mutations [$\chi^2(1, N=217)=27.14$, p<.001]. **Bicluster-B**: (1) **GATA3** had a significantly higher proportion of Type-1 and Type-3 mutations [$\chi^2(1, N=217)=75.63$, p<.001]; (2) **AKT1** had a significantly higher proportion of Type-1 and Type-3 [$\chi^2(1, N=217)=30.67$, p<.001].

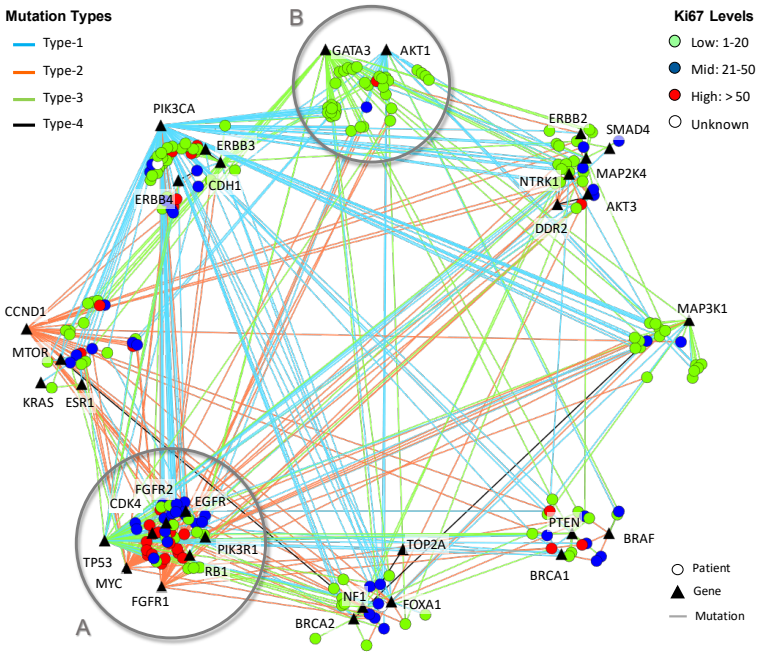


Fig. 1. 8 biclusters showing the co-occurrence of 32 genes mutated across 217 ER/PR+ HER2- Chinese breast cancer patients.

***Interpretation*.** (1) **Bicluster-A**. This bicluster had a significantly higher proportion of patients with high Ki67, with Type-1 and Type-3 mutations more likely to be associated with loss of function, and Type-2 with gain of function. This is because the bicluster contained TP53 and RB1, two well-known tumor suppressor proteins, suggesting a loss of function as a driver event. Furthermore, the MYC and EGFR oncogenes could be additional drivers explaining the relative enrichment for high Ki67 in this cluster. Finally, the co-occurrence of FGFR1/2 and CDK4 suggests potential efficacy for CDK4/6 inhibitors (e.g., palbociclib), and for FGFR inhibitors (e.g., erdafitinib), requiring prospective validation. (2) **Bicluster-B.** This bicluster had a significantly higher proportion of patients with low Ki67, and contained AKT1 and GATA3, which might define a distinct subtype of patients with differential response to treatment.[8] Although treatment data is not yet available, the overall results suggest that treatment-emergent resistance could explain the biclustering and node/edge enrichment. For example, the bicluster containing ESR1 might represent patients with acquired endocrine therapy resistance[9]; the bicluster containing ERBB2 may represent acquired HER2 mutations[10], and the biclusters containing BRCA1/2 are likely enriched for germline variants of these mutations, which are mechanistically distinct from other etiologies of breast cancer.

Conclusions and Future Research.

The results suggest that node and edge enrichment analysis used in combination with bipartite network analysis and visualization enabled a multi-channel interpretation of the data. Our future research will integrate data related to treatment responses with the goal of generating testable hypotheses for interventions targeted to specific patient subgroups.

Acknowledgements. Funded in part by UTMB Cancer Center, UTMB CTSA (UL1 TR001439), and NIH U01 (CA231840).

References

1. Collins FS, Varmus H. A new initiative on precision medicine. *The New England journal of medicine.* 2015;372(9):793-795.
2. Dai X, Li T, Bai Z, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research.* 2015;5(10):2929-2943.
3. Yang L, Ye F, Bao L, et al. Somatic alterations of TP53, ERBB2, PIK3CA and CCND1 are associated with chemosensitivity for breast cancers. *Cancer science.* 2019;110(4):1389-1400.
4. Newman MEJ. *Networks: An Introduction.* Oxford, United Kingdom: Oxford University Press; 2010.
5. Chen B, Zhang G, Wei G, et al. Heterogeneity of genomic profile in patients with HER2-positive breast cancer. *Endocrine-related cancer.* 2020;27(3):153-162.
6. Vera-Badillo FE, Chang M, Kuruzar G, et al. Association between androgen receptor (AR) expression, Ki-67, and the 21-gene recurrence score in early breast cancer. *Journal of Clinical Oncology.* 2014;32(15_suppl):547-547.
7. Bhavnani SK, Chen T, Ayyaswamy A, et al. Enabling Comprehension of Patient Subgroups and Characteristics in Large Bipartite Networks: Implications for Precision Medicine. *Proceedings of AMIA Joint Summits on Translational Science.* 2017:21-29.
8. Smyth LM, Zhou Q, Nguyen B, et al. Characteristics and outcome of AKT1 E17K-mutant breast cancer defined through AACR GENIE, a clinicogenomic registry. *Cancer discovery.* 2020:CD-19-1209.
9. Jeselsohn R, Buchwalter G, De Angelis C, Brown M, Schiff R. ESR1 mutations—a mechanism for acquired endocrine resistance in breast cancer. *Nature Reviews Clinical Oncology.* 2015;12(10):573-583.
10. Nayar U, Cohen O, Kapstad C, et al. Acquired HER2 mutations in ER+ metastatic breast cancer confer resistance to estrogen receptor–directed therapies. *Nature genetics.* 2019;51(2):207-216.

Impact of variable look-back period length on cardiovascular outcomes when using electronic health record data for epidemiologic research

Nrupen A. Bhavsar, PhD[1]; John Pura, PhD[1]; Ann Marie Navar, MD, PhD[1]; Anne Hellkamp, PhD[1]; Paul Muntner, PhD[2]; Matthew Maciejewski, PhD[1]; Sudha Raman, PhD[1]; Brad Hammill, DrPH[1]; Lesley Curtis, PhD[1]; L. Ebony Boulware, MD, MPH[1];
[1]Duke University, Durham, NC; [2]The University of Alabama at Birmingham, Birmingham, AL

Introduction:
In contrast to traditional epidemiologic cohort studies where researchers prospectively collect information during predefined study visits to use as covariates in statistical analyses, studies that use electronic health record (EHR) data are limited to data recorded following patient disclosure or an encounter. These data can be incomplete because individuals were not under observation (i.e., receiving care), they recently moved or had encounters in another health system, or because they were healthy and had no interactions with the health system. With these constraints, a central consideration for conducting epidemiological studies using EHR data is the availability of sufficient data to define study eligibility and assess baseline covariates. Researchers often only have EHR data available from a discrete time period (e.g., 3 or 5 years) to conduct analyses and there is a trade-off between having a short or long look-back period as it affects the duration of follow-up time available. This study aims to quantify the impact of 6, 12, and 24 month look-back periods on the association between diabetes and risk of an incident major cardiovascular event using EHR data alone and EHR data linked to Medicare claims.

Methods*:*
Data Sources: EHR data from the Duke University Health System (DUHS) and Lincoln Community Health Center (LCHC)from 2009-2014 were linked to Medicare claims data. The DUHS consists of two community hospitals, one large referral hospital, and a network of outpatient clinics. LCHC is a federally qualified health center serving the uninsured, underinsured, and undocumented residents of Durham County. *Eligibility criteria:* Eligible patients were those 65 years of age or older with a Durham County address at the time of their first Medicare claim in 2009. We restricted the analysis to patients who had 24 months of continuous enrollment after their first claim in 2009 in Medicare fee for service Parts A and B and an encounter in the DUHS in 2011 (i.e., the index date) (N=10076); this allowed for up to a 24-months of time prior to the index data (i.e., 24-month look-back) to identify baseline covariates. Medicare fee for service Part D was used to identify medications if available. Patients were excluded if they were diagnosed with cardiovascular disease (CVD) or cancer, with the exception of melanoma and basal cell carcinoma, during the 24 months preceding the index date, had incongruous data across multiple visits (e.g., two distinct genders, dates of birth more than 1 year apart, or two valid death dates) or were living outside of the United States at any time. A total of 5473 patients were included in this analysis.
Exposure: Diabetes was defined in the EHR using a previously developed algorithm that incorporated ICD-9 codes, hemoglobin A1c >=6.5%, and use of medication to treat diabetes. Diabetes was defined in Medicare claims data using diagnosis codes and medication to treat diabetes. *Outcomes:* The health outcome of interest was a composite indicator of major adverse cardiovascular events, including myocardial infarction, stroke, and cardiac procedures (i.e., angiography, percutaneous coronary intervention, or a coronary artery bypass graft). These were identified in the EHR and Medicare claims using diagnosis (i.e., ICD-9) and procedure (i.e., CPT) codes. *Covariates:* Covariates of interest included demographic (i.e., age, race, sex) and clinical characteristics. Clinical characteristics included hypertension which was defined in the EHR using ICD-9 codes, systolic blood pressure >=140 mmHg, diastolic blood pressure >=90 mmHg, or use of antihypertensive medication; hypertension was defined in Medicare claims using ICD-9 codes and antihypertensive medication. Only blood pressure measured through inpatient or outpatient encounters were used as prior work suggests that blood pressure measured in the emergency department may be systematically biased. Hyperlipidemia was defined by use of statin medication in EHR and Medicare claims.
Statistical Analysis: Descriptive analyses were conducted to compare baseline characteristics of the study population across 6, 12, and 24 month look back periods using EHR data only were compared to 24 months of EHR data with Medicare claims data. The total number of diagnoses identified through EHR and claims data (i.e., pooled diagnoses), the number of diagnoses in common and not in common and percent of pooled diagnoses found in each dataset were compared using EHR and claims data. Cox proportional hazards models were used to calculate hazard ratios (HR) and 95% confidence intervals (95% CI) to compare the time to the composite cardiovascular event between individuals with and without diabetes. Models were adjusted for demographic and clinical characteristics, and number of encounters[1]. Follow-up time was calculated from first EHR encounter in 2011 until an incident cardiovascular event, death, or administrative censoring on December 31, 2014, whichever occurred first. All statistical analyses were performed in R 3.4.3.

Results:

There were fewer diagnoses present in EHR data, as compared to claims data, with the number of diagnoses present increasing with longer look back periods (Table 1). The median number of total diagnoses per patient found in EHR and claims data together for 6, 12, and 24 month look back periods was 5, 10, and 29, respectively. The median proportion of pooled diagnoses present in the EHR alone was less than the median proportion of pooled diagnoses present in Medicare claims data.

Table 1: Median (interquartile range [IQR]) unique baseline diagnoses per patient present in EHR and Medicare claims data

	24 months	12 months	6 months
Sample size	5,467	5,386	5,131
Median No. diagnoses pooled (EHR + Medicare)	29 (20-43)	10 (6-17)	5 (3-9)
In EHR	15 (8-24)	6 (3-11)	3 (2-6)
In Medicare	29 (19-42)	10 (6-16)	5 (3-9)
Median No. diagnoses in both data sources	15 (8-24)	5 (3-10)	3 (1-5)
Median No. diagnoses in one but not both data sources	13 (6-24)	5 (3-10)	3 (2-6)
% pooled diagnoses present in EHR	62 (34-80)	67 (43-90)	78 (50-100)
% pooled diagnoses present in Medicare	100 (97-100)	100 (98-100)	100 (94-100)

A greater proportion of comorbidities and medications were identified with longer look back periods. Among the 5473 EHR-eligible patients, the mean age was 77 years; 67% were female and 28% were Black. The proportion of patients with diagnoses (i.e., diabetes, hypertension, and chronic kidney disease) and taking medications (i.e., statins, diabetes medications, antihypertensive medication, and ACE/ARBs) was higher with longer look back periods (Table 2).

Table 2: Proportion of diagnoses and medications identified in 6, 12, and 24 month look back period

	24 mo. (EHR + claims)	24 mo. (EHR)	12 mo. (EHR)	6 mo. (EHR)
N	5,473	5,473	5,473	5,473
Comorbidities (n, %)				
Diabetes	1,546 (28)	1,501 (23)	1,346 (21)	1,230 (19)
Hypertension	4,631 (85)	4,378 (68)	3,942 (61)	3,442 (53)
CKD	632 (12)	611 (9)	454 (7)	336 (5)
Medications (n, %)				
Statin	2,294 (42)	2,346 (36)	2,165 (33)	1,908 (29)
Diabetes medication	951 (17)	961 (15)	894 (14)	828 (13)
Anti-hypertensives	3,451 (63)	3,484 (54)	3,261 (50)	2,869 (44)
ACE/ARBs	2,648 (48)	2,688 (42)	2,511 (39)	2,214 (34)

Longer look back periods with EHR data alone resulted in associations similar to EHR and claims data combined. In a fully adjusted model – with covariates for demographic and clinical characteristics, and number of encounters - the risk for CVD was higher among individuals with diabetes as compared to individuals without diabetes for all look-back periods. Shorter look-back periods in the EHR resulted in point estimates for the hazard ratios that are overestimated as compared to longer look-back periods using EHR data alone and all available data from the EHR supplemented with Medicare claims (Table 3). However, the confidence intervals for these associations overlap.

Table 3: HR and corresponding 95% CI for association between diabetes and CVD by look-back period

	24 mo. (EHR + Claims)	24 months (EHR)	12 months (EHR)	6 months (EHR)
N	5,473	5,473	5,473	5,473
HR (95% CI)	1.43 (1.08, 1.90)	1.41 (0.98, 2.02)	1.55 (1.07, 2.25)	1.64 (1.12, 2.39)

Discussion:

The extent of data present in the EHR varied by type of data element; demographic data were present at higher rates across look back periods than comorbidity and medication data. These data were more present with increasing length of look back period. The association between diabetes and CVD was higher with shorter look-back periods but the overall inference remained unchanged regardless of look-back period length. Future CVD studies that leverage EHR data to identify baseline conditions may want to consider the use 12 to 24 month look back periods in the absence of additional administrative data that can improve data completeness. There may be minimal impact of look-back period length on inferences obtained from EHR data that are interested in CVD outcomes. This may be impacted by the extent to which the local population interacts with only a single health system.

References

1. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. Am J Epidemiol 2016;184:847-55.

Applying a computational transcriptomics-based drug repositioning pipeline to identify therapeutic candidates for endometriosis

Arohee Bhoja[1,2], Dan Bunis[1], Brian Le[1,3], Idit Kosti[1,3], Christine Li[1], Sahar Houshdaran[4], Sushmita Sen[4], Julia Vallve Juanico[4], Juan Irwin[4], Wanxin Wang[4], Linda Giudice[4], Marina Sirota[1,3]
[1]Bakar Computational Health Sciences Institute, UCSF, San Francisco, CA, [2] The Harker School, San Jose, CA, [3]Department of Pediatrics, UCSF, San Francisco, CA, [4]Department of Obstetrics, Gynecology and Reproductive Sciences, UCSF, San Francisco, CA

Introduction

Endometriosis is a reproductive disease characterized by growth of endometrial tissue outside its normal location, which can lead to symptoms of pelvic scarring, pain, and infertility. Although it is extremely common, affecting over 200 million people worldwide, current treatments mostly focus on symptom management and are not always effective [1]. Traditional drug discovery methods tend to be expensive and time consuming; it can take up to 15 years and $1 billion to bring a new drug to market [2]. Additionally, many of these novel drugs fail in later stages of testing, resulting in a substantial loss of money and time. In this study, a transcriptomics based computational drug repurposing pipeline [3] [4] was applied to endometriosis gene expression data in order to identify potential new therapeutics from existing drugs based on expression reversal. In addition, concordance of therapeutic predictions between signatures of disease from distinct disease stages or particular menstrual cycle phases was utilized to identify whether different subtypes of disease should be treated with different drugs or whether different drugs should be administered to endometriosis patients at different points of the cycle.

Methods

Microarray-based transcriptional profiling data, from eutopic endometrial tissues, of women either with endometriosis or no uterine or pelvic pathology was obtained from GEO. [5] This data was normalized with the R package justRMA [6] and batch corrected using the package ComBat [7]. Using the package limma, [8] differential gene expression analysis between disease and control samples was performed on the unstratified data and the data stratified by cycle phase and disease stage. The significant genes from each signature were identified using the cutoffs adj P-val < 0.05 and logFC > 1.1 **(Table 1).** A transcriptomics-based drug repurposing pipeline (previously applied in the Sirota et al. 2011 study) [3] was then applied to the signatures. The pipeline utilizes a rank-based pattern-matching method, leveraging gene expression data for both diseases and drugs, in order to identify disease-drug pairs with opposite transcriptional effects. On the drug side, the Connectivity Map (CMAP) [9] dataset from the Broad Institute was used, which consists of gene expression profiles from over 1000 small-molecule drugs. Reversal scores were calculated for each drug in the CMAP dataset and permutation analysis was carried out to assess significance. Drug hits with q-values < 0.0001 or reversal scores < 0 (indicating signature reversal) were examined further resulting in 190-291 drug hits per signature **(Table 1).** Network analysis on DrugBank data was used to visualize the relationships between the top significant hits [10].

Results

Table 1: Significant genes and total drug hits

Signatures	**Significant Genes** (adj P-val < 0.05, log2FC > 1.1)	**Total Drug Hits** (q-value < 0.0001, reversal score < 0)
Unstratified	324 genes	255 drug hits
Stage 1/2	571 genes	270 drug hits
Stage 3/4	284 genes	270 drug hits
PE	645 genes	264 drug hits
ESE	171 genes	190 drug hits
MSE	456 genes	291 drug hits

We found high levels of overlap in the drug hits for each signature despite the similarities not being as pronounced on the gene side. 181 of 298 total drug hits were common across all six signatures **(Figure 1A)**. There was an extensive amount of overlap between the top 20 drugs across all six signatures, with many drugs returning high reversal scores across the board **(Figure 1B).** As would be predicted, several of the identified drug hits within the top 20 (by reversal score) have been used in clinical trials to treat endometriosis; including fenoprofen (average reversal score: -0.799), an NSAID commonly used for pain relief and used successfully to treat dysmenorrhea, a symptom of endometriosis; levonorgestrel (average reversal score: -0.705), a progesterone-based emergency contraceptive, which has also been applied to endometriosis successfully in clinical trials; and menadione (average reversal score: -0.562), which has been used to alleviate pain in dysmenorrhea patients. Additionally, two of the largest categories of drugs returned by the pipeline (anti-inflammatory and sex hormone drugs) have been extensively used to treat endometriosis previously **(Figure 1C)**.

Upon looking into the protein targets for the top 20 drugs **(Figure 1D),** 15 proteins were identified that were targeted by two or more drugs. Out of those proteins, three were found to have a direct link to endometriosis. PPARG and PPARA, which are commonly targeted by NSAIDs (including fenoprofen, which is shown in the network plot), inhibit the growth of endometrial tissue when activated [11] and NR3C1 is consistently expressed in endometrial cells [12]. Finally leveraging this approach we have identified a number of novel therapeutic candidates such as flumetasone (average reversal score: -0.766), a corticosteroid

meant for topical use, and primaquine (average reversal score: -0.697), an anti-malaria drug. Chloroquine, a chemically similar drug to primaquine, has been shown to have therapeutic effects on endometriosis.

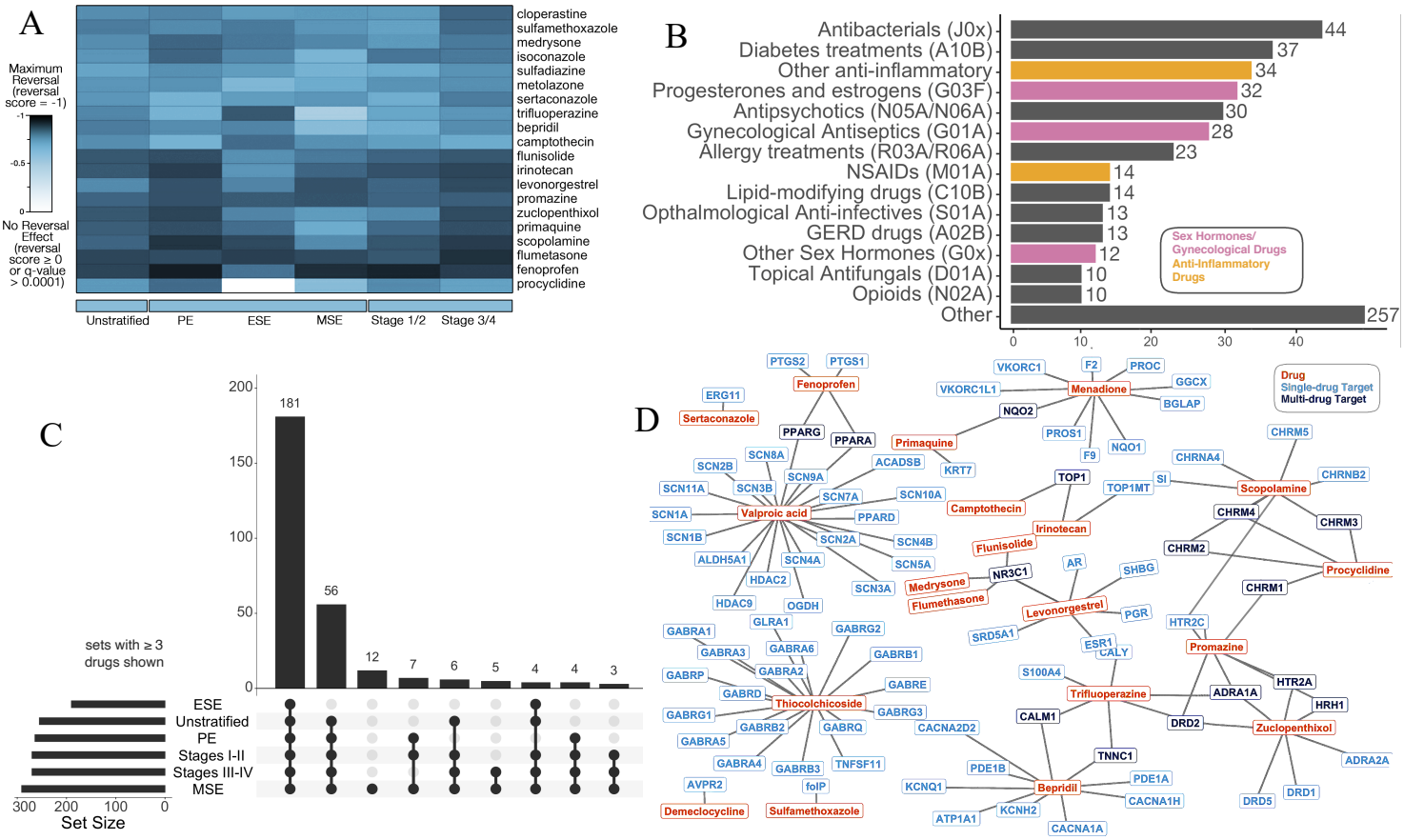


Figure 1: A) Heatmap showing reversal scores for top 20 drugs across signatures, B) Bar plot showing distributions of drug classes, C) UpSet plot showing overlap in drug hits across signatures D) Drug target network for top 20 combined drug hits (data from DrugBank, plot created using ggnetwork)

Discussion

In this work, we leveraged a transcriptomics-based drug repurposing approach to identify known and novel therapeutic candidates for endometriosis. We found that therapeutic predictions were relatively consistent across disease stage and menstrual cycle phase and included many known treatments and novel candidates. The highest ranking candidates across all signatures made sense from a clinical perspective as many are NSAIDs or hormonal drugs, and many have been taken to clinical trials. Identification of similar drugs across the signatures suggests that the therapeutic predictions are robust and similar compounds could be used to treat different stages of the disease targeting common disease mechanism across various phases of the menstrual cycle. Although further experimental work needs to be done to validate the drug hits identified from this study, the results are promising in that the pipeline returned several drugs and drug categories with known therapeutic effects on endometriosis.

References

[1] Acog.org. 2020. Endometriosis. [online] Available at: <https://www.acog.org/patient-resources/faqs/gynecologic-problems/endometriosis> [Accessed 27 August 2020].
[2] Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*. 2020;323(9):844–853. doi:10.1001/jama.2020.1166
[3] Sirota M, Dudley JT, Kim J, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data [published correction appears in Sci Transl Med. 2011 Sep 28;3(102):102er7]. *Sci Transl Med*. 2011;3(96):96ra77. doi:10.1126/scitranslmed.3001318
[4] Huang, Chen-Tsung et al. "Perturbational Gene-Expression Signatures for Combinatorial Drug Discovery." *iScience* vol. 15 (2019): 291-306. doi:10.1016/j.isci.2019.04.039
[5] *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51981.
[6] Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. 2004. affy---analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 3 (Feb. 2004), 307-315.
[7] Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Elana J. Fertig, Andrew E. Jaffe, Yuqing Zhang, John D. Storey and Leonardo Collado Torres (2020). sva: Surrogate Variable Analysis. R package version 3.36.0.
[8] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.
[9] J. Lamb *et al.*, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, Sep. 2006, doi: 10.1126/science.1132939.
[10] D. S. Wishart *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, doi: 10.1093/nar/gkx1037.
[11] Lebovic DI, Kavoussi SK, Lee J, Banu SK, Arosh JA. PPARγ activation inhibits growth and survival of human endometriotic cells by suppressing estrogen biosynthesis and PGE2 signaling. *Endocrinology*. 2013;154(12):4803-4813. doi:10.1210/en.2013-1168
[12] Simmons RM, Satterfield MC, Welsh TH Jr, Bazer FW, Spencer TE. HSD11B1, HSD11B2, PTGS2, and NR3C1 expression in the peri-implantation ovine uterus: effects of pregnancy, progesterone, and interferon tau. *Biol Reprod*. 2010;82(1):35-43. doi:10.1095/biolreprod.109.07960

Ontologizing Health Systems Data at Scale: Making Translational Discovery a Reality

Tiffany J. Callahan, MPH[1], Jordan M. Wyrwa, DO[1], Nicole A Vasilevsky, PhD[2], Peter N. Robinson, MD, PhD[3], Melissa A Haendel, PhD[4], Lawrence E. Hunter, PhD[1], Michael G. Kahn, MD, PhD[1]

[1]University of Colorado Anschutz Medical Campus, Aurora, CO, USA; [2] Oregon Health Sciences University, Portland, OR, USA; [3]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA; [4]Oregon State University, Corvallis, OR, USA

Introduction

A significant promise of electronic health records (EHRs) lies in the ability to perform large-scale investigations of mechanistic drivers of complex diseases. Despite significant progress in biomarker discovery, this promise remains largely aspirational due to its disconnectedness from biomedical knowledge[1]. Linking molecular data to clinical data will enable biologically meaningful analysis by integrating knowledge about biology and pathophysiology from multiple ontologies. Similar to clinical terminologies, ontologies are classification systems that provide detailed representations of a specific domain of knowledge consisting of a set of concepts and logically defined relationships[2]. Unlike most clinical terminologies, ontologies are computable and interoperable, which means they can be easily integrated with other data including data from basic science and clinical research.

The usefulness of normalizing (i.e. mapping or annotating) clinical data to ontologies, like those in the Open Biomedical Ontology (OBO) Foundry, has been recognized as a fundamental need for the future of deep phenotyping[1]. Existing work has largely focused on using ontologies to improve phenotyping in specific diseases[3] and for the enhancement of specific biological and clinical domains[4,5]. Prior work has largely been limited to one-to-one mappings and rarely includes robust evaluation or external validation. Unfortunately, learning algorithms are not yet able to capture the complex semantics underlying these concepts and their relationships. Until a comprehensive resource that includes mappings between multiple clinical domains and ontologies is created and validated, automatic generation of inference between patient-level clinical observations and biological knowledge will not be possible.

To address these limitations, we have developed OMOP2OBO, the first health system-wide integration and alignment between the Observational Medical Outcomes Partnership (OMOP) standardized clinical terminologies and eight OBO ontologies. To verify that the mappings are both clinically and biologically meaningful, we have performed extensive validation with assistance from multiple domain experts. Here, we present preliminary results examining the coverage of the mappings in two institutions' EHR data.

Methods

Standard clinical terminology concepts were extracted from a PEDSnet OMOP v5 de-identified instance of the Children's Hospital Colorado EHR. Clinical concepts (and their ancestor concepts) included all OMOP standard terminology identifiers from the Condition Occurrence, Drug Exposure, and Measurements tables. Additional metadata for each concept identifier included source codes, labels, and synonyms at both the concept and concept ancestor levels. Ontologies were selected under the advice of domain experts and included diseases, phenotypes, anatomical entities, cell types, organisms, chemicals, hormones, metabolites, vaccines, and proteins. Clinical data use was approved by the Colorado Multiple Institutional Review Board (#15-0445). Additional details, data, and code are available on GitHub: https://github.com/callahantiff/OMOP2OBO.

Mapping OMOP Concepts to OBO Concepts. Clinical concepts were mapped at the concept and ancestor level, drug exposures concepts were mapped at the ingredient level, and measurement concepts were mapped at the result level according to their LOINC scale type. One-to-one and one-many mappings were created using a combination of automatic and manual strategies, for each clinical concept to applicable concepts in each ontology. The automatic approach employed database cross-reference mapping, exact string mapping (using concept labels and synonyms), and word embedding-based cosine similarity scoring (using all clinical and ontology concept labels, synonyms, and definitions). All concepts unable to be mapped automatically were manually mapped. For all mappings, evidence was generated and includes the mapping source, metadata/provenance (e.g. cross-referenced identifiers, exact match strings), and validation source (e.g. expert review). Mappings were converted to Resource Description Framework

(RDF) and logically validated by running a deductive logic reasoner. Additionally, a random 20% sample of the most challenging manual condition, drug exposure, and measurement mappings were verified by a panel of domain experts spanning molecular biology, clinical pharmacology, pediatric and adult medicine, and biomedical ontology curation. Several iterations of review were performed until reaching a consensus.

Results

The full set of mapped clinical concepts included 29129 condition concepts, 1697 unique drug exposure concepts, and 4083 measurement concepts. For conditions, 20850 concepts were mapped to 4661 phenotypes and 3614 diseases. For drug ingredients, 1574 concepts were mapped to 1422 chemicals, 91 proteins, 39 organisms, and 54 vaccines. Expanding measurement concepts by result type yielded 11072 results which mapped to over 920 phenotypes, 25 anatomical entities, 27 cell types, 338 chemicals/hormones/metabolites, 194 organisms, and 113 proteins. The ratio of automatic to manual mappings differed by clinical domain and ontology with conditions and drug ingredients having more automatic mappings than measurements. These findings are likely a result of the ontologies for these domains providing significantly more metadata ($\chi^2(14) = 2{,}664{,}853.82, p < 0.0001$). Agreement between the domain experts and the mapping annotators was moderate to excellent with 90.9% on measurements, 75% on drug ingredients, and 82.5% on conditions. Coverage analysis of the OMOP2OBO concepts on clinical data obtained from two independent health systems revealed 80-92% coverage of condition occurrence concepts, 91-96% coverage of drug exposure concepts, and 50-55% coverage of measurement concepts. Finally, the RDF version of the mappings was found to be logically consistent by a deductive logic reasoner.

Discussion and Conclusion

OMOP2OBO is the first health system-wide resource to provision interoperability between clinical EHR concepts and OBO ontologies. OMOP2OBO presents unprecedented opportunities to improve clinical decision making and computational phenotyping by providing additional insight into the molecular mechanisms underlying each patient's unique set of observations at hospital scale. Currently, OMOP2OBO contains 23824 standardized OMOP clinical terminology concepts and 42249 concepts in eight OBO biomedical ontologies. Although evaluation is still ongoing, preliminary results suggest excellent coverage of OMOP2OBO condition and drug concepts and excellent coverage of measurement concepts when examined in two health systems. It is important to note that the frequently updated ontologies which also contained detailed metadata on each concept (e.g. labels, definitions, synonyms, and database cross-references) tended to result in a larger number of accurate automatic mappings. These types of ontologies were also easier for both the researchers and domain experts to navigate and utilize when performing manual annotation. Additionally, it appears that concepts which are frequently used in clinical practice may also be more likely to be represented by an existing ontology. We will be exploring both of these observations further in follow-up experiments. Additional work currently underway includes expanding mapping provenance, performing an expanded coverage study on 24 national and international hospital databases and health systems, conducting chronic and rare disease cohort studies in pediatric and adult populations, and developing a novel machine learning pipeline with the goal of improving the accuracy of automatic annotation.

References

1. Weng C, Shah NH, Hripcsak G. Deep phenotyping: Embracing complexity and temporality-towards scalability, portability, and interoperability. J Biomed Inform. 2020;105:103433.
2. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med. 2018;379:1452-62.
3. Thompson R, Papakonstantinou Ntalis A, Beltran S, Töpf A, de Paula Estephan E, et al. Increasing phenotypic annotation improves the diagnostic rate of exome sequencing in a rare neuromuscular disorder. Hum Mutat. 2019;40:1797-812.
4. Zhang XA, Yates A, Vasilevsky N, Gourdine JP, Callahan TJ, et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. NPJ Digit Med. 2019;2.
5. Raje S, Bodenreider O. Interoperability of disease concepts in clinical and research ontologies: Contrasting coverage and structure in the Disease Ontology and SNOMED CT. Stud Health Technol Inform. 2017;245:925-9.

The Impact of Sickle Cell Status on Adverse Delivery Outcomes Using Electronic Health Record Data

Silvia P. Canelón, PhD[1], Samantha Butts, MD, MSCE[2], Mary Regina Boland, MA, MPhil, PhD, FAMIA[1,3-4]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania,
[2]Division of Reproductive Endocrinology and Infertility, Penn State College of Medicine and Penn State Health,
[3]Institute for Biomedical Informatics, University of Pennsylvania,
[4]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia

Abstract. *This study investigates the effect of sickle cell trait and sickle cell disease, on adverse pregnancy outcomes at Penn Medicine. The risk of a Cesarean section (C-section), preterm birth, stillbirth, pain crisis, blood transfusion, and hemorrhage during delivery were all found to be significantly correlated with race/ethnicity, sickle cell disease, the number of pain crises before delivery, and the number of blood transfusions before delivery. Multiple birth was also found to significantly increase the risk of these same outcomes.*

Introduction. Sickle cell (SC) disease is a severe and complex inherited genetic disorder and the most common hemoglobinopathy in the United States, affecting roughly 100,000 individuals.[1] It primarily affects those of African ancestry and is associated with high lifetime morbidity and premature mortality.[2] In addition, those pregnant with SC have been shown to be at an increased risk of adverse outcomes.[3] As a result, it can be difficult to determine if the adverse pregnancy outcomes are due to SC or health disparities common among the Black or African American community.[4,5] Our study investigates this relationship by assessing the impact of sickle cell status and race/ethnicity on the risk of pregnancy-related complications by leveraging electronic health records (EHRs). EHRs contain rich information on patient medical history and treatment and can be used to study effects of prenatal exposures on delivery-related outcomes. This study utilizes a previously developed algorithm that extracts delivery episode details and delivery dates from the EHR to identify adverse outcomes for each of a patient's deliveries.[6] These delivery episode details allowed us to examine the contributions of sickle cell trait and disease as well as racial/ethnic health disparities[4] on a patient's likelihood of experiencing a pregnancy-related complication – specifically Cesarean delivery (C-section), preterm birth, stillbirth, pain crisis, blood transfusion, and hemorrhage. The Institutional Review Board of the University of Pennsylvania approved this study.

Methods. We obtained EHR data for 1,060,100 female patients with visits to inpatient or outpatient clinics within the Penn Medicine health system 2010-2017. We used encounter records to extract patient demographic and blood type information. We used *International Classification of Diseases version 9/10* (ICD-9/10) codes to identify C-section, preterm birth, stillbirth, hemorrhage, blood transfusion, and pain crisis diagnoses and procedures as *outcomes* occurring during delivery. The same approach identified multiple birth, sickle cell trait, and sickle cell disease diagnoses, as *predictors*. We used a custom algorithm[6] to extract delivery episode details, and an outcome code assigned within a delivery episode was used to link the outcome to a specific delivery. This allowed the study of pregnancy outcomes at the pregnancy-level rather than just the patient-level, while adjusting for the number of pregnancies per patient. Additional predictors included patient age, marital status (single, yes/no), race/ethnicity as binary variables, repeat C-section (yes/no), the number of pain crises before delivery, the number of blood transfusions before delivery, blood type, and blood factor Rh (negative, yes/no). The interaction between Black race/ethnicity and sickle cell mutation (sickle cell trait or disease) was also included. Logistic regression models were constructed to evaluate the impact of predictors on delivery outcomes, with $\alpha = 0.05$ defined as significant.

Table 1. Patient characteristics

	Patients (%)	Deliveries (%)
Total Population	50560 (100)	63334 (100)
Sickle cell status		
No sickle cell	48492 (95.9)	60637 (95.7)
Sickle cell mutation	2068 (4.1)	2697 (4.3)
Sickle cell trait	1904 (3.8)	2482 (3.9)
Sickle cell disease	164 (0.3)	215 (0.3)
Patient age (Mean ± SD)	29.5 ± 6.1	N/A
Marital Status Single	28186 (55.7)	34823 (55.0)
Race/Ethnicity[a]		
Black/African American	23777 (47.0)	29965 (47.3)
White	17034 (33.7)	21443 (33.9)
Hispanic	4031 (8.0)	4985 (7.9)
Asian	3305 (6.5)	4073 (6.4)
Other	1644 (3.3)	2022 (3.2)
Native Hawaiian/Pacific Islander	75 (0.2)	94 (0.2)
American Indian/Alaskan Native	61 (0.1)	81 (0.1)

[a]Descriptions are "Non-Hispanic" unless otherwise indicated.

Results. The algorithm identified 50,560 patients with delivery diagnoses or procedures at Penn Medicine and a total of 63,334 deliveries between 2010 and 2017. The predominant race/ethnicity descriptions of the patients with

deliveries were non-Hispanic Black or African American ("Black", 47.3% of deliveries), non-Hispanic White ("White", 33.9%), Hispanic (7.9%), non-Hispanic Asian ("Asian", 6.4%). Due to privacy concerns, the minority of patients that identified as either Mixed were considered to identify as "Other" for the purposes of this study (**Table 1**). Of patients with a sickle cell mutation, 92.9% were Black. In regards to *C-sections*, Native Hawaiian/Pacific Islander patients were at greatest risk (aOR 1.65, $p < 0.05$) followed by Black patients (aOR 1.26, $p < 0.001$), and multiple birth, sickle cell disease, and age were also significant risk factors. For *preterm birth*, multiple birth was the strongest predictor (aOR 11.1, $p < 0.001$), followed by SC disease (aOR 2.95, $p < 0.01$), and identifying as Black (aOR 1.66, $p < 0.001$). Black patients with a SC mutation were at decreased risk (aOR 0.54, $p = 0.040$). The strongest predictor for *stillbirth* was SC disease (aOR 7.25, $p < 0.01$) followed by multiple birth (aOR 4.68, $p < 0.001$), while having a prior C-section significantly decreased the risk (aOR 0.43, $p < 0.01$). The risk of *hemorrhage* was greater for multiple birth and among Asian patients (aOR 1.70, $p < 0.001$ and aOR 1.48, $p < 0.01$). Patients with SC disease were at greatest risk of a *blood transfusion* during delivery (aOR 8.48, $p < 0.001$) and correlated with the number of prior blood transfusions (aOR 1.66, $p < 0.01$). The risk was lower in the case of prior pain crises and prior deliveries (aOR 0.89, $p = 0.021$ and aOR 0.52, $p < 0.001$). Prior pain crises were associated with an increased risk of *pain crises* during delivery (aOR 1.3, $p = 0.019$). Blood types B and AB were associated with an increased risk of preterm birth (aOR 1.11, $p = 0.028$ and aOR 1.26, $p < 0.01$) and blood factor Rh negative increased risk of a blood transfusion (aOR 1.28, $p = 0.022$). The average maternal age at time of delivery was 29.5 ± 6.1 years. Maternal age at the time of delivery correlated with blood transfusions during delivery and C-sections (aOR 1.01, $p = 0.044$ and aOR 1.05, $p < 0.001$) (**Figure 1**).

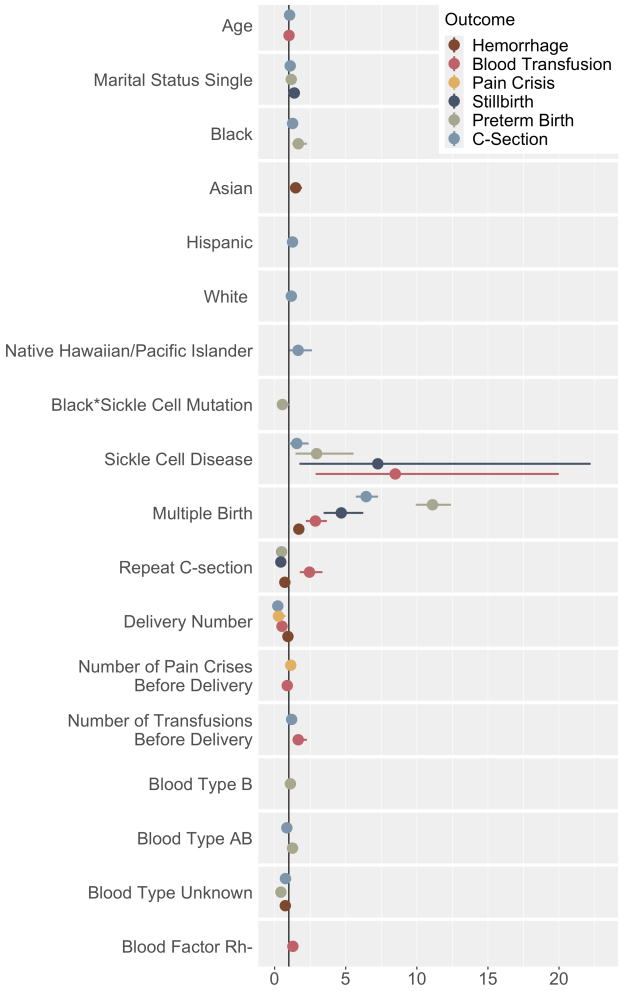


Figure 1. Adjusted odds ratio estimates with 95% confidence intervals, for each adverse delivery outcome. Only significant predictors are displayed due to space constraints.

Discussion. Incorporating sickle cell status into the model revealed that patients with SCD were at an increased risk of C-section, preterm birth, stillbirth, and blood transfusion during delivery. Patients with sickle cell disease were twice as likely to have a preterm birth (aOR 2.95 vs. 1.49), 2.7x more likely to have a stillbirth (aOR 7.25 vs. 2.65), and 9.7x more likely to require a blood transfusion during delivery (aOR 8.48 vs. 0.88), compared to those with sickle cell trait. These findings confirm that SCD pregnancies are at high risk of adverse outcomes[3] and that patients would benefit from greater systemic support for comprehensive coordinated care. We confirmed the existence of racial/ethnic disparities in adverse pregnancy outcomes among patients at Penn Medicine, most notably for C-section, preterm birth, and hemorrhage outcomes. Native Hawaiian/Pacific Islander patients were 1.4x more likely than White patients to have a C-section (aOR 1.65 vs. 1.17). Likewise, Black patients were 1.8x more likely to have a preterm birth (aOR 1.66 vs. 0.89) and Asian patients were 1.8x more likely to experience a hemorrhage during delivery (aOR 1.48 vs. 0.82). Reasons for these differences are multifactorial and could include adverse socioeconomic circumstances and lived experiences with racism contributing to maternal stress, among others.[4] In the future, we hope to increase our work detailing the causes of these disparities and exploring potential mitigation approaches to address them.

References

1. Centers for Disease Control and Prevention. Data & Statistics on Sickle Cell Disease [Internet]. 2017. Available from: https://www.cdc.gov/ncbddd/sicklecell/data.html
2. Kuo K, Caughey AB. Contemporary outcomes of sickle cell disease in pregnancy. Am J Obstet Gynecol. 2016;215:505.e1-505.e5.
3. Barfield WD, Barradas DT, Manning SE, Kotelchuck M, Shapiro-Mendoza CK. Sickle Cell Disease and Pregnancy Outcomes. Am J Prev Med. 2010;38:S542–9.
4. Bryant AS, Worjoloh A, Caughey AB, Washington AE. Racial/ethnic disparities in obstetric outcomes and care: prevalence and determinants. Am J Obstet Gynecol. 2010;202:335–43.
5. Martin JA, Hamilton BE, Osterman MJK, Driscoll AK, Drake P. Births: Final data for 2017. Vol. 67, Natl Vital Stat Rep. Hyattsville, MD; 2018.
6. Canelón SP, Burris HH, Levine LD, Boland MR. Development and Evaluation of MADDIE: Method to Acquire Delivery Date Information from Electronic Health Records. Int J Med Inform. 2020;145:104339.

Mapping the Unmapped Readspace in Whole Genome Sequencing Data:
The Blood Microbiome of 1,000 Families

Brianna Chrisman[1], Chloe He[2], Jae-Yoon Jung[2], Kelley Paskov[2], Nate Stockham[3], Peter Washington[1], Maya Varma[4], Dennis P. Wall[2,5]

[1] Department of Bioengineering, Stanford University, [2] Department of Biomedical Data Science, Stanford University, [3] Department of Neuroscience, Stanford University, [4] Department of Computer Science, Stanford University, [5] Department of Pediatrics (Systems Medicine), Stanford University

Introduction: High coverage whole genome sequencing (WGS) of a human subject generates hundreds of GB of data, allowing researchers or clinicians to reconstruct an individual's genome and to identify disease-causing variants. However, in most WGS pipelines, millions of reads are discarded during the process of reconstruction/assembly. Known as the unmapped read space, these are reads that do not correspond to known human DNA sequences. Some of these unmapped reads correspond to the human blood virome or non-reference insertions in the human genome. Previous studies have harnessed these unmapped reads to measure the prevalence of different viruses in the human virome [1], and to catalogue alternative human haplotypes [2]. However, these studies used WGS of unrelated individuals. Using related individuals, we catalogue the unmapped read space, focusing in particular on the blood microbiome. By utilizing the structure of nuclear families we gain the unique ability to detect virus chromosomal integration events, intra-family transmission of infectious diseases, and household-specific water, food, and environmental contaminants.

Methods: We used WGS from lymphocytes or whole blood samples from over 1004 nuclear families (over 4000 individuals), from the iHART dataset previously collected and curated by our group [3]. Originally for autism research, the iHART dataset consists of WGS from children with autism, their siblings, and their parents, providing a powerful family structure for analysis (Fig. 1C). Using `bwa-mem`, we realigned the reads that did not map or aligned poorly to the human reference genome (hg38 + decoy) to a database of human viruses on NCBI [4], bacteria native to the human microbiome [5], and alternative haplotypes [2]. From the read counts of viruses and bacteria, we tested association between abundance and sequencing plate or sample type (whole blood vs. LCL) to quantify contaminants, and association between abundance and family to quantify intra-household transmissibility. An outline of the pipeline is shown in Fig. 1A.

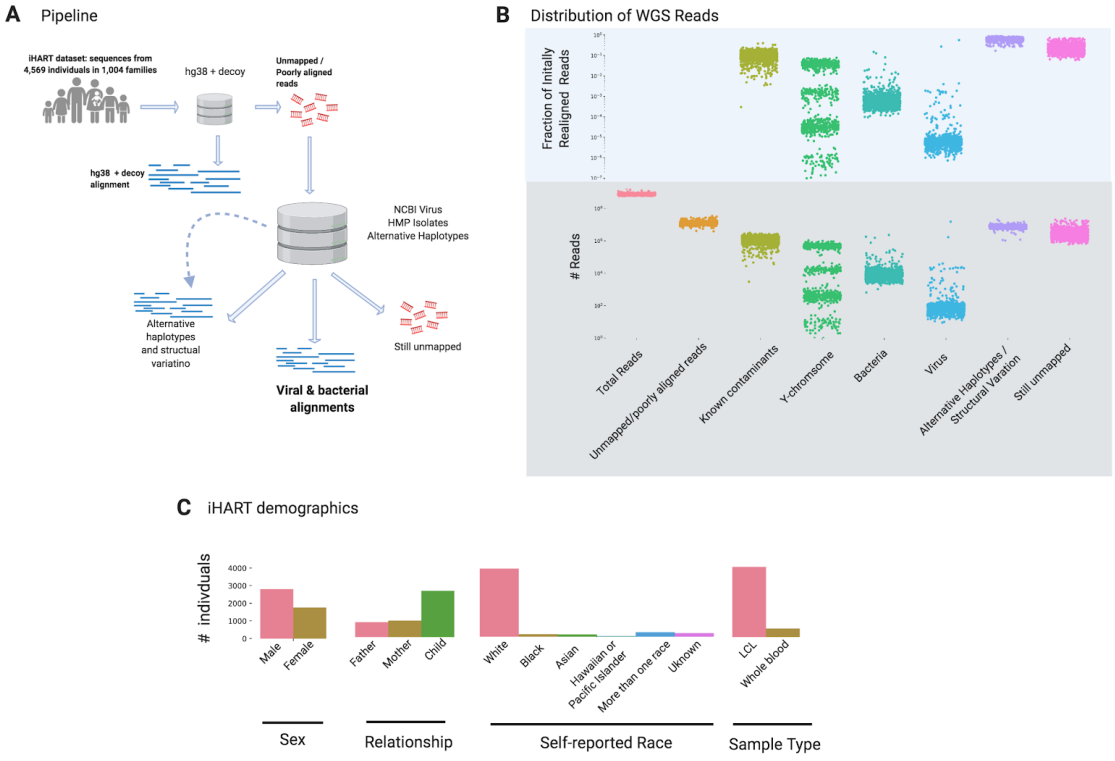


Figure 1: (A) Pipeline of project. (B) Distribution of how WGS reads poorly aligned to hg38 realigned. (C) Demographics of the iHART dataset.

Results: We found that the unmapped read space in WGS provides valuable insight into the human blood microbiome. Of the unmapped or poorly aligned reads in samples (average 11 million reads, 1.5% of all reads), on average , 64% (70M) are alternative haplotypes or structural variation poorly captured in the reference genome, 11% (1M) contaminants, 2% (20k) likely y-chromosome fragments mapping to bacterial contigs, .08% (6k) bacteria, .02% virus (100) and 22% (20M) reads ultimately remain unmapped (Fig. 1B). Although a seemingly small percentage,, the amount of reads mapped to bacteria or viruses varied dramatically between families and individuals, illustrating the wide variability in human blood microbiome composition. In addition to extensively cataloguing the viruses and bacteria detected in WGS of human blood, we present several notable results.

(1) Household drives transmission of many blood microbes. We found that the abundance of bacteria and viruses, such as *Burkholderia*, *Shewanella*, Torque Teno Virus, and parvovirus, were strongly associated with family. General blood microbiome composition is strongly dictated by family. Many bacterial species, such as *Raphidiopsis Brooki, Leuconostoc Gasicomatitum*, and *Herpspirillum Seropedicaeare*, were found in only a handful of households and infect multiple family members, and many of these have been previously reported as water and food contaminants.

(2) Human herpesviruses (HHV) 6A, 6B, and 7 show unique patterns of chromosomal integration and latency. HHV-6A has a pattern consistent with inherited chromosomally integrated herpes (iciHHV), at a rate of .4%. HHV-6B has a pattern consistent with iciHHV-6B at a rate of .2%, and also seems to establish different levels of latency in 1% of LCLs via chromosomal integration. HHV-7 does not show evidence of heritability but also seems to establish latency in 2% of LCLs, though not through chromosomal integration.

(3) In addition to known contaminants such as Epstein-Barr virus and phiX, WB and LCL sequencing results are plagued with many other contaminants, especially bacteria. *Pseudomonas, Mesorhizobium, and Bradyrhizobium* commonly contaminate lymphoblastoid cell lines, and *Ralstonia*, *Streptococcus*, *Burkholderia*, and *Acinebacter* are strongly associated with sequencing plate, indicating contamination from a previous sequencing run, or from a human handler.

(4) Probable fragments of repetitive Y-chromosome not well catalogued in the reference genome mismap to two bacteria, *F. ulcerans* and *M. Bacterium*. These bacterial abundances strongly associate with sex, with males having 1,000 - 100,000 more reads mapping to thse contigs than females. Furthermore, we found strong heritability between father and son abundances of each contig, suggesting that these reads map to a repetitive region in the y-chromosome, with high variability in the number of repeats.

Discussion: The unmapped read space of WGS from families provides valuable insight into the structure of the human blood microbiome, intra-family transmission dynamics, and viral chromosomal integration. Additionally, WGS data, and possibly reference contigs, are plagued with contaminants that could compromise WGS studies. As WGS is becoming ubiquitous, quality control of sample prep, storage, sequencing, and data processing must become more standardized. Regardless, as the amount of viral and microbiome research and WGS studies continue to rise, unmapped read space will be a promising untapped data source for exploring the mysterious contributions of non-human sequences to the human form.

References

1. Moustafa, Ahmed, et al. "The blood DNA virome in 8,000 humans." PLoS pathogens 13.3 (2017): e1006292.
2. Wong, Karen HY, Michal Levy-Sakin, and Pui-Yan Kwok. "De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations." Nature communications 9.1 (2018): 1-9.
3. Ruzzo, Elizabeth K., et al. "Inherited and de novo genetic risk for autism impacts shared networks." Cell 178.4 (2019): 850-866.
4. Brister, J. Rodney, et al. "NCBI viral genomes resource." Nucleic acids research 43.D1 (2015): D571-D577.
5. Turnbaugh, Peter J., et al. "The human microbiome project." Nature 449.7164 (2007): 804-810.

Impact of COVID-19 Pandemic on the Use of Telemedicine in Academic Medical Center in New York City

Wanting Cui, MA, Joseph Finkelstein, MD, PhD
Icahn School of Medicine at Mount Sinai, New York, NY

Introduction:

De-identified data set extracted from electronic health records at Mount Sinai Health System was used to analyze telemedicine services between January 2019 and July 2020. It contained 136,497 unique patients and 225,136 sessions during the 1.5-year period. The COVID-19 pandemic significantly changed the demand for telemedicine services. The average age of patients increased since the pandemic and there were significantly more White patients using the service than African American patients. In addition, telemedicine has expanded to more disciplines since the pandemic. Reasons for these disparities are yet to be established, and will be explored in future analyses.

Method:

A de-identified study dataset was generated by querying electronic health records at the Mount Sinai Health System to identify all patients who used telemedicine services between January 2019 and July 2020 with subsequent removal of protected health information. Variables in the dataset includes patients' demographics, encounter diagnosis, medical history, and the corresponding care providers' primary specialty. We only included patients who are 18 years and older, and eliminated patients with missing variables. The different types of diagnoses were all coded in ICD-10 codes. We defined encounter month as the month of encounter since January 2019, and quarter of encounter as the quarter of encounter since January 2019. We calculated patients' Charlson comorbidity index based on patients' medical history, using the patients' age and ICD-10 code [1]. We also mapped patients' primary encounter diagnosis into corresponding body systems using ICD-10 codes [2]. In addition, since a patient could have multiple issues to resolve in one session and each patient could have multiple visits in a period of time, we defined that each patient will have one session with each health care provider per day.

We performed comparative analyses of the number of telemedicine sessions over time, patients' demographics, comorbidities, primary diagnoses and health care providers' primary specialties. All analyses were performed in Python (Python version 3.7). All statistical tests were two-sided, unless otherwise specified, with $p<0.05$ being considered statistically significant. March 2020 was used in analysis as pandemic commencement time in New York City.

Results:

The dataset contained 136,497 unique patients, and 225,136 telemedicine sessions that were conducted during the year and half period. There was a significant increase in demand for telemedicine services triggered by COVID-19 pandemic. In 2019, on average patients used the services 273.5 times per month. In contrast, patients on average have used the services around 30,000 times per month since the pandemic commencement in New York City. According to Figure 1, the demand of telemedicine increased over 100-fold in March 2020 and reached the peak in May 2020. There were over 66,000 services in that month.

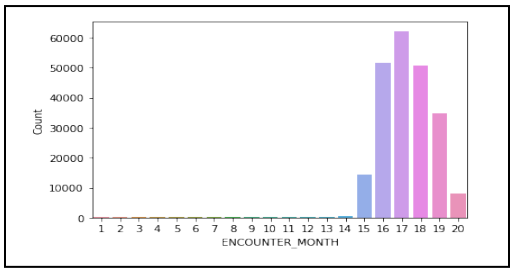


Figure 1. Number of sessions by month.

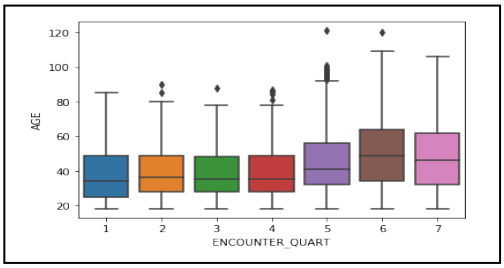


Figure 2. Box plot of patients' age by calendar quarter.

Patients who used telemedicine prior to the COVID-19 pandemic were younger, with an average age of 39.96 years old (Figure 2). Since the pandemic, older patients started to use the telemedicine service, and the average age increased to 50.45 years old. In addition, in 2019, patients who used telemedicine had a significantly lower comorbidity index (0.51), compared to the comorbidities index (1.24) of patients in 2020. Furthermore, White patients (46.5%) were more likely to use the telemedicine service than African American patients (46.5%) both prior and after the start of pandemic.

In terms of body system, prior to the pandemic, patients with mental disorders and digestive system diseases used the telemedicine services the most. In the fifth quarter (January to March 2020), there were more patients with infectious diseases or respiratory problems. Furthermore, patients with immunity disorder, musculoskeletal diseases, skin and subcutaneous tissue problems were more likely to use the telemedicine service since the pandemic.

Table 1. Mapped body systems based on patients' primary diagnoses.

	Quarter of Encounter						
Body System	1	2	3	4	5	6	7
1. Infectious and parasitic disease	5.99%	4.73%	3.19%	2.28%	11.21%	2.54%	1.65%
2.Neoplasms	0.73%	0.16%	0.73%	0.98%	3.45%	5.45%	5.14%
3. Endocrine, nutritional, and metabolic diseases and immunity disorders	2.36%	5.22%	4.50%	6.62%	5.72%	9.52%	9.80%
4. Diseases of blood and blood-forming organs	0.18%	0.00%	0.15%	0.00%	0.70%	0.98%	0.90%
5. Mental disorders	28.31%	22.35%	24.67%	18.55%	12.91%	11.76%	17.37%
6. Diseases of the nervous system and sense	5.44%	7.50%	9.43%	8.89%	5.66%	8.28%	9.85%
7. Diseases of the circulatory system	1.27%	2.45%	3.19%	1.08%	4.08%	8.62%	5.97%
8. Diseases of the respiratory system	10.16%	6.04%	3.63%	6.29%	15.27%	5.58%	4.21%
9. Diseases of the digestive system	31.40%	27.90%	26.71%	34.60%	7.28%	5.48%	5.88%
10. Diseases of the genitourinary system	0.73%	2.61%	2.76%	2.39%	3.14%	5.01%	4.46%
11. Complications of pregnancy, childbirth, and the puerperium	0.18%	0.16%	0.29%	0.11%	0.81%	0.96%	0.83%
12. Diseases of the skin and subcutaneous tissue	1.09%	2.28%	1.74%	1.74%	1.90%	5.15%	4.98%
13. Diseases of the musculoskeletal system	1.81%	3.43%	3.05%	2.71%	5.09%	8.16%	7.35%
14. Congenital anomalies	0.18%	0.00%	0.00%	0.54%	0.20%	0.32%	0.32%
15. Certain conditions originating in the	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.01%
16. Symptoms, signs, and ill-defined conditions	7.62%	8.48%	10.45%	8.35%	16.62%	14.35%	12.30%
17. Injury and poisoning	0.73%	1.63%	1.45%	1.52%	1.08%	1.76%	1.31%
18. Factors influencing health status and contact with health services	1.63%	4.57%	3.77%	3.15%	4.83%	5.90%	7.47%
Body System None	0.18%	0.49%	0.29%	0.22%	0.05%	0.17%	0.18%

Conclusion:

The COVID-19 pandemic changed the landscape of telemedicine drastically. The demand for the service increased significantly and it has reached a wider range of patients. The average age of patients increased since the pandemic and there were significantly more White patients using the service than African American patients. In addition, telemedicine has expanded to more medical disciplines since the pandemic. In future studies, we plan to study the accessibility of telemedicine to older adults and the disparities of telemedicine usage between different races. Thus, future analyses of telemedicine are warranted.

Reference

1. Ho CH, Chen YC, Chu CC, Wang JJ, Liao KM, Age-adjusted Charlson comorbidity score is associated with the risk of empyema in patients with COPD, *Medicine (Baltimore)* **96(36)** (2017), e8040.
2. H, Ghali WA, New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality, *J Clin Epidemiol* **57(12)** (2004), 1288-94.

The Impact of Data Availability on Algorithmic Diabetes Type Detection using Electronic Health Records: The SEARCH for Diabetes in Youth Study

Franck Diaz-Garelli, Ph.D.,[1] Kristin M. Lenoir, MPH,[2] Lynne E. Wagenknecht, DrPH,[2] Dana Dabelea, MD, PhD,[3] Tessa Crume, PhD,[3] Catherine Pihoker, MD,[4] Eva Lustigova, MPH, [5] Jasmin Divers, PhD,[6] Brian J. Wells, M.D. Ph.D. [2]

[1]University of North Carolina at Charlotte, Charlotte, NC, [2]Division of Public Health Sciences, Wake Forest School of Medicine, Winston Salem, NC, [3]University of Colorado Denver, Aurora, CO, [4]University of Washington, Seattle, WA, [5]Kaiser Permanente Research, Pasadena, CA [6]NYU Long Island School of Medicine, Mineola, NY, USA

Introduction

Disease-based surveillance using electronic health records (EHR) relies heavily on accurately identifying patients with medical conditions in EHRs. Multiple methods have been developed to support this research task, and patient selection algorithms which rely heavily on structured data are broadly used due to ease of implementation. These approaches often utilize structured diagnosis data,[1] though data availability, completeness, and quality may vary over time and between patients.[2] Still, the literature lacks examples with varying levels of data completeness and the numerous definitions that can be ascribed to completeness.[3] This work is necessary to inform algorithm development and implementation for "real word," EHR-based surveillance applications.

To address this gap, we explored the impact of EHR data availability on a diagnosis-based patient classification algorithm developed within SEARCH as a test case. SEARCH validated a rule-based ICD-10 diagnosis code algorithm to identify diabetes status (yes/no) and diabetes type (type 1, type 2 and other type) from diabetes-related diagnosis data available in EHRs of pediatric patients.[4] This algorithm achieved an accuracy of 0.955 in correctly classifying status and type. It had a positive predictive value (PPV) of 0.969 to determine diabetes status, and a PPV of 0.980 to classify type 1 cases when compared to an adjudicated gold-standard diabetes status and diabetes type. We sought to identify variability in algorithm performance metrics for correctly classifying diabetes type based on cumulative diagnosis code data over time. Our overarching goal was to establish minimum data availability requirements to achieve acceptable predictive accuracy.

Methods

We employed a data set of youth (<20 years old in 2017) procured by 3 different sites in the US that participate in the SEARCH Study: Cincinnati Children's Hospital, Cincinnati, OH; Seattle Children's Hospital, Seattle, WA; and Children's Hospital Colorado, Denver, CO. From a population of youth in the EHR (n=729,272), a sub-population of 8,682 potential cases of diabetes was identified with the requirement of an in-person encounter and some EHR-based evidence of diabetes occurring in 2017: hemoglobin A1c >=6.5% (>=48 mmol/mol), or fasting plasma glucose >=126 mg/dl (>=7.0 mmol/L), or random plasma glucose >=200 mg/dl (>=11.1 mmol/L),or at least one diabetes-related ICD-10 code (E08-E13), or a diabetes-related medication. EHR data (inpatient and outpatient) for potential 2017 prevalent diabetes cases were extracted as far back in time as records were available. Study coordinators reviewed the medical records of each of these patients according to the SEARCH protocol to identify a gold standard diabetes status and type. The current study was limited to 5,426 patients who were considered probable diabetes cases based upon the rule-based ICD-10 algorithm (presence of 2 or more cumulative diabetes-related ICD-10 codes as of 12/31/2017). The rule-based ICD-10 algorithm was also used to predict diabetes type (type 1, type 2, or other diabetes type) based upon a preponderance of cumulative diabetes type codes.

We sub-divided single-patient EHR records over time by number of months since the patient's first ICD-10 diabetes-related code to simulate stages/levels of data availability as would naturally evolve in an active clinical EHR setting. For each subset, we also calculated secondary completeness metrics[3] as the number of months with new diagnosis data (i.e., data density) and the number of diagnosis codes entered accumulated (i.e., documentation completeness). For each subset, we applied the rule-based ICD-10 algorithm to predict patients' diabetes type and assessed the performance using standard classification metrics (e.g., sensitivity[Se], specificity[Sp], positive predictive value [PPV] and negative predictive value [NPV]). We selected an optimal performance threshold by maximizing Youden's J (i.e., J=Se+Sp-1).[5] Performance measures were based upon a classification matrix where predicted diabetes type was compared to the gold standard type. False positives (non-diabetes cases) identified as cases by the rule-based algorithm were automatically considered misclassified by type. We employed R's 'cutpointr' package[6] for maximization. Visual

exploration and analyses were performed using Tableau (version 2020.2, Tableau Software, Inc., Seattle, WA). Statistical analyses and data manipulation were done in R version 3.6.2 and RStudio (version 1.2.5033, RStudio, Inc., Boston, MA). The study was approved by Institutional Review Boards with waivers of informed consent and in accord with the Health Insurance Portability and Accountability Act authorization.

Results

Our dataset contained 5,426 unique individuals identified by the rule-based ICD-10 algorithm. Of those, 161 were false positives, 4,711 were type 1 diabetes cases, 381 were type 2 diabetes cases and 173 were other type diabetes cases. After synchronizing each patient's timeline based upon first diagnosis, the data were sub-divided into 220 different datasets containing incremental number of months of data (e.g., month 1, month 1-2, month 1-3, etc.) for all patients.

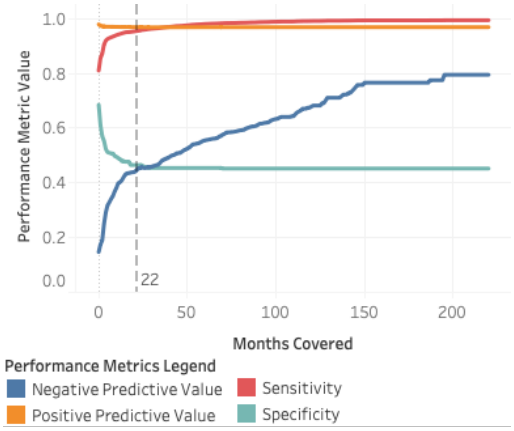


Figure 1 – Algorithm performance metric change over months of available data. Optimal performance threshold marked at 22 months of available data.

Running the ICD-10 rule-based classification algorithm over these 220 monthly slices, we observed that predictive performance metrics (correctly classifying diabetes type) varied with data availability (Figure 1). Se increased from 81% in the first month of data to 99% by month 220. Sp decreased from 68% to 45% due to the increasing number diagnoses for non-diabetic patients. NPV increased from 15% to 80%. PPV did not change (97% to 96%). We determined the optimal cutpoint for the number of months since first diabetes diagnosis code by Youden's J (J=0.026) was 22 months (Figure 1). Using this cutpoint, accuracy was 86% with Se=91%, PPV=94%. However, the Sp and NPV were very low (11% and 7%, respectively). An AUC value of 0.51 confirmed the weakness of this predictor.

Exploring our two other definitions of completeness improved the threshold selection task and allowed for a deeper understanding of data availability. On one hand, a completeness definition based on number of months with new diagnoses (i.e., the data density or number of time slices with new data) improved the balance between prediction metrics while providing more concrete guidance on minimal data completeness requirements. The optimal cutpoint was set at 2 distinct months with diagnosis observations (cutpoint=1.5, J=0.21), yielding an accuracy of 81% with Se=0.83, PPV=0.96. Sp (38%) and PPV (13%) improved, and an AUC of 0.60 underscored this improvement. On the other hand, exploration of a completeness metric based on the number of recorded diagnosis codes (i.e., documentation completeness) revealed an optimal cut point of 10 diagnosis codes (J=0.33), an accuracy of 85%, a Se of 88%, and a PPV of 96%. Sp and PPV were further improved (46% and 19%, respectively). An AUC of 0.7 revealed significantly improved prediction performance when the threshold was set using this definition of completeness.

Conclusion

Our analysis showed that data availability and completeness can substantially affect the performance of EHR-based patient classification algorithms. Our results indicate that completeness definitions impact algorithm performance and highlight the need to explore multiple definitions and select the most "fit for purpose" definitions[7] upon implementation. Given the ever-changing nature of EHR data and inconsistent completeness across patient charts,[3] this sort of phenomenon must be accounted for when evaluating predictions relying on such data. Our findings provide preliminary evidence that "safety net" EHR data completeness thresholds must be defined by algorithm developers to prevent misclassification that can lead to downstream analytical errors. Our results also provide a primer on the impact and importance of data quality metric definition for algorithm evaluation, tuning and calibration.

References

1. Zhong, V. W. *et al.* An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: the SEARCH for Diabetes in Youth Study. *Journal of the American Medical Informatics Association* **23**, 1060–1067 (2016).
2. Weiskopf, N. G., Rusanov, A. & Weng, C. Sick Patients Have More Data: The Non-Random Completeness of Electronic Health Records. *AMIA Annual Symposium Proceedings* **2013**, 1472–1477 (2013).
3. Weiskopf, N., George Hripcsak, Swaminathan, S. & Weng, C. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics* **46**, 830–836 (2013).
4. Wells, B. J. *et al.* Detection of Diabetes Status and Type in Youth Using EMRs. *Diabetes Care (In Press)* (2020).
5. Fluss, R., Faraggi, D. & Reiser, B. Estimation of the Youden Index and its associated cutoff point. *Biom J* **47**, 458–472 (2005).
6. Thiele, C. *cutpointr: Determine and Evaluate Optimal Cutpoints in Binary Classification Tasks*. (2020).
7. Holve, E., Kahn, M., Nahm, M., Ryan, P. & Weiskopf, N. A comprehensive framework for data quality assessment in CER. *AMIA Jt Summits Transl Sci Proc* **2013**, 86–88 (2013).

Performance of Automatic De-identification Across Different Note Types

Nicholas Dobbins[1], David Wayne[2], Kahyun Lee[4], Özlem Uzuner, Ph.D.[4], Meliha Yetisgen, Ph.D.[1,3]
[1]Biomedical and Health Informatics, [2]School of Medicine, [3]Linguistics, University of Washington, Seattle, WA
[4] Information Sciences and Technology, George Mason University, Fairfax, VA

Introduction

Free-text clinical notes detail all aspects of patient care and have great potential to facilitate quality improvement and assurance initiatives as well as advance clinical research. However, concerns about patient privacy and confidentiality limit the use of clinical notes for research. As a result, the information documented in these notes remains unavailable for most researchers. De-identification (de-id), i.e., locating and removing personally identifying protected health information (PHI), is one way of improving access to clinical narratives. However, there are limited off-the-shelf de-identification systems able to consistently detect PHI across different data sources and medical specialties. In this abstract, we present the performance of a state-of-the art de-id system called NeuroNER[1] on a diverse set of notes from University of Washington (UW) when the models are trained on data from an external institution (Partners Healthcare) vs. from the same institution (UW). We present results at the level of PHI and note types.

Dataset

UW-Dataset: We created a dataset of 600K notes from patients who received treatment at University of Washington Medical Center and Harborview Medical Center between 2007 and 2017. From this dataset, we randomly sampled 1000 notes from 10 note types (100 notes each). The selected subset of 1000 notes contains a total of 1,890,849 tokens from the following note types: (1) admit notes, (2) discharge notes, (3) emergency department (ED) notes, (4) nursing notes, (5) pain management (PM) notes, (6) progress notes, (7) psychiatry notes, (8) radiology notes, (9) social work notes, and (10) surgery notes. The length of notes varies across note types (token count - avg: 1890.9; max: 2492.58 (progress notes); min: 948.55 (radiology); std-dev: 706.13). We annotated this set with 25 personal health identifiers (PHI). 8 graduate students from UW Biomedical Health Informatics departments and 2 medical students from UW School of Medicine completed the annotation. All notes were double-annotated and all conflicts were resolved. We grouped the 25 PHI under 6 PHI types (e.g., NAME: patient, doctor, user name). Table 1 includes the entity counts per PHI found in these notes.

PHI Type	PHI	Admit Notes	Discharge	ED	Nursing	Pain Mgmt.	Progress	Psychiatry	Rad.	Social work	Surgery
NAME	Patient	360	353	252	11	419	598	1351	24	844	31
	Doctor	531	1505	2791	54	3475	1075	894	319	1334	177
	UserName	0	0	1	0	3	0	0	0	0	0
LOCATION	Room	3	108	4	62	4	38	114	0	173	1
	Department	81	662	72	5	222	282	89	170	196	38
	Hospital	383	1189	327	192	995	694	609	128	1326	77
	Organization	56	140	45	5	166	136	328	3	1493	0
	Street	7	170	4	0	1110	117	10	84	36	0
	City	116	162	45	0	939	136	93	21	313	5
	State	41	94	6	1	833	85	28	21	183	1
	Country	5	1	0	0	3	12	61	0	8	0
	Zip	0	26	1	0	410	44	4	21	8	0
	Other	19	23	6	0	588	40	13	0	153	0
AGE	Age	335	241	534	9	256	263	269	73	190	207
DATE	Date	4107	3463	2267	877	4566	4195	3052	683	903	2829
CONTACT	Phone	380	1786	259	6	384	287	369	120	1740	55
	Fax	0	17	0	0	6	62	5	0	5	0
	Email	0	2	0	0	0	7	2	0	46	0
	URL	1	151	66	8	61	38	16	0	5	4
IDs	MRN	2	2	4	0	0	13	1	5	48	3
	ID-Number	14	105	288	0	76	10	6	5	26	47
	Account	2	0	0	0	0	0	0	0	0	0
	Health Plan	0	8	0	0	0	5	9	0	119	0
PROFESSION	Profession	84	61	13	0	129	63	85	0	80	0
	Total:	6527	10269	6985	1230	14645	8200	7408	1677	9229	3475

Table 1. Annotation statistics for PHI entities across different not types.

As can be seen from the table, each note type has different distributions of PHI content. Pain management, discharge, and social work notes are the top three note types for PHI content density. Radiology and nursing notes include significantly fewer PHI when compared to other note types. There is also variance in the density of PHI types across note types. For example, patient name is mentioned more frequently in psychiatry and to a lesser extent social work notes, while age, medical record number, and profession are rarely mentioned in nursing and radiology notes.

i2b2 Dataset: We used the 976 longitudinal notes, including doctor's notes, discharge summaries, doctor-patient correspondence, and lab results, available in the training set of 2014 i2b2/UTHealth shared task[2] as external data to measure the domain adaptability between Partners Healthcare and UW. Our UW data were annotated according to the annotation guidelines associated with the 2014 i2b2/UTHealth shared task data, providing two consistently annotated data sets for de-identification across institutions.

Methods

We define de-identification as a named entity recognition (NER) task. We evaluated the performance of an NER system, called NeuroNER[1], on our dataset. NeuroNER uses long short-term memory (LSTM)-based recurrent neural networks (RNN) for non-overlapping label prediction and achieves state-of-the-art performance on a number of tasks.

In our experimental setup, we split the 1000 annotated UW notes with a 4:1 ratio (training set: 800, test set: 200). We ran the following experiments: (1) trained and tuned on external data (i2b2 de-id dataset[2]) and tested on UW-test set, (2) trained and tuned on in-house data (UW-training set) and tested on UW-test set, and (3) trained and tuned on external and in-house data (i2b2 de-id dataset + UW-training set) and tested on UW-test set. We compared performance differences across models at the PHI and note type levels.

Results and Conclusion

Table 2 presents extraction results at the PHI type level. Overall training only with i2b2 notes provides a reasonable starting point for de-identification on UW data. However, performance is significantly better when the models were trained with the UW training set. The combined i2b2 + UW-training set achieved the best performance across 5 of 6 PHI types, with a mean per-label F1-score improvement of 1.2 over training with UW notes alone. Among PHI types, dates showed the best performance, with 97.5 F1-score, while professions showed poorest performance with 50.0 F1-score. We believe the comparatively high performance of date identification can be attributed to (1) the large number of date instances w.r.t. other labels, and (2) the relatively small number of date patterns (e.g., '2019-4-8', or 'March 8th'). For the worst-performing type, profession, we believe the main reason for its poorer performance is due to a smaller number of training samples.

Table 3 presents F1-scores at the note type level. Among note types, the addition of UW training notes to i2b2 notes boosted F1-scores by a mean 12.3 (std-dev: 5.8). Radiology notes showed the greatest improvement between training with i2b2 notes versus UW notes (+27.5) – i2b2 data contained no radiology notes. Across nearly all note types, the relatively large number of dates raises the overall F1 scores significantly, despite comparatively lower performance of other types, such as professions. Our results suggest that training NeuroNER using multi-institutional corpora for de-identification tasks can improve identification of certain types of PHI.

Training set	Name	Location	Age	Date	Contact	IDs	Profession
I2b2	75.9	59.8	81.0	92.2	66.5	40.2	25.0
UW	91.0	**83.1**	89.7	**97.4**	**86.9**	82.8	45.6
I2b2 + UW	**92.6**	**83.0**	**91.4**	**97.5**	**87.0**	**83.6**	**50.0**

Table 2. De-identification performance (F1-score) across different types of PHI.

Training set	Admit	Discharge	ED	Nursing	Pain Mgmt.	Progress	Psychiatry	Rad.	Social work	Surgery
I2b2	88.6	82.41	83.1	90.0	82.5	82.9	82.6	70.2	79.6	87.2
UW	97.2	94.4	93.7	97.7	95.3	93.6	95.9	97.7	90.8	95.8
I2b2 + UW	96.7	92.8	95.6	95.9	97.4	94.2	95.5	97.5	91.0	95.6

Table 3. De-identification performance (F1-score) across different note types.

Acknowledgements

This study was supported by the National Library of Medicine under Award Number R15LM013209 and by the National Center For Advancing Translational Sciences of National Institutes of Health under Award Number UL1 TR002319. Experiments were run on computational resources from the UW Department of Radiology.

References

1. Dernoncourt, Franck, Ji Young Lee, and Peter Szolovits. 2017. "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks." arXiv preprint 1705.05487.
2. Stubbs, Kotfila C, Uzuner Ö. 2015. "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1." J Biomed Inform 58 Suppl: S11-S19.

Community-based Organizations as Partners in Addressing Racial Disparities in Infant Mortality: Establishing an Infrastructure for Routine Data Collection.

Naleef Fareed, PhD[1], Christine Swoboda, PhD[1], John Lawrence, MS[1], Tyler Griesenbrock, BA[1], and Timothy Huerta, PhD[1]

[1]Ohio State University, College of Medicine, Columbus, Ohio

Introduction

Ohio was ranked ninth-worst in the U.S. for infant mortality rate (IMR) in 2018 and has a significant disparity in the IMR between black and white infants.[1] A multi-agency effort to address infant mortality (IM) in Ohio, known as the Ohio Equity Institute (OEI), was created to focus infant vitality efforts in nine Ohio counties. Coordinating these entities helped bring focus to the number of independent efforts, but also highlighted an absence of consistent data collection and infrastructure to facilitate evaluation. To address this lack of consistency, the Ohio Department of Medicaid (ODM) developed the OEI Community-Based Organization (CBO) Evaluation project. The project's main goal was to create a statewide system for routine data collection and analysis to determine the extent to which the selected interventions serve high-risk pregnant women and assess the effect of these interventions on health care utilization and birth outcomes among these CBO participants. Active data collection processes have been shown to improve prevention programs and increase evaluative capacity to address multiple public health problems.[2-3] A robust data collection system for OEI was determined to have the following characteristics: 1) multimodality of data collection, 2) a standardized data model, and 3) systematic educational outreach efforts to train those reporting data. The Ohio State University research team, composed of health and bioinformatics subject matter experts, was tasked with developing and deploying a multimodal data collection system for CBOs across the state of Ohio.

Methods

The OEI CBO Evaluation project focused on nine Ohio counties with large disparities in IMR and significant urban populations. Our needs assessment in 2018 involved interviews with CBOs and literature reviews, in which we learned about existing data collection activities. The research team at OSU used this needs assessment process to inform the development of the OEI data collection system, which gathers data on demographics, environmental and behavioral risks, and care received by the mother and infant. This data is meant to supplement the birth data in Vital Statistics without being duplicative, and especially focuses on risk factors a mother in a CBO may have including housing status, smoking/drug/alcohol use, depression, stress, financial difficulties, vitamin intake, and food insecurity.

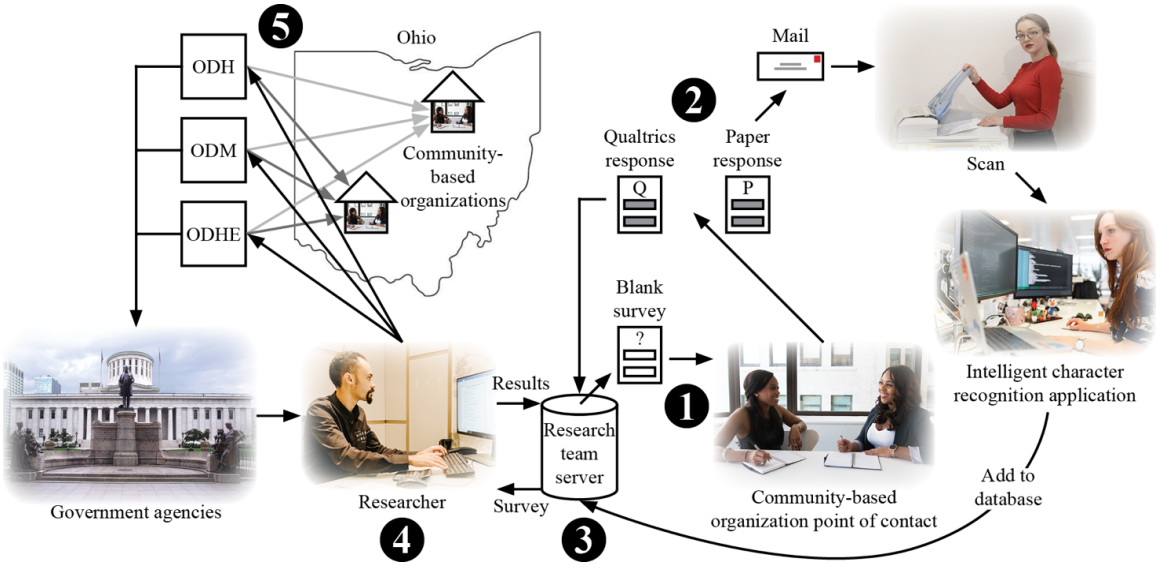


Figure 1. Overall vision for the OEI data collection infrastructure.

Figure 1 illustrates our initial vision for the data collection infrastructure and shows how the data collection, curation, and reporting activities are integrated. The first step involves a CBO provider collecting participant

information using a collection form. The provider then uses a data entry mechanism – an online data entry system, scanning and faxing forms, mailing forms, secure email of spreadsheet data or forms, or uploading to a secure online portal – to report the data to the research team. The third step involves the curation of a database on a server that, in step four, can be used by the team to access data via a portal. Step five illustrates how the team can use the database to develop a conceptual data model for use by researchers and government agencies to develop queries for reports and dashboards. The CBOs can also use this data to interact with the researchers and government agencies for decision making purposes.

Between February and July 2018, our team at OSU developed and deployed the OEI data collection infrastructure. Critical milestones were developing data collection forms, piloting the data collection system, testing system integration, and making refinements based on CBO feedback. Data collection began in September 2018 for 30 programs, with more programs enrolling as training and data use agreements were finalized. The OEI common data model consists of elements from our data collection infrastructure (demographic and risk data) linked to state datasets that provide detailed medical and birth data (i.e., Ohio Vital Statistics and Medicaid Claims). The primary objectives of developing our data model for OEI were 1) to generate reports that quantify the funded CBOs based on key metrics and 2) to dynamically visualize these key metrics on a Tableau dashboard.

The OSU team trained CBO employees on how to use the data collection materials in July and August 2018. The training consisted of webinars with demonstrations of how to use the Qualtrics data portal, a validated Excel spreadsheet, and the paper forms. After these demonstrations, CBOs were given a choice of how to submit data between those options, and paper forms, the spreadsheet file, or login information was given to the CBO to start data collection. After the first month of data collection, surveys and phone calls were conducted to discuss this process. CBOs were given the opportunity to change data submission preferences, provide input about portal changes desired, and receive additional training.

Results

To develop the OEI system and integrate its data with our data model, we identified critical assumptions (e.g., capacity for CBOs to report data) that provided us with the basis for our multimodal data collection and integration approach. In our presentation, we will discuss the workflow that we developed to account for the idiosyncratic constraints that exist across the OEI CBOs, and the integration of the data across CBOs with other key databases. The data model created by our research team has been leveraged to generate data visualizations using Tableau dashboards. We will present success metrics that include details about the data we have collected from approximately 20,000 participants to date. Metrics will be evaluated using descriptive statistics regarding enrollment, risk factors, and data missingness.

Discussion

Data collection continues to be one of the biggest challenges for community-based organizations.[4] Even when the CBOs have systems in place, there are data points that are difficult to collect for certain programs. In our case, programs that meet occasionally or exist mainly to refer people to services often see participants only once or twice. They do not have time to collect extensive data, nor has a trusting relationship been built with participants. Variables frequently missing include information about the other biological parent, risk factors, and information about prenatal care attendance. This is not missing systematically; these programs ask the questions and do not receive answers from all of the participants. These data quality concerns have motivated continuous changes to the data collection system.

This data collection system allows assessment of risk factors for women enrolled in public health programs, which will help describe populations experiencing disparities and influence interventions. This data collection system also supplements birth data, allowing comparisons of clinical outcomes from birth data with behavioral and environmental risk information to assess who is more likely to experience poor outcomes. Although there are national reporting requirements for certain IM prevention programs, there is little comparison of programs or pooling of program data. By collecting data for multiple programs in Ohio, the effects of IM prevention efforts can be more formally assessed and programs can be compared. IM is a rare outcome and difficult to assess statistically; pooling data will allow for comparisons of birth outcomes throughout the state and between program types. These comparisons will help highlight which program types and components have the most potential for improving outcomes. The common data model

created is generalizable across multiple maternal and infant health programs in Ohio, and could be shared with other state or local efforts to help them adopt similar data collection systems and compare efforts.

Conclusion

The OEI data collection infrastructure our team designed and deployed is still in its early stages of system maturity. Our team continues to learn from its implementation and use as the system evolves. The system, although requiring more effort from CBOs, is demonstrating signs of collective action at the local and state levels to better coordinate and share information on how to best use programmatic resources to reduce IM and its associated disparities in Ohio.

References

1. Centers for Disease Control and Prevention. 2020. Infant Mortality Rate by States. Available at: https://www.cdc.gov/nchs/pressroom/sosmap/infant_mortality_rates/infant_mortality.htm
2. Weir SS, Baral SD, Edwards JK, Zadrozny S, Hargreaves J, Zhao J, Sabin K. Opportunities for enhanced strategic use of surveys, medical records, and program data for HIV surveillance of key populations: Scoping review. JMIR public health and surveillance. 2018;4(2):e28.
3. Dixit S, Arora NK, Rahman A, Howard NJ, Singh RK, Vaswani M, Das MK, Ahmed F, Mathur P, Tandon N, Dasgupta R. Establishing a demographic, development and environmental geospatial surveillance platform in India: planning and implementation. JMIR public health and surveillance. 2018;4(4):e66.
4. Holden RJ, Scott AM, Hoonakker PL, Hundt AS, Carayon P. Data collection challenges in community settings: Insights from two field studies of patients with chronic disease. Quality of Life Research. 2015 May 1;24(5):1043-55.

Clinical Impact of Satisfying Group Fairness Constraints in Revised Pooled Cohort Equations

Agata Foryciarz[1,2], Stephen R Pfohl[2], Nigam H Shah, PhD, MBBS[2]

[1]Department of Computer Science, Stanford University, Stanford, CA;

[2]Center for Biomedical Informatics Research, Stanford University, Stanford, CA

Introduction

In 2013, ACC/AHA guidelines started recommending the use of a 10-year risk model for Atherosclerotic Cardiovascular Disease (ASCVD) to guide statin prescriptions for individuals at risk of developing ASCVD[1]. The risk stratification model, Pooled Cohort Equations (PCEs), showed improved calibration compared to previous models[2], but still underestimated risk for Black patients[3], and overestimated risk for female patients[4]. To address those disparities, updated PCEs, which use newer statistical methods and representative datasets were proposed[3]. It is also possible to apply algorithmic fairness methods to reduce variability in error rates of the risk score across groups[5]. However, it remains unclear what effect such fairness methods would have on treatment decisions[6]. In this work, we use a common fairness criterion - equalized odds - to build fairness-constrained ASCVD risk prediction models using updated statistical methods and representative pooled cohorts to examine its effect on the changes in the risk categories assigned to patients and the resulting treatment decisions.

Methods

We use four longitudinal cohorts used to derive the original Pooled Cohort Equations: ARIC (Atherosclerosis Risk in Communities Study, 1987-2011), CARDIA (Coronary Artery Risk Development in Young Adults Study, 1983-2006), CHS (Cardiovascular Health Study, 1989-1999), FHS OS (Framingham Heart Study Offspring Cohort, 1971-2014), as well as two modern longitudinal cohorts: MESA (Multi-Ethnic Study of Atherosclerosis, 2000-2012), and JHS (Jackson Heart Study, 2000-2012).

We include individuals aged 40 to 79, self-identifying as White or Black, with no past history of myocardial infarction, stroke, coronary bypass surgery or angioplasty, congestive heart failure or atrial fibrillation. We extract variables for total cholesterol, HDL cholesterol, treated and untreated systolic blood pressure, BMI, diabetes, cholesterol medication, age, as well as binary sex and race, recorded at the initial examination. An ASCVD event is defined as the presence of myocardial infarction, lethal or non-lethal stroke, or lethal coronary heart disease within 10 years of the initial examination. We exclude individuals with extreme values of systolic blood pressure, total cholesterol and high-density lipoprotein cholesterol, and remove records with missing covariates. We split data into train (81%), validation (9%) and test (10%) sets through sampling stratified by race, sex, study, presence of outcome and censoring.

To predict the risk of an ASCVD event within 10 years, we build logistic regression models with inverse propensity score weighting (IPSW) to account for censoring[3], and include two-way variable interactions. We estimate propensity scores w with a Kaplan-Meier estimator. To build fairness-constrained models, we consider the equalized odds criterion, which requires that the predictions be independent of the group (of which we consider four: Black women (BW), White women (WW), Black men (BM) and White men (WM)), conditioned on the observed outcome[7]. This results in a loss function of the form $L(w^T X, y) + \lambda R$, where $w^T X$ are the weighted covariates, L is the cross-entropy loss, and R is a regularized objective. As in our prior work[6], we use a regularized objective that penalizes the difference in the mean predicted probability of the outcome for each group with the marginal distribution of predicted probabilities within strata defined by observed outcomes.

If this criterion is satisfied exactly, the ROC curves for all groups will match, implying that both sensitivity and specificity are equalized across groups at all decision thresholds. Among individuals who would have developed ASCVD if untreated, the rate at which treatment decisions are made at some risk level would be the same across groups. We build 10 fairness-constrained models, varying the parameter λ, which controls the extent to which violation of equalized odds is penalized, uniformly on the log scale from 0.001 to 10.

We use the 2019 ACC/AHA guidelines to define risk categories[8]: Low/Borderline Risk (below 7.5%), Intermediate Risk (7.5%-20%) and High Risk (above 20%). The guidelines recommend prescription of statins for patients at High Risk, as well as for some patients at Intermediate risk, following a clinician-patient discussion.

Results

The baseline revised PCE model achieved high performance on the validation set, which varied between groups (Fig 1). For fairness-constrained models, group-level performance varied without dropping significantly for $\lambda \in (0.001, 0.1)$; however, AUC, precision and loss degraded for higher values of λ (Fig 1). Across all groups, across increasing values of λ, the implementation of fairness constraints reclassified individuals into the Intermediate Risk category, and away from Low/Borderline and High Risk categories (Fig 2).

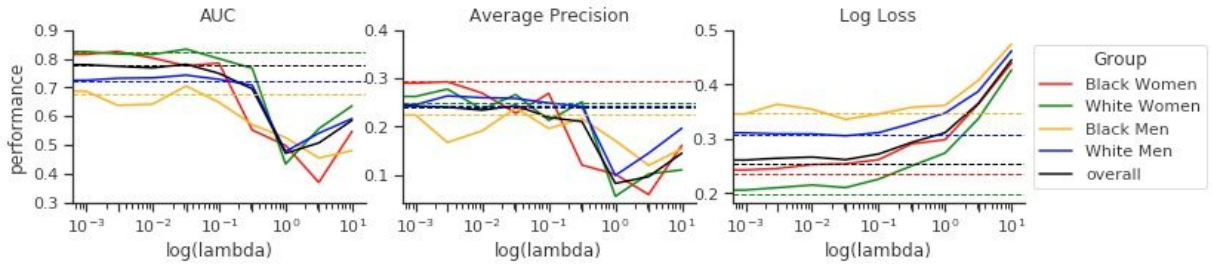


Figure 1. **Group-level performance measures** as a function of the parameter λ. Dashed lines correspond to the result for the unpenalized training procedure ("baseline model").

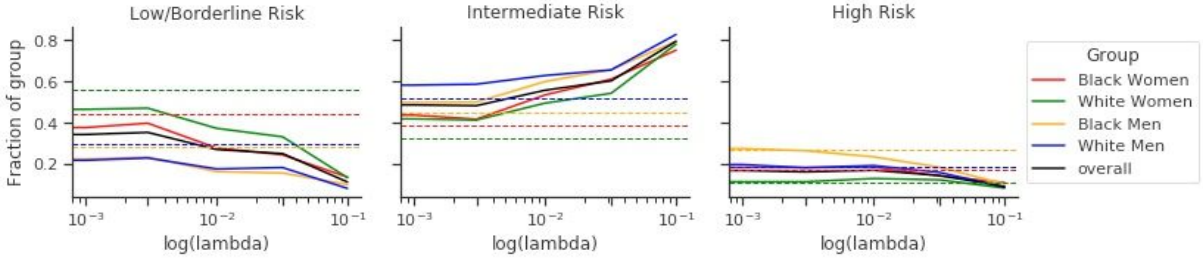


Figure 2. **Fraction of groups classified into risk categories** as a function of the parameter λ. Dashed lines correspond to the result for the unpenalized training procedure ("baseline model").

Discussion

We describe the changes in the risk category assignment that result from the re-deriving the Pooled Cohort Equations augmented with a fairness constraint in the form of an equalized odds penalty. While there are heterogenous changes across four demographic groups examined, our results suggest that imposing equalized odds would lead to more people being considered for treatment at the Intermediate Risk category, and fewer people treated at the High Risk category, resulting in a recommendation of statin treatment for more individuals who would not end up developing ASCVD, and placing uncertainty on treatment decisions for people that would be classified as High Risk at baseline that would be likely to benefit from treatment. Further investigation is needed to examine the net-effect of imposing fairness constraints on the ASCVD risk-estimation models. We note that imposing fairness criteria should not replace efforts to ensure diversity in training datasets.

References

1. Stone, N. J. *et al.* 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* **63**, 2889–2934 (2014).
2. Qureshi, W. T. *et al.* Impact of Replacing the Pooled Cohort Equation With Other Cardiovascular Disease Risk Scores on Atherosclerotic Cardiovascular Disease Risk Assessment (from the Multi-Ethnic Study of Atherosclerosis [MESA]). *Am. J. Cardiol.* **118**, 691–696 (2016).
3. Yadlowsky, S. *et al.* Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann. Intern. Med.* **169**, 20–29 (2018).
4. Mora, S. *et al.* Evaluation of the Pooled Cohort Risk Equations for Cardiovascular Risk Prediction in a Multiethnic Cohort From the Women's Health Initiative. *JAMA Internal Medicine* vol. 178 1231 (2018).
5. Pfohl, S. *et al.* Creating fair models of atherosclerotic cardiovascular disease. *AIES 2019* 271–278 (2019).
6. Pfohl, S. *et al*. An Empirical Characterization of Fair Machine Learning For Clinical Risk Prediction. *arXiv* (2020).
7. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *NeurIPS* 3323–3331 (2016).
8. Arnett, D. K. *et al.* 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **74**, e177–e232 (2019).

Creating a national COVID-19 Limited Data Set: Regulatory and governance innovations within the National COVID Cohort Collaborative (N3C)

Melissa Haendel, PhD, Christine Suver, PhD, Julian Solway, MD, John Wilbanks, and Joni Rutter, PhD

Oregon Clinical and Translational Science Institute, Oregon Health & Science University, Portland, OR, USA; Sage Bionetworks, Seattle, WA; University of Chicago, Chicago, IL; National Center for Advancing Translational Sciences, Bethesda, MD

Introduction

The COVID-19 pandemic has revealed many cracks in US public health and healthcare systems. One key issue is the lack of access to a large amount of patient-level observational data, upon which machine learning and statistical analytics can be performed to discover underlying mechanisms for optimal patient care. Recent retractions of large observational studies in high impact journals[1,2] have highlighted the need for reproducible, transparent, accountable science—no small challenge for sensitive clinical data. However, recent innovations and deep collaboration have made this possible for the large scale of participants and data needed to address the pandemic. N3C (covid.cd2h.org) provides the security required to house the extensive limited data set and broad community access for collaborative analytics, an experiment in regulatory engineering and open science. N3C's goal is to demonstrate that a "multi-site collaborative learning health network can overcome barriers to rapidly build a scalable infrastructure incorporating multi-organizational clinical data for COVID-19 analytics."[3]

Methods - Data Governance

Data governance comprises (a) principles, policies and strategies to be adopted, (b) creation of functions and roles to implement these policies and strategies, and (c) architectural designs that provide both a home for the data and an operational expression of policies in the form of controls and audits ("Research Data Governance, Roles, and Infrastructure" Ch 14 of [4]). The value of information is threatened both by loss of integrity—the principal internal threat—and by its potential for theft or leakage, compromising privacy and failure to meet regulatory requirements—the external threats. As a management discipline, "data governance" delineates the principles, policies, strategies, functions, and actions that guide the establishment of a coherent governance program. As a management practice, data governance enhances and defends the value of the data in the organization, both inwards and outwards. The internal goal is to establish best practices to support and assure the integrity of the data so that it maintains its value and maximized scientific outcomes. The external goal is to protect the data from deliberate theft, accidental leakage, and inappropriate disclosure while simultaneously providing access to experts for whom the data are a necessary resource to improve societal outcomes. Data governance in "open" science must balance these goals with widespread dissemination, where closed organizations can simply use trade secrecy and contracts to protect information. N3C opted to submit to a single IRB at Johns Hopkins Medicine to minimize burden. To respect autonomy, sites were given the option to submit to their local IRB; only three sites chose to.

Results - Data Security

Biomedical research data governance adds additional protections to the general principles to ensure that patient privacy and confidentiality are not breached, that categories of data are precisely specified, that a "minimum necessary" standard is observed, that there is ethical oversight, and that all data users are appropriately authorized and authenticated. Participant consent was not required for N3C and a HIPAA waiver of consent was granted. Oversight issues for N3C had to be prominent, to assure stakeholders of adequate protection and that they were built into the design from the very start. They must also factor in

broad data access by a wide variety of users in the service of open innovation.

The urgency for a research response to the pandemic led to highly streamlined approval processes for research projects at many institutions. In the case of N3C, the value and the organization were immediately deemed sufficient. It was accepted that the design of the Enclave would provide adequate protection monitored by an independent assessor, but the question was raised as to whether the contracted commercial entities for the deployed software components had a federal "operating directive", and whether all components that necessitated a search of federal resource status (FedRAMP) commercial entities had the proper federal "authorization to operate" (ATO).

Discussion - Key Conclusions for Collaboration and Publication Ethics

Because N3C is a large community (>1000 members) with complex regulatory and security regulations, Community Guiding Principles were developed to define: partnership, inclusivity, transparency, reciprocity, accountability, and security[5].

There was a need to split regulatory repercussions from data misuse into an NIH provisioned User Code of Conduct (ncats.nih.gov/n3c/resources/data-user-code-of-conduct), that could be overseen by government regulation, versus the behavioral and collaborative norms that ensure community expectations for good community citizenship. This was followed by a similar process to define Publication and Attribution Principles[5]. There is a tension between providing transparency and efficiency of collaboration for resource reuse (e.g., code sets, mapping files, software tools, etc) and the ability to allow investigators the privacy and time to realize their ideas and not get scooped. The community has maximized transparency by public sharing of project titles and investigator names, community-available resources, and Domain Teams support for community development and inclusion of junior investigators.

Acknowledgments

The work presented in this panel reflects the collaboration of many individuals from across the N3C, CTSAs, NCATS, OHDSI, and the many organizations that provided ongoing participation. Specific thanks go to Ken Gersing, Christopher Chute, Warren Kibbe, Adam Wilcox, Tony Solomonides, Lilli Portia, Meredeth Temple-O'Connor, Anita Walden, Julie McMurry, Connor Cook, and Andrew Neumann. This work has been funded through the National Center for Advancing Translational Sciences, National Institutes of Health, under award number U24 TR002306.

References

1. Mehra MR, Ruschitzka F, Patel AN. Retraction-Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. Lancet. 2020;395: 1820.

2. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621. N Engl J Med. 2020;382: 2582.

3. Haendel M, Chute C, Gersing K, N3C Consortium. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. J Am Med Inform Assoc. 2020. doi:10.1093/jamia/ocaa196

4. Daniel C, Kalra D, Section Editors for the IMIA Yearbook Section on Clinical Research Informatics. Clinical Research Informatics. Yearb Med Inform. 2020;29: 203–207.

5. N3C Consortium. Community Guiding Principles for the National COVID Cohort Collaborative (N3C). Zenodo; 2020. doi:10.5281/ZENODO.397872

FHIR Used to Expedite sIRB Workflow

W. Ed Hammond, PhD; James Topping, MS; Kirubel Asfaw, MS; Diane Rodden, MSIS; Vivian L. West, PhD; Lawrence Muhlbaeir, PhD; Eric Eisenstein, DBA
(Duke University, Durham, NC)

Introduction

Increasingly Randomized Controlled Trials are being conducted at multiple sites. The revised Common Rule now requires all federally funded, multisite studies to use a single Internal Review Board (sIRB) model. This requirement went into effect on January 20, 2020. [1] Prior to this mandate, institutions independently created their own forms and processes to meet IRB requirements. The revised Common Rule is specific in what is required for the IRB process.

We are conducting research to create a set of standardized forms, and distribute and manage multiple documents used in sIRB review and approval, using HL7's Structural Data Capture (SDC) to manage the forms. This project will decrease manual efforts and redundancy in capturing document content by using HL7 International ® FHIR® (Fast healthcare Interoperable Resources) to automate the processes. Figure 1 illustrates the problem we address.

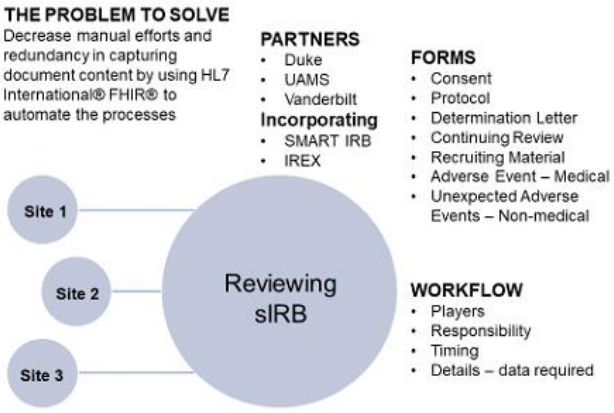


Figure 1. The Problem to Solve

Methods

First, we obtained documents currently used in the IRB process from 58 CTSA sites. Currently there is no consistency in the organization, wording, or content of the various documents required in the sIRB process for multi-site research grants. In addition, the same data have to be entered redundantly in different but related forms. There is no national source for assigning the required identification numbers.

We used the FHIR Resources Questionnaire and the FHIR Questionnaire Resource Response Resource[2] to create a template for each form required for the sIRB review. We then used HL7's SDC[3] to manage the various forms, including auto-population and manual population of the forms. The forms can then be distributed to required sites and participants. Figure 2 illustrates the process for creating templates for each form. The forms created were consent, protocol, determination letter, recruiting material, continuing review, and adverse events – both medical and non-medical. The templates will be globally accessible to all researchers.

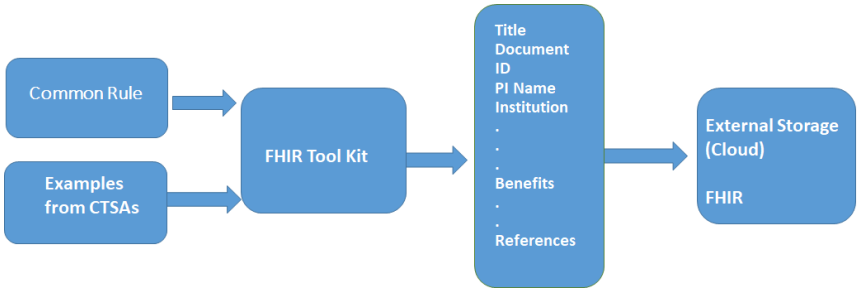


Figure 2. Process for creating template.

The templates are managed using Structured Data Capture, which is a result of an ONC funded project that was developed in both HL7 and IHE. A researcher selects the form to be populated. If it is the first form, a universal identifier (UID) is assigned and the PI fills in an initial set of data such as the title of the research, the investigators, the sites for the research, and any other required data. Those data are stored in a local database to auto-populate other forms that use the same data. Data about the sites are extracted from a national database that includes the names of the sites, an identifier, and any additional administrative data including address. The UIDs are hierarchical, so sub-sites within an institution may be identified. Figure 3 shows the flow for this process.

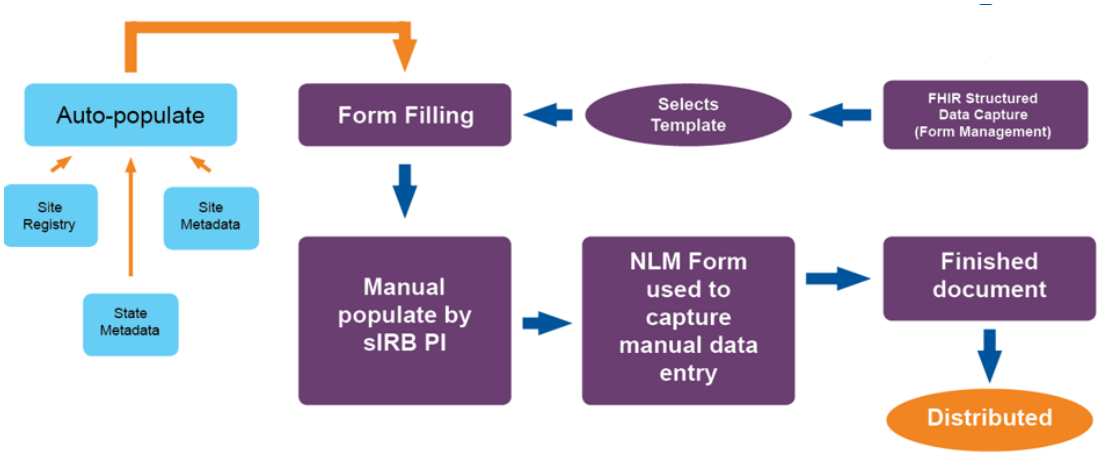


Figure 3. Process for completing a form for sIRB

All of the standardized fill-in options for each question are included within the form template as well as text boxes to add non-standard language from the template. These will be looked at to further standardize the form based on necessity and commonality. State and site requirements are identified in the site database and are incorporated into the form.

Results

Templates for all the forms were developed and demonstrated at an HL7 connectathon. We also developed a set of data elements and a glossary for each form, as well as a common set across all forms. This approach permitted us to create a common set of data elements that appeared in multiple forms.

Discussion

We are in the process of creating a FHIR Implementation Guide and plan to ballot it in HL7. A next step is to get multiple groups to use the forms and provide feedback. Clearly the use of the forms will reduce the effort of creating and using the forms. The standardization of forms across multiple sites will contribute to understanding through well-defined content. The question remains "Will sites be willing to switch from current systems and methods and implement the standards?" Further, through the use of FHIR, the data contained within the forms can be used for secondary purposes. The challenge now is to define what and how we will interface or integrate with the IRB vendors.

References

1. Government Publishing Office. Electronic Code of Federal Regulations, January 17, 2020. https://www.ecfr.gov/cgi-bin/retrieveECFR?gp=&SID=83cd09e1c0f5c6937cd9d7513160fc3f&pitd=20180719&n=pt45.1.46&r=PART&ty=HTML.
2. HL7.org. FHIR Release 3 (STU; v3.0.2-11200): Oct 24, 2019. http://hl7.org/fhir/stu3/questionnaire-operations.html.

This work is supported by CTSA Grant number: UL1TR002553

ctGATE: A Clinical Trial Generalizability Assessment Toolbox

Zhe He, PhD[1], Arslan Erdenasileng, MS[1], Xiang Tang, MS[1], Yiqi Xu, MS[1], Qian Li, MS[2], Neil Charness, PhD[1], William Hogan, PhD[2], Thomas J. George, MD[2], Yi Guo, PhD[2], Jiang Bian, PhD[2]

[1]Florida State University, Tallahassee, FL, USA; [2]University of Florida, Gainesville, FL, USA

Abstract

Rigorously designed clinical trials often tend to overemphasize internal validity and subsequently diminish the generalizability of their results to the real-world population. This tendency may partly be due to inadequate resources for generalizability assessment. We performed a systematic review of the literature and identified 187 studies relevant to generalizability assessment. We further developed a web-based toolbox called ctGATE to enable quick search of related papers, implementation codes, and associated tutorials for generalizability assessment methods.

Introduction

Clinical studies are often conducted under idealized and rigorously controlled conditions to ensure their internal validity, though at the expense of compromising their external validity or generalizability. These idealized conditions sometimes result in overly restrictive eligibility criteria. Subsequently, certain population subgroups are often excluded through questionable criteria and are subsequently underrepresented. For example, older adults have been especially underrepresented in cancer studies. The underrepresentation of these population subgroups may lead to overestimates of the treatment effects and increase the likelihood of adverse outcomes in diverse populations when the interventions are moved into real-world clinical practice. It is imperative to rigorously assess the generalizability of a clinical study, so that stakeholders including pharmaceutical companies, policymakers, providers, and patients would be able to understand and anticipate the possible effects of the interventions in the real world. In the past two decades, many studies have assessed generalizability, but mostly were after the fact, ad hoc, not systematic, and focused on specific diseases and sets of trials without a formalized approach. So far, there has been a significant knowledge gap between the available methods for generalizability assessment and their adoption in research practice. Most generalizability assessments have been conducted as an ad hoc auditing effort by a third party after the fact. We believe the key barriers to generalizability assessment are two-fold: (1) the lack of evidence to demonstrate their validity, which also leads to the lack of consensus on the best practice for generalizability assessments; and (2) the lack of readily available, well-vetted statistical and informatics tools. Motivated to fill this gap, we systematically reviewed the extant methods for generalizability assessments [1]. We also developed a web-based Clinical Trial Generalizability Assessment Toolbox (ctGATE) with its accompanying documentations and tutorials.

Methods

We performed a systematic literature search over the following 4 databases: MEDLINE, Cochrane, PychINFO, and CINAHL. We conducted the scoping review in the following six steps: 1) gaining an initial understanding about clinical trial generalizability assessment, population representativeness, internal validity, and external validity, 2) identifying relevant keywords, 3) formulating four search queries to identify relevant articles in the 4 databases, 4) screening the articles by reviewing titles and abstracts, 5) reviewing articles' full-text to further filter out irrelevant ones based on inclusion and exclusion criteria, and 6) coding the articles for data extraction.

Study Selection: We used an iterative process to identify and refine the search keywords and search strategies. Following the Institute of Medicine's standards for systematic review and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), we identified 5,352 articles as of February 2019 from MEDLINE, CINAHL, PychINFO, and Cochrane using relevant keywords. After removing duplicates, 3,569 records were assessed for relevancy by two researchers (ZH and XT) through reviewing the titles and abstracts against the inclusion and exclusion criteria. Conflicts were resolved with a third reviewer (JB). During the screening process, we also iteratively refined the inclusion and exclusion criteria. Out of the 3,569 articles, 3,275 were excluded through the title and abstract screening process. Subsequently, we reviewed the full texts of 294 articles, among which 106 articles were further excluded based on the exclusion criteria. The inter-rater reliability of the full-text review between the two annotators is 0.901 (i.e., Cohen's kappa, $p < .001$). 187 articles were included in the study.

Data Extraction: We coded and extracted data from the 187 eligible articles according to the following aspects: (1) whether the study performed an *a priori* generalizability assessment or *a posteriori* generalizability assessment or

both; (2) the compared populations and the conclusions of the assessment; (3) the outputs of the results (e.g., generalizability scores, descriptive comparison); (4) whether the study focused on a specific disease. If so, we extracted the disease and disease category; (5) whether the study focused on a particular population subgroup (e.g., elderly). If so, we extracted the specific population subgroup; (6) the type(s) of the real-world patient data used to profile the target population (i.e., trial data, hospital data, regional data, national data, and international data). Note that trial data can also be regional, national, or even international, depending on the scale of the trial. Regardless, we considered them in the category of "trial data" as the study population of a trial is typically small compared to observational cohorts or real-world data. For observational cohorts or real-world data (e.g., electronic health records [EHRs]), we extracted the specific scale of the database (i.e., regional, national, and international). For the studies that compared the characteristics of different populations to indicate generalizability issues, we further coded the populations that were compared (e.g., enrolled patients, eligible patients, general population, ineligible patients), and the types of characteristics that were compared (i.e., demographic information, clinical attributes and comorbidities, treatment outcomes, and adverse events). We also identified the statistical/informatics methods used for generalizability assessment.

ctGATE: To allow researchers to easily find appropriate generalizability assessment methods and tools for their studies given the available data, we developed the prototype of ctGATE on WordPress with the OceanWP theme and the TablePress plugin. The TablePress plugin allowed us to display, sort, filter and search the included publications.

Results

Figure 1 shows the interface of ctGATE. The work-in-progress tool can be accessed at https://ctgate.cci.fsu.edu/. With ctGATE, the user can search clinical trial generalizability assessment papers using the data source, disease category, types of generalizability assessment methods (score/non-score output, a priori / a posteriori generalizability assessment), PMID, and title. Then the filtered papers will be displayed in a table. The user can (1) click on the PMID to view the entry of the article in PubMed; (2) click on the title of the paper to view all the coded information about the study; and (3) view the R/Python tutorials for the generalizability assessment. Currently, it has four tutorials: (1) GIST 2.0 (in Python) [2], (2) Traditionally statistical methods (in R); (3) Standardized mean difference of propensity scores (in R), and (4) Propensity score weighting/matching (in R). The tutorials were developed in Jupyter Noteboook or R Markdown, allowing the users to interact with the codes. The users can also directly access these tutorials in the "Tutorials" tab. In the next phase, we will conduct user evaluations with end users to assess its usability and usefulness.

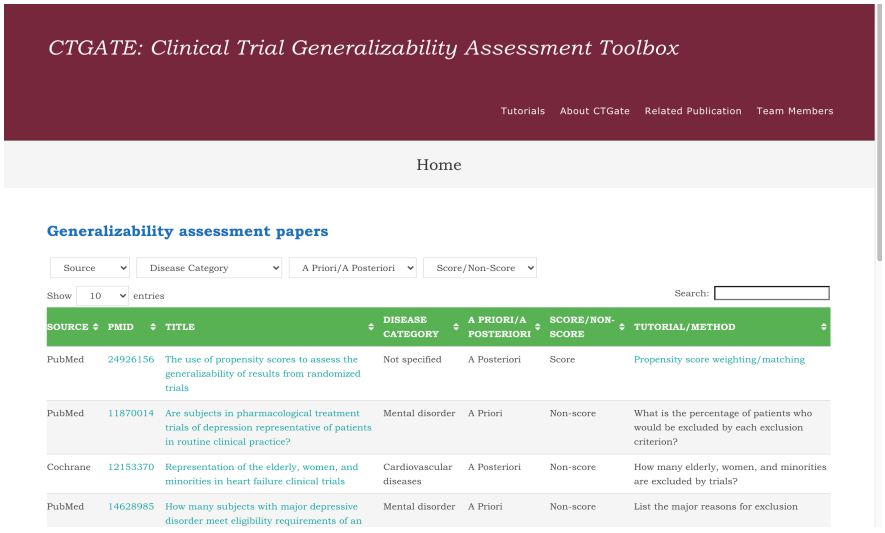


Figure 1. ctGATE interface

Acknowledgments

This study was supported by the National Institute on Aging of the National Institutes of Health (NIH) under Award Number R21AG061431; and in part by NIH Award UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] He Z, Tang X, Yang X, Guo Y, George TJ, Charness N, Quan Hem KB, Hogan W, Bian J. Clinical Trial Generalizability Assessment in the Big Data Era: A Review. Clin Transl Sci. 2020;13(4):675-84.

[2] Sen A, Chakrabarti S, Goldstein A, Wang S, Ryan PB, Weng C. GIST 2.0: A scalable multi-trait metric for quantifying population representativeness of individual clinical studies. J Biomed Inform. 2016;63:325-36.

Privacy-Preserving Transitive Record Linkage

Andrew Hill, BS[1], Michael G. Kahn, MD, PhD[2], Shaun Grannis, MD[3,4], Chan Voong, MUSA[2], Lisa Schilling, MD, MSPH[2], Toan C. Ong, PhD[2]
University of Colorado Denver[1], University of Colorado Anschutz Medical Campus[2], Indiana University[3], Regenstrief Institute[4]

Introduction:

Record linkage (RL) is a process that determines which records belong to the same individual across datasets.[1] Linking patient records across multiple clinical data sources can improve data quality, especially data completeness. Clear-text RL algorithms use human-readable personally identifiable identifiers (PII) to determine linkages. Privacy-Preserving Record Linkage (PPRL) is an alternative linkage method that obfuscates all PII data before performing linkage, ensuring minimal risk of sensitive information being exposed.[2] Transitive record linkage (TRL) captures linkages between multiple linked record pairs that may be missed using traditional pairwise linkage methods.[3] For example, pairwise linkage may link Record A to Record B and Record B to Record C. TRL adds a link between Record A and Record C. In this work, we developed and implemented a novel scalable method called Privacy-Preserving Transitive Record Linkage (PPTRL) to generate both pairwise and transitive linkages while maintaining the privacy and security of patient data.

Methods:

PPTRL has two steps: 1) performing traditional deterministic or probabilistic PPRL to generate pairwise linkages (**Table 1a and 2**) using an undirected graph to create transitive linkages. The strength of each pairwise linkage is measured by confidence score values from a probabilistic linkage method. Probabilistic linkage methods generate confidence scores based on the textual similarity of one or more data fields within each pair of records. Pairwise linkages are used as input to determine records that are linked together in a record cluster. Members of a record cluster are assigned with a shared identifier called Network ID (**Table 1b**). A Network ID is an identifier used to refer to a single entity or individual. For example, consider the direct linkages $\{A, C\}$ and $\{B, C\}$ from Rows 1 and 2 in **Table 1a**. Records A and B both link to C, but do not link to each other, since there is no direct $\{A, B\}$ linkage in the table. However, records A and B are indirectly linked, since both connect to record C. A and B are transitively linked and records A, B and C are in the same record cluster. **Table 1b** represents the additional transitive linkage between A and B by assigning both records the same Network ID = 4001.

Table 1(a) pair-wise linkage; (b) transitive linkage.

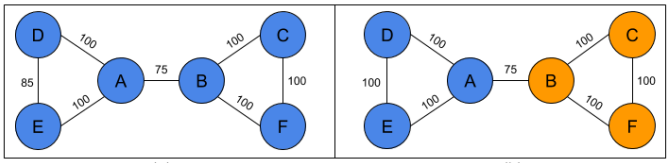
#	ID 1	ID 2	Confidence
1	A	C	100
2	B	C	85
3	D	E	100

(a)

Record ID	Network ID
A	4001
B	4001
C	4001
D	4002
E	4002

(b)

Figure 1: Transitive linkage (a) before and (after) linkage refinement.



Transitive Linkage Computation: The pairwise linkage result (**Table 1a**) is formulated as an undirected graph G. Each record r_i is represented as a vertex in G, and each pairwise linkage between records r_1 and r_2 is represented as an undirected edge $E = \{r_1, r_2\}$. Finding the Network IDs for each record is to find the maximally connected subgraphs or connected components (CC) of G. A CC of G is a subgraph, such that every vertex in the CC is reachable by traversing edges from any other vertex. After partitioning graph G into a set of CCs, a unique Network ID is generated for each CC and assigned the same Network ID for all the nodes in this subgraph.

Transitive Linkage Refinement: Transitive Linkage identifies unique entities from a pairwise RL result. However, naively assuming that all vertices in a connected component should be linked together may inadvertently generate false-positive linkages resulting in records that belong to different individuals being assigned a shared Network ID. In this study, we focus primarily on the reduction of these false-positives in the transitive linkage result. To illustrate a false-positive error, consider a linkage scenario with two true entities $G1 = \{A, D, E\}$ and $G2 = \{B, C, F\}$. **Figure 1a** is a graphical representation of the linkage graph G described above. The number above each edge represents the pairwise linkage confidence, a value between 0 (no match) - 100 (identical match) generated from a probabilistic linkage method. We can observe that entities $G1$ and $G2$ are both well connected within their vertex groups but that both entities are connected by a linkage $\{A, B\}$ having a weak confidence score = 75. Because we know that $G1$ and $G2$ are separate entities, the linkage $\{A, B\}$ represents a false positive linkage. The Transitive Linkage algorithm will assign all records the same Network ID, since all records reside in the same CC. **Figure 1b** presents the correct output where the graph has been partitioned into two groups, represented by different colors. Each group

will receive a separate Network ID. To avoid problematic scenarios like **Figure 1a**, we apply a greedy correlation clustering algorithm, which was first proposed in [3]. This greedy algorithm has been shown to perform better in both accuracy and runtime than other methods.[3] This algorithm modifies the Network IDs in each connected component of the linkage graph to reduce the effect of False Positive linkages. The algorithm begins with the Network IDs assigned by the original Transitive Linkage result as the initial set of cluster labels, which the algorithm then refines. The refinement algorithm first assigns each CC an initial penalty score based on the similarities of disconnected vertices and the dissimilarities of connected vertices. Then, the algorithm will attempt to find a cluster label assignment with a lower penalty score by changing the cluster labels of each vertex or by creating new cluster labels. If the algorithm discovers a penalty lower than the initial score, it will move to this new cluster label configuration and continue optimizing until no changes can be made to decrease the penalty score. Notably, each CC optimization is an independent operation, since each CC is disconnected from the rest of the graph. Thus, after identifying the CCs within the output graph, the refinement process can be executed in parallel.

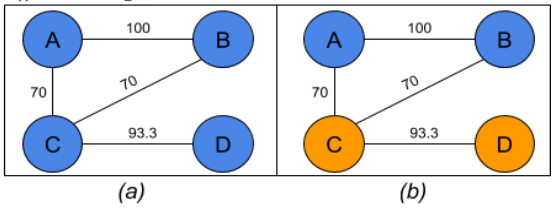
Results:

To test our transitive linkage method, we created a synthetic dataset with 115,000 records. We applied multiple corruption methods to simulate real-world data quality issues (missingness, typos, character transpositions, etc.). The data were linked using both probabilistic and deterministic PPRL methods. Network IDs were recorded for each record before and after the transitive linkage refinement process. Before refinement, the transitive linkage result on the dataset produced 70 false positive linkages. The refinement algorithm was able to remove 12 false positive linkages, while only removing 2 true positive linkages. This result shows that the refinement algorithm is capable of reducing the number of false positive linkages, while not breaking a large number of existing true positive linkages.The pairs of linkages which shared a Network ID prior to refinement but were assigned distinct Network IDs post-refinement (i.e., records which changed Network ID because of the refinement step) were recorded for analysis. **Table 2** highlights one example with four records that were initially contained in the same CC. In the simulated data ground-truth labels (the PID column in **Table 2**), Records *A* and *B* are known to belong to the same entity (PID=1), while the remaining 2 records belong to a separate, unrelated entity (PID=2). Before refinement, all four records are linked transitively because they reside in the same connected component, as illustrated in **Figure 2a**. Records A and B are linked with a confidence of 100, indicating a perfect match. Additionally, records C and D are linked with a confidence score of 93.3, indicating a high (though not exact) similarity between the records. The links (A, C) and (B, C) are false positive linkages, generated because of the similarities between the values for First Name, Last Name, and DOB. These linkages have low confidence scores, indicating that the similarity between the records is weaker. Following the refinement step from **Figure 2b**, the algorithm splits records C and D into a separate Network ID (indicated by the color of the graph nodes). This cluster assignment was chosen because records C and D are relatively dissimilar from the records A and B, yet are closely related to each other by the edge $\{C, D\}$. After refinement, the algorithm has removed the false positive linkages, correctly assigning records A and B to one cluster, and records C and D to a separate cluster.

Table 2: Raw data using simulated corrupted records with known pairs.

ID	PID	First Name	Last Name	DOB	Address	City
A	1	Roseanna	Bernucci	2017/12/06	50 Judy Way	Knoxville
B	1	Roseanna	Bernucci	2017/12/06	50 Judy Way	Knoxville
C	2	Rosana	Bernucci	2017/06/16	3672 Gateway Drive	Montgomery
D	2	Rosana	Bernucci	2017/11/16	3672 Gateway Drive	Montgomery

Figure 2: (a) initial and (b) refined CC. Colors represent different assigned Network IDs



Conclusion:

PPTRL adds additional true positive linkages to the traditional linkage results by including both direct and indirect linkages. After the initial graph partitioning, the PPTRL result is refined by a parallel greedy clustering algorithm to reduce the occurrence of false positive linkages.

References

1 Herzog, Thomas & Scheuren, Fritz & Winkler, William. (2007). Data Quality and Record Linkage. 10.1007/0-387-69505-2.

2. Verykios, Vassilios & Christen, Peter. (2013). Privacy-preserving record linkage. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 3. 10.1002/widm.1101.

3. Anja Gruenheid, Xin Luna Dong, and Divesh Srivastava. 2014. Incremental record linkage. Proc. VLDB Endow. 7, 9 (May 2014), 697–708. DOI:https://doi.org/10.14778/2732939.2732943

System for High Intensity EvaLuation During Radiation Therapy (SHIELD-RT): A prospective randomized study of machine learning-directed clinical evaluations during outpatient cancer radiation and chemoradiation

Julian C. Hong, MD, MS,[1,2] Neville C.W. Eclov, PhD,[2] Nicole Dalal, BA,[2] Samantha M. Thomas, MS,[3,4] Sarah J. Stephens, MD,[2] Mary Malicki, ACNP,[2] Stacey Shields, ANP-BC,[2] Alyssa Cobb, RN,[2] Yvonne M. Mowery, MD, PhD,[2] Donna Niedzwiecki, PhD,[3,4] Jessica D. Tenenbaum, PhD,[3] Manisha Palta, MD[2]

[1]Department of Radiation Oncology and Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA; [2]Department of Radiation Oncology, [3]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC; [4]Duke Cancer Institute – Biostatistics, Duke University, Durham, NC

Introduction

Patients undergoing outpatient radiotherapy (RT) or chemoradiation (CRT) frequently require acute care (emergency department evaluation or hospitalization). This can impact treatment outcomes, patient quality of life and preferences, and costs to patients and the healthcare system, making it a priority to the Centers for Medicare and Medicaid Services. Machine learning (ML) may guide interventions to reduce this risk. There are limited prospective, randomized studies investigating the clinical benefit of ML in healthcare, which has represented an obstacle to their implementation. The objective of this study was to determine whether ML identification of high risk patients driving mandatory twice-weekly clinical evaluation could reduce acute care visits during treatment.

Methods

During this single institution randomized quality improvement study (NCT04277650)[1], 963 outpatient adult courses of RT and CRT started from January 7 to June 30, 2019 were evaluated by a gradient boosted tree model. This model had previously been trained and validated based on electronic health record and cancer treatment plan data for the prediction of acute care during an RT or CRT course.[2] Top predictive factors in this model were broad, including treatment parameters (RT dose, schedule, and modality; systemic therapy), encounter history (ED and admission history), vitals (weight and pain), age, and labs (albumin).

ML identified 361 courses as high risk (>10% risk of acute care during treatment). Among these, 314 were randomized to standard once-weekly clinical evaluation or mandatory twice-weekly evaluation. The remaining courses did not undergo randomization due to physician opt out (N=31) or treatment completed within one week (N=16). Both arms allowed additional evaluations based on clinician discretion. Of the non-randomized courses, 602 were low risk and included as a comparator group. The primary endpoint was the rate of acute care visits during RT. Model performance was evaluated using receiver operating characteristic area under the curve (AUC) and decile calibration plots. For courses on the intervention arm, clinician predictions for acute care probability were collected at the first mandatory supplemental evaluation. These predictions were unblinded, and clinicians had knowledge that the patient had been assigned to intervention (ML risk >10%). ML risk was blinded for high and low risk courses without intervention.

Results

Among randomized courses, 311 were eligible and allocated to protocol treatment with once-weekly (N=157) or mandatory twice-weekly evaluation (N=154). Twice-weekly evaluation reduced rates of acute care during treatment from 22.3% to 12.3% (difference -10.0%, 95% CI -18.3 to -1.6; relative risk 0.556, 95% CI 0.332-0.924; p = 0.02). For comparison, low risk patients had a 2.7% acute care rate. Model discrimination was good in high and low risk patients undergoing standard once-weekly evaluation (AUC 0.851). Clinicians completed predictions for 145 of 154 intervention courses (94%). Unblinded clinician predictions had a narrow distribution centered around a median of 10% (IQR 5%-15%) with wide confidence intervals at most deciles (AUC 0.704).

Discussion

In this prospective randomized study, ML accurately triaged patients undergoing RT and CRT, directing clinical management with reduced acute care rates versus standard of care. This prospective study demonstrates the potential benefit of ML in healthcare and offers opportunities to enhance care quality and reduce healthcare costs.

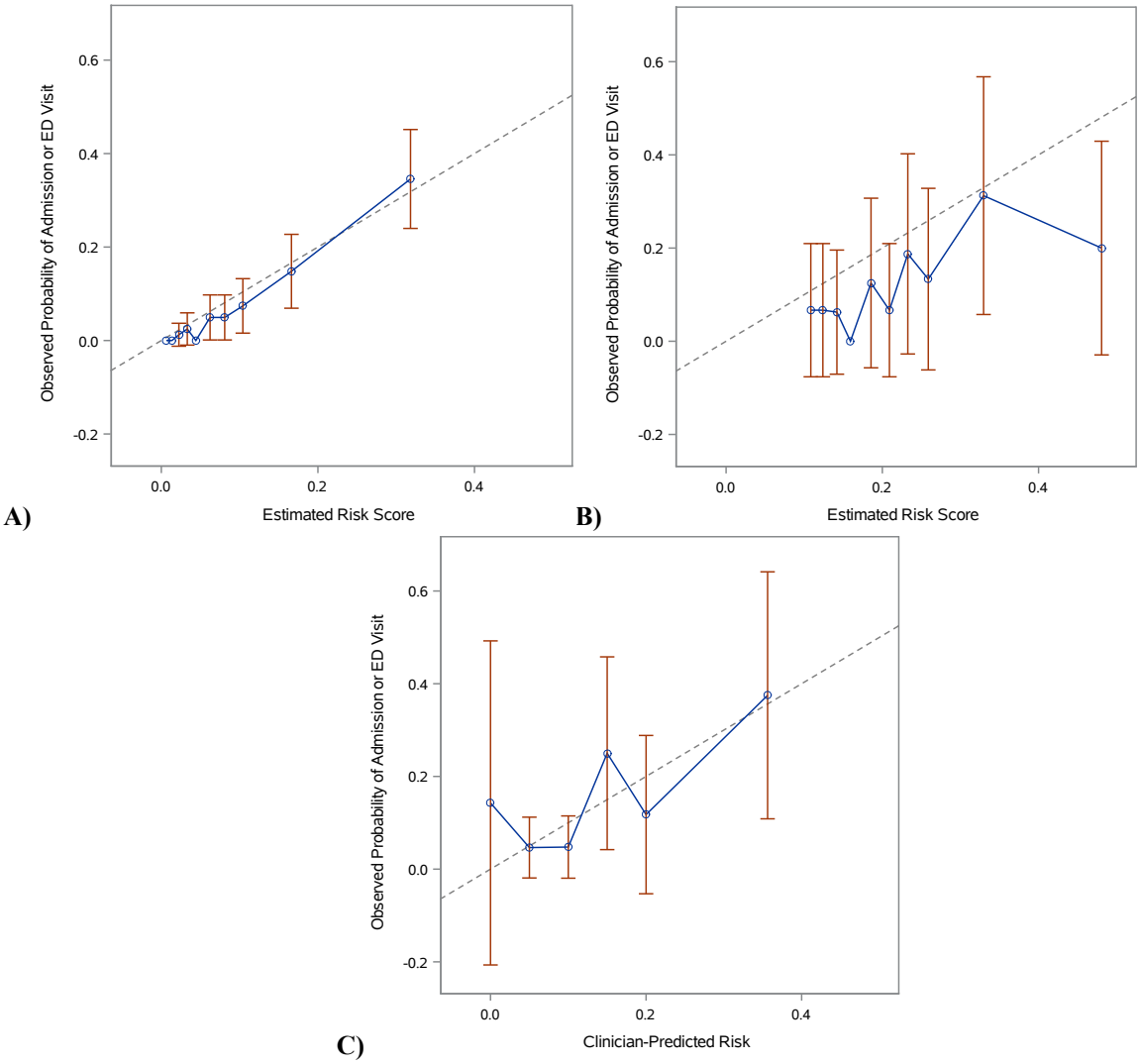


Figure 1. Calibration plot of the machine learning model prediction versus the actual event rate split into deciles demonstrates good calibration for all non-interventional courses (A; control and non-randomized courses; n = 809). For patients undergoing intervention, calibration demonstrates slight underestimation consistent with the interventional decrease in acute care (B; n = 154). Unblinded clinician predictions were centered around 10%, the known cut-off for randomization (C; n = 145). Error bars represent 95% confidence interval.

References

1. Hong JC, Eclov NCW, Dalal NH, et al. System for High-Intensity Evaluation During Radiation Therapy (SHIELD-RT): A Prospective Randomized Study of Machine Learning–Directed Clinical Evaluations During Radiation and Chemoradiation. *J Clin Oncol*. Published online September 4, 2020:JCO.20.01688. doi:10.1200/JCO.20.01688

2. Hong JC, Niedzwiecki D, Palta M, Tenenbaum JD. Predicting Emergency Visits and Hospital Admissions During Radiation and Chemoradiation: An Internally Validated Pretreatment Machine Learning Algorithm. *JCO Clin Cancer Inform*. 2018;2(2):1-11. doi:10.1200/CCI.18.00037

Acknowledgments

This study was funded in part by the Duke Endowment, which had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Improving social determinants of health screening: Examination of patients with limited English proficiency

Eun Ji Kim MD MS MS[1,2], Austin Fischer BS[2], Khadeja Kausar MHA[1], Simita Mishra PhD MHA[1], Shaun Allicock MS[1], Sabina Zak BS[1], Michael I. Oppenheim MD[1,2], Joseph Conigliaro MD MPH[1,2]

[1]Northwell Health, Manhasset, NY; [2]Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY

Introduction

Social determinants of health (SDH), the conditions into which people are born, grow, live, work, and age, have been consistently shown to influence health outcomes. Recent efforts have focused on systematically documenting SDH, particularly unmet social needs, as these factors have been shown to play a larger role in the health of individuals than either their insurance status or access to care.[1-3] Northwell Health implemented the SDH screening and referral program in June 2019 to capture population level data and offer tailored interventions for vulnerable patients. Positive responses to the program's screening questions automatically generate case manager and social work referrals.

To evaluate the SDH screening program and better understand how SDH screening can be improved, a committee with stakeholders from various departments at Northwell (Community Health, Medical Informatics, and Research) was created. Here, we compared the presence of social needs by limited English proficiency (LEP) status, which is associated with lower healthcare utilization and unfavorable health outcomes.[4,5] In particular, we sought to evaluate the presence of social needs in Spanish speakers.

Methods

We obtained Northwell Health SDH screening data from June 25th, 2019 to February 29th, 2020. Our sample included all adults (aged 18 years or older) that responded to the screening. We excluded participants with missing demographic information and preferred language. LEP status was designated based on patients' self-reported preferred language: patients who preferred non-English as their primary language were categorized to have LEP. The SDH screening captures the presence material need, employment, medical-legal assistance, health insurance, public benefit, health literacy, public transportation, medical care, utility bill, poor housing quality, food insecurity and housing insecurity. We performed descriptive analysis of all participants, and then sub-analyses of the participants based upon English proficiency status. We calculated t-tests for continuous variables and chi-squared tests for categorical variables.

Results

We found that there was a significant difference between English proficient individuals and those with LEP status across all sociodemographic domains examined, including age, gender, race/ethnicity, and health insurance ($p<0.001$) (Table 1). There were also differences in sociodemographic characteristics by Spanish speaking LEP versus non-Spanish speaking LEP patients.

Table 1. Patient characteristics by English proficiency, n (%) for categorical variables and mean (standard deviation) for continuous variables

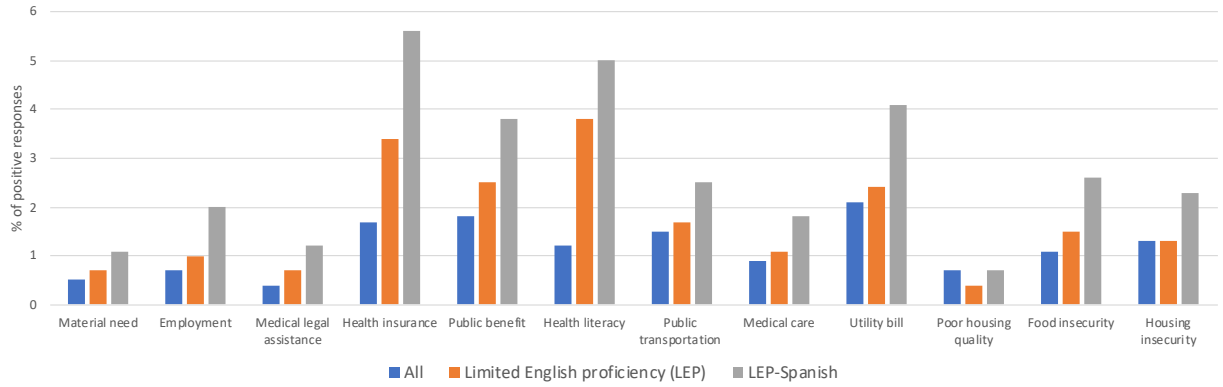
	All (n=92,958)	English proficient (n=83,445)	Limited English proficiency (n=9,513)*	Spanish (n=4,425)*
Age, years	65.2 (18.3)	64.8 (18.3)	69.0 (17.2)	63.8 (18.9)
Gender				
Female	49211 (52.9%)	43767 (52.5%)	5444 (57.2%)	2544 (57.5%)
Race/ethnicity				
White, non-Hispanic	50692 (54.5%)	48917 (58.6%)	1775 (18.7%)	122 (2.8%)
Black, non-Hispanic	15130 (16.3%)	14753 (17.7%)	377 (4.0%)	14 (0.3%)
Hispanic	11156 (12.0%)	7274 (8.7%)	3882 (40.8%)	3838 (86.7%)
Asian	6327 (6.8%)	4428 (5.3%)	1899 (20.0%)	6 (0.1%)
Other/Multiracial	6971 (7.5%)	5760 (6.9%)	1211 (12.7%)	357 (8.1%)
Unknown/Decline	2682 (2.9%)	2313 (2.8%)	369 (3.9%)	88 (2.0%)
Health insurance				

Commercial	29065 (31.3%)	27533 (33.0%)	1532 (16.1%)	931 (21.0%)
Medicare	36176 (38.9%)	32194 (38.6%)	3982 (41.9%)	1649 (37.3%)
Medicaid	11215 (12.1%)	8665 (10.4%)	2550 (26.8%)	1334 (30.2%)
Self-pay	735 (0.8%)	588 (0.7%)	147 (1.6%)	112 (2.5%)
Unknown	15767 (17.0%)	14465 (17.3%)	1302 (13.7%)	399 (9.0%)

* P-values comparing 1) patients with English proficiency to limited English proficiency and 2) Spanish speaking versus non-Spanish speaking versus English proficiency, were all <0.001.

A higher percentage of patients with LEP reported social needs compared to patients with English proficiency ($p \leq 0.05$), excluding housing insecurity ($p=0.87$), public transportation ($p=0.16$) (Figure 1). Furthermore, patients with Spanish speaking LEP had even higher percentages of social needs compared to patients with English proficient.

Figure 1. Percentage of patients with positive responses to social need screening questions



Conclusion

Our evaluation program identified that LEP status, specifically Spanish speaking LEP speakers, was associated with the increased presence of social needs. The study is limited by individual social needs being documented as yes/no responses; "no" to each social need question can be due to a lack of social need or a missing response. We also identified LEP based on their preferred language, which can result in identifying patients who are fluent in English and non-English language to be labeled as LEP patients. This evaluation paves the way for an intervention aimed at addressing social needs of LEP patients. Patients with LEP were found to utilize resources to address social needs,[6] therefore, it will be important to expand the screening of LEP patients to better identify patients with social needs and subsequently guide referrals to appropriate services.

Acknowledgement: We thank Yulia Kogan, and Yuval Romm for their support.

Reference

1. Cottrell EK, Gold R, Likumahuwa S, et al. Using Health Information Technology to Bring Social Determinants of Health into Primary Care: A Conceptual Framework to Guide Research. *J Health Care Poor Underserved.* 2018;29(3):949-963.
2. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. *Public Health Rep.* 2014;129 Suppl 2:19-31.
3. Buitron de la Vega P, Losi S, Sprague Martinez L, et al. Implementing an EHR-based Screening and Referral System to Address Social Determinants of Health in Primary Care. *Med Care.* 2019;57 Suppl 6 Suppl 2:S133-S139.
4. Sentell T, Braun KL. Low health literacy, limited English proficiency, and health status in Asians, Latinos, and other racial/ethnic groups in California. *J Health Commun.* 2012;17 Suppl 3:82-99.
5. Karliner LS, Jacobs EA, Chen AH, Mutha S. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. *Health Serv Res.* 2007;42(2):727-754.
6. Uwemedimo OT, May H. Disparities in Utilization of Social Determinants of Health Referrals Among Children in Immigrant Families. *Front Pediatr.* 2018;6:207.

A Landscape Survey of Planned SMART/HL7 Bulk FHIR Data Access API Implementations and Tools

James Jones,[1] Daniel Gottlieb,[1,2] Joshua C. Mandel,[1,3]
Vladimir Ignatov,[1] Alyssa Ellis,[1] Kenneth D. Mandl[1,2,3]
[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA;
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA;
[3]Department of Pediatrics, Harvard Medical School, Boston, MA

90AX0022/01-00 & 90AX0019/01-01, Office of the National Coordinator of Health Information Technology

Introduction

Recently published by the Office of the National Coordinator of Health Information Technology (ONC), the 21st Century Cures Act "Final Rule" regulates application programming interface (API) requirements for certified health information technology and defines protections against "information blocking." The rule covers two open APIs and authorization frameworks, standardized by Health Level Seven International (HL7) and designed to enable a robust app ecosystem. One, SMART on FHIR, provides patient- and provider-facing data access for individual patients and small cohorts. The second, the SMART/HL7 FHIR Bulk Data Export, enables system-level access to data for larger cohorts and patient populations. Compliance with the standardized API functionality in the rule is required by 2022. Without a standardized API, Bulk data operations normally require expensive and customized extract-transform-load pipelines, with myriad business needs requiring more efficient, standardized access where possible. Only advanced healthcare organizations can participate in these activities. Feedback from initial adopters will inform advancement of the standard and effective usage in 2022.

We surveyed 22 developer teams across major healthcare and IT sectors to ascertain information about their progress and solicit feedback on deployment of FHIR servers and tools implementing and advancing the SMART/HL7 FHIR Bulk Data Access Implementation Guide. We requested information on standardized features likely to be supported, and on timelines for any implementations planned ahead of the regulated timeline.

Methods

We approached key stakeholder organizations across the health enterprise for comments and timelines for adoption of the FHIR Bulk Data API as described in its implementation guide (https://hl7.org/fhir/uv/bulkdata/index.html). Thirty-seven representatives from organizations with known plans to build to the standard were contacted and asked to complete a survey online or over the phone.

Results

We received survey responses from 22 of the 37 organizations; seven were developing both FHIR bulk data servers and clients based on the API, 12 were developing servers alone, and three were developing clients alone. Relevant use cases were probed, alongside comments, timelines, and intentions to implement individual elements of the specification. Respondents were classified as payors (5), research institutions (3), EHR vendors (3), FHIR tooling developers (4), cloud vendors (5), and other purveyors of health IT platforms (2). Only EHR vendors responded that a motivation was to meet regulatory requirements. Respondents developing FHIR tooling expressed interest in advancing the standard itself. Use cases commonly being addressed with the Bulk FHIR API included value-based care, public health, and data sharing with--payors, researchers, accountable care organizations (ACOs), across health information exchanges, and across sites within an organization. Machine learning and business analytics were also of interest.

Progress toward individual elements of the bulk FHIR spec were requested, with respondents able to select *Currently supported*, *Planning to support by May 2022*, and *Not planning to support*. Notably, 15 of the 19 servers reported plans to be ready ahead of the deadline. Answers are shown in Figure 1. Features commonly expected for early rollout include the bulk data operations for exporting all patients and individual groups, along with basic filtering on resource type and last-updated timestamp. The SMART backend authorization framework and asynchronous request pattern are in scope for 15 of the 19 servers, while the optional '_typeFilter' query parameter and exporting all server contents were flagged as out of scope or too costly by seven respondents.

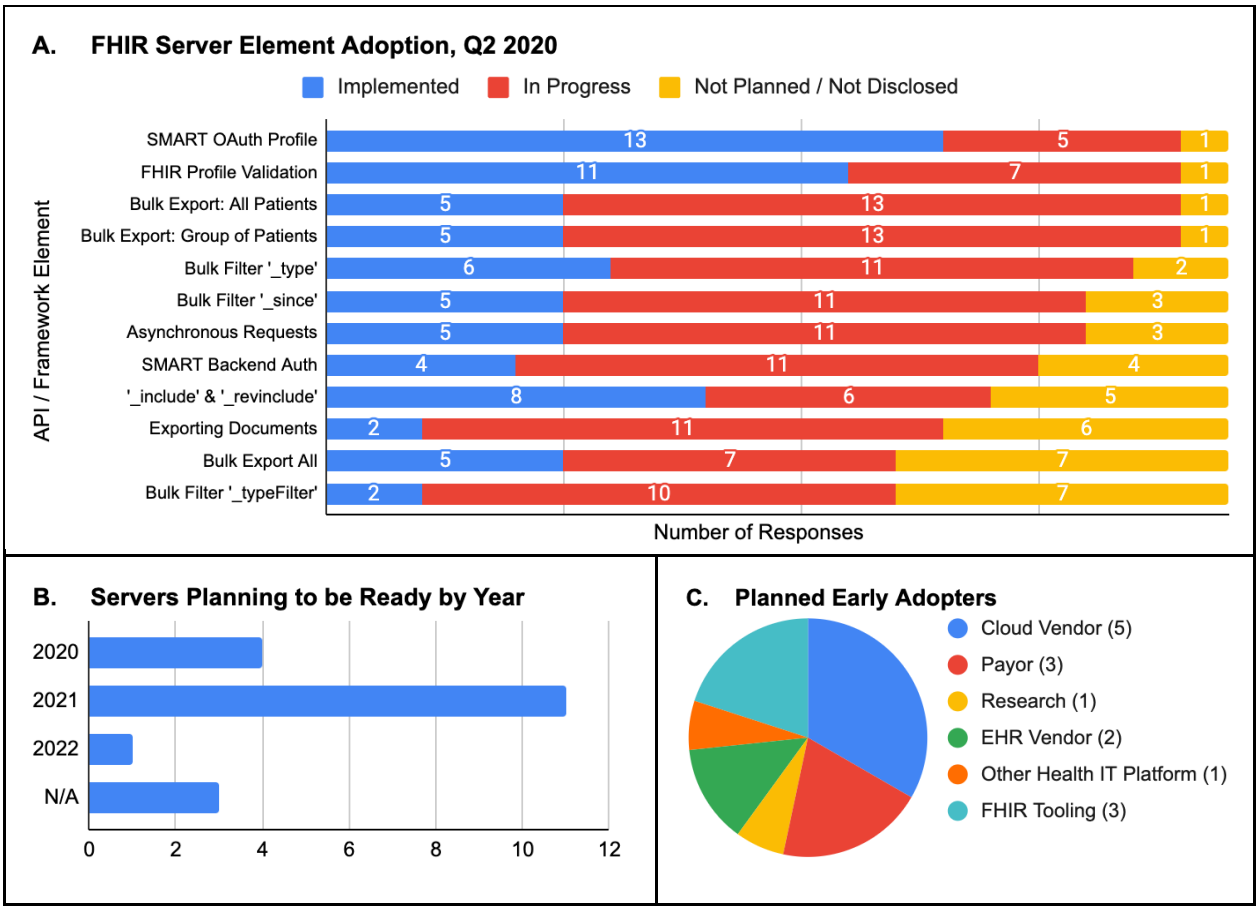


Figure 1. FHIR Bulk data early adoption landscape. **A**, Progress toward individual API and framework elements of 19 servers. **B**, Planned timelines for bulk data implementations to be ready for interoperability. "N/A" indicates no answer given. **C**, Breakdown of the 15 servers planning to have complete implementations before 2022.

Further, seven implementers are planning to support uploading data directly to cloud buckets where possible, and support for serializing data in the Parquet format is in scope in five implementations (in addition to the required NDJSON format). Notable implementation hurdles listed in free text were: i) hardware limitations and logistics for moving large datasets, including error handling, processing time, deciding where to split large files, and how best to load them; ii) managing granular access, particularly in federated systems and where the user requesting a bulk export needs to be explicitly identified for audit purposes; and iii) de-identifying data stored in documents and free text fields when leveraging the exported data for some use cases.

Discussion

The initial SMART on FHIR standard took 11 years from conception to regulation. In contrast, the SMART/HL7 bulk FHIR access project has moved much more rapidly since its inception in December 2017, being regulated under ONC's rule by 2020. Of note, this survey was conducted after March 2020, when the ONC rule was published requiring certified health IT to implement the API by May 2022. In this study we discovered that many organizations not directly affected by the regulation were planning to rollout support ahead of that deadline, which has since been extended to December 2022. We also identified implementation challenges, some of which are already under open development, such as patient group management and optimization of bulk data import. Federated access control and de-identification tools may merit more attention. Lastly, the resources and costs for moving and storing data by EHR and cloud vendors will have a large impact on the robustness of the health information economy. Monitoring these factors, encouraging participation of siloed data sources, and continuing progress toward common use of an interoperable FHIR data model will greatly aid systems and developers in leveraging the emerging bulk data landscape.

Frequent Sampling of LCLs may Affect Genotyping Quality in Whole Genome Sequencing

Jae-Yoon Jung, PhD [1], Brianna Chrisman [2], Kelley Paskov [3], Nate Stockham [4], Peter Washington [2], Maya Varma [5], Min Woo Sun [3], Sepideh Maleki [6], Kevin Tabatabaei [7], Dennis P. Wall, PhD [1,3]

Departments of [1]Pediatrics and Systems Medicine, [2]Bioengineering, [3]Biomedical Data Science, [4]Neuroscience, [5]Computer Science, Stanford University, Stanford, CA 94305, USA
[6]Department of Computer Science, University of Texas, Austin, TX 78712, USA
[7]Faculty of Health Sciences, McMaster University, Hamilton, ON L8S 4L8, Canada

Introduction

As a continuous source of genomic DNA from immortalized B cells, lymphoblastoid cell lines have been extensively used in genomic experiments and sequencing projects. LCL samples show high genotype concordance with their parental whole blood (WB) samples in their early passage[1]. However, passage information is often unavailable and there have been few studies on the characteristics of LCL genotyping in whole genome sequencing (WGS)[2,3]. In our previous study[4], we found a significant discrepancy in the number of rare *de novo* variants between LCL and WB-based samples in our WGS data set and hypothesized that such difference was mainly caused by LCL-derived artifacts. To examine this hypothesis, we selected a single LCL sample that is among the most frequently sequenced (NA12878, or Coriell Institute GM12878), and seven sets of this cell line's separately sampled WGS data released in between 2009 and 2019. We assume that the release date of each data set is proportional to the actual sampling date, and examine whether there are genotyping quality changes, while minimizing other confounding differences, over time.

Methods

We selected seven WGS sets of NA12878, and the selection criteria include high coverage ($>$ 30x), common sequencing platform (Illumina), and technology (paired-end reads). We downloaded all raw sequences from the NCBI Sequence Read Archive[5] or Genome in a Bottle[6] site as aligned BAM or CRAM format files. All data sets not already aligned to the GRCh38 reference genome were converted into FASTQ format using *samtools fastq* (v1.10) and realigned with *bwa mem* (v0.7.17-r1188) and the GRCh38 reference. Following the GATK best practice guidelines[7], each realigned BAM file was sorted by *samtools sort* and duplicate-marked by *picard MarkDuplicates* (v2.23.3). We ran *gatk HaplotypeCaller* (v4.1.4.1) to generate intermediate genotype calling files (gvcf) for each data set, merged per-set gvcf files by *gatk CombineGVCFs*, and ran joint genotyping by *gatk GenotypeGVCFs*. We used *samtools idxstats* to count mapped reads per chromosome shown in Table 1. For defining the types of variants, we used calculated variant consequences from VEP[8] tool (v100.2). To measure the relative genotyping discrepancy between two variant sets, we define the discordance rate of variant data set S_1, given S_2 as the ratio of mismatching variants, where S_2 has a variant (non-missing and non-HOM-REF genotype). Table 1 shows the discordance rate of six data sets as compared to *Pilot2*, the earliest sequenced set.

$$Discordance\_rate(S_1 \mid S_2) = \frac{Count(S_1 \neq S_2 \mid S_1, S_2 \notin \{./., 0/0\})}{Count(S_2 \notin \{./., 0/0\})}$$

Results

After applying the same genotyping pipeline for all data sets, $Pilot2$ shows the smallest number of total (non-missing and non-HOM-REF) variants in chromosome 1 among all tested sets. While $Pilot2$ was sequenced using shorter reads than the others, the average coverage (45.3x) is comparable to the other sets, so we conclude that the difference in the total number of variants does not originate from read length differences. Variant calling change happens not only among common variants but also in rare ones: as shown in Table 1, *Pilot2* has significantly smaller number of protein-truncating variants (PTVs) than all other variant sets. These PTV counts are not dependent on the total number of variants (e.g., *Pl.200x*), thus separately suggesting possible genotyping quality changes over time. Interestingly, the discordant rate against $Pilot2$ monotonically increases relative to release year. It makes another independent

observation supporting our hypothesis that some of the true variants available in the earlier sets have been subjected to arbitrary quality and/or calling changes in later sequenced data sets.

Discussion

Here we demonstrated that overall genotyping quality of LCL samples is likely affected by cell passage increase over time. Given that LCL bio-materials have been frequently sampled for over a decade, efforts to remove artifacts that are a consequence of cell passage will be needed. We will further examine whether there are regions where LCL genotype quality/calling changes commonly happen, and test methods to identify and correct such LCL artifacts in WGS.

Table 1: Variant Counts in Chromosome 1 of NA12878.

ID	Released	#Mapped Reads	Avg. Cov.*	#Variants	#PTV	Disc. Rate (%)
Pilot2	2009	236,963,395	45.3	323,292	22	-
High.cov	2012	201,760,367	81.8	373,583	27	4.71
Pl.200x	2013	532,063,931	215.8	410,392	29	5.04
Pl.30x	2013	130,584,986	52.9	384,894	30	5.11
Pcr.free	2013	62,236,108	60.5	405,105	32	5.34
CCDG	2018	77,071,137	46.7	395,589	30	5.34
Phase3	2019	63,199,231	38.1	395,590	29	5.39

* Average coverage here is defined as (mean read length in bp) x (#total mapped reads)/(total chr1 length).

References

1. J. H. Oh, Y. J. Kim, S. Moon, H. Y. Nam, J. P. Jeon, et al. Genotype instability during long-term subculture of lymphoblastoid cell lines. *J. Hum. Genet.*, 58(1):16–20, Jan 2013.

2. C. M. Schafer, N. G. Campbell, G. Cai, F. Yu, V. Makarov, et al. Whole exome sequencing reveals minimal differences between cell line and whole blood derived DNA. *Genomics*, 102(4):270–277, Oct 2013.

3. Lena M. Joesch-Cohen and Gustavo Glusman. Differences between the genomes of lymphoblastoid cell lines and blood-derived samples. *Advances in genomics and genetics*, 7:1–9, 2017.

4. Elizabeth K. Ruzzo, Laura Pérez-Cano, Jae-Yoon Jung, Lee-Kai Wang, Dorna Kashef-Haghighi, et al. Inherited and de novo genetic risk for autism impacts shared networks. *Cell*, 178:850–866, August 2019.

5. Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39:D19–D21, January 2011.

6. Justin M. Zook, Brad Chapman, Jason Wang, David Mittelman, Oliver Hofmann, Winston Hide, and Marc Salit. Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. *Nature biotechnology*, 32:246–251, March 2014.

7. Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20:1297–1303, September 2010.

8. William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17:122, June 2016.

Identifying Biases in Clinical Decision Models Designed to Predict Need of Wraparound Services

Suranga N. Kasthurirathne PhD[1,2], Joshua R. Vest PhD[1,2], Shaun J. Grannis MD,MS[1,2]
[1]Regenstrief Institute, Indianapolis, IN, USA; [2]Indiana University, Indianapolis, IN, USA

Introduction. Evidence of systemic biases in Artificial Intelligence (AI) solutions[1,2] have led to calls to rigorously investigate AI models for biases that impact marginalized and vulnerable populations. However, there has been limited efforts to investigate systemic biases present in AI models for the clinical domain. Previously, we developed a series of AI models capable of predicting need of wraparound services, which are defined as additional non-medical services that are provided in conjunction with primary care[3]. We developed AI models predicting need of referrals to wraparound services for behavioral health, social work, and dietitian visits, as well as other services such as respiratory therapy, financial planning, medical-legal partnership assistance, patient navigation, and pharmacist consultations. These models were implemented across nine federally qualified healthcare centers in Indianapolis, IN to predict need of referrals[3]. In this study, we inspect each AI model for evidence of harmful biases across multiple demographic factors.

Materials and methods. We identified a population of adults (>=18 years) with at least one outpatient visit at Eskenazi Health, a county-owned urban safety net provider located in Indianapolis, IN. We extracted a comprehensive list of patient-level demographic, diagnosis, medication, and past visit history data from the Indiana Network for Patient Care (INPC), one of the largest, continuously operated statewide Health Information Exchanges (HIE) in the United States[4]. We used the Gradient Boosting (XGBoost) classification algorithm to develop four AI models capable of predicting need of referrals for behavioral health, social work, dietitian visits, and all other referral types. As with our previous efforts, we restricted the dietitian referral model to a subset of patients with specific risk conditions[3]. For bias detection, we identified three demographic features (race, age, and gender) as 'protected attributes' which present considerable risk of causing biases[5]. We will use these protected attributes to partition the population into two groups, patients who may be advantaged or disadvantaged based on each attribute. We evaluate biases by investigating statistical measures assumed to be equal across groups partitioned using each attribute (Table 1). Fairness and bias measures are context-dependent constructs. A variety of metrics have been proposed to investigate biases across these constructs. We used the fairness tree method[1] to select the most appropriate bias detection metric for our use case and applied this metric to each AI model using the AI Fairness 360 framework[2], which supports a wide variety of well-established bias detection metrics (Figure 1).

Results. We identified a total of 72,484 adult patients from an urban, primary care safety-net population: predominantly female 47,187 (65.1%), ethnically diverse (~25% white, non-Hispanic), and with high chronic disease burdens. Of these, 15,867 (21.9%) were eligible for inclusion in the dietitian model. Need of referrals, which constituted our gold standard reference, were behavioral health (12,162/72,484, 16.8%), social work (4104/72,484, 5.7%), dietitian counseling (4330/15,867, 27.3%), and other services (17,877/72,484, 24.7%). Performance of each AI model, as optimized for F1-score, was high, and compatible to prior modelling efforts[3] (table 2). We selected False Negative Rate (FNR) parity, which characterizes the degree to which model predictions report similar false negatives scores across advantaged and disadvantaged populations defined by each protected attribute. The fairness tree method recommended this metric because our AI models were designed to be assistive in nature, to prioritize predictive equity for patients in need, and because our interventions were designed to be applied to a broader population[1]. We found that FNR parity for each protected attribute and AI model were considerably low (< 0.07), indicating no evidence of biases (table 3).

Discussion. We were able to reproduce AI models with predictive performance metrics which were both high and comparable to our original effort[3]. Investigation using the AI fairness 360 framework found no indication of biases based on patient age, gender or race across any of the models under test. Therefore, we conclude there is a low likelihood that patient age, gender and race are introducing bias into our algorithms. Next steps include expansion of our analysis to investigate biases caused by social determinants such as homelessness, poverty, and unemployment, and individual-level bias metrics, which contrary to group based metrics used in this effort, investigate biases on the principal that similar individuals should be treated similarly irrespective of any protected attributes[6]. Further, our investigation may be further refined by use of additional advantaged and disadvantaged categories for each protected attribute. While our results indicated considerably low FNR parity scores, determining threshold of bias for larger scores requires a broader conversation with a multi-stakeholder group. In the event that models are found to be biased, they can be re-calibrated using a variety of bias mitigation methods also available via the AI Fairness toolkit.

Table 1. Advantaged vs. disadvantaged values for each protected attribute.

Protected attribute	Advantaged vs disadvantaged values
Gender	Advantaged value: male. Disadvantaged value: all others
Race	Advantaged value: non-Hispanic whites. Disadvantaged value: all others
Age	Advantaged value: 18 - 65 years. Disadvantaged value: >= 65 years

Figure 1. The complete study approach from data collection, AI model development to evaluation of biases.

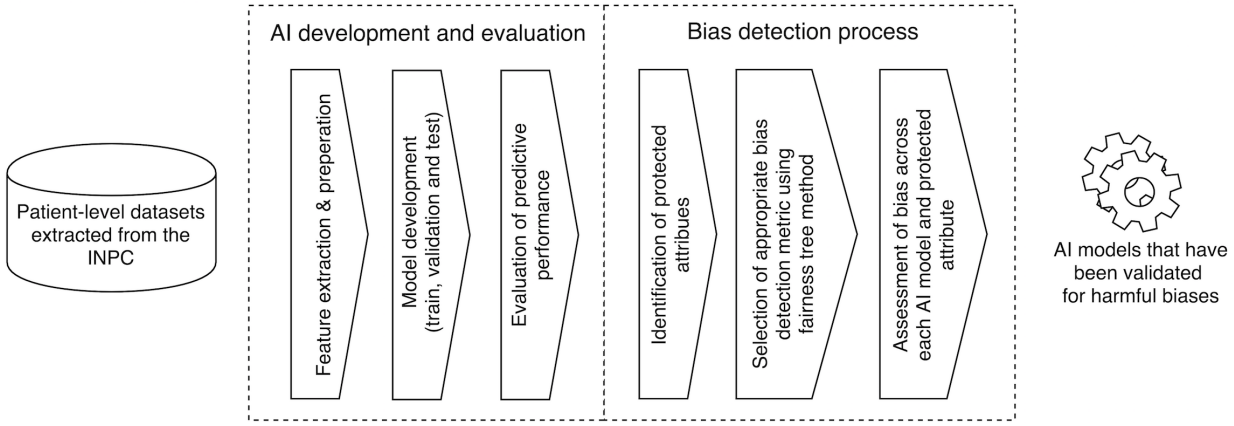


Table 2. AI model performance metrics.

Performance metric	Behavioral health (%)	Social work (%)	Dietitian (%)	Other (%)
Sensitivity	83.5 (83.0, 88.9)	72.5 (69.4, 75.7)	75.13 (70.6, 77.2)	59.1 (56.7, 63.5)
Specificity	99.2 (98.6, 99.8)	99.2 (99.1, 99.5)	93.2 (90.7, 94.4)	92.6 (89.5, 96.2)
F1-score	90.3 (87.5, 93.6)	82.3 (79.5, 85.4)	77.9 (73.2, 80.6)	64.9 (62.7, 67.7)
Precision	95.1 (92.1, 98.2)	95.5 (93.4, 97.5)	79.6 (76.3, 84.1)	73.6 (70.7, 77.3)
AUROC	98.2 (97.5, 98.6)	93.6 (92.7, 95.3)	91.3 (90.2, 92.4)	85.6 (84.5, 86.1)

Table 3. FNR parity for each AI model and protected attribute.

Protected attribute	Behavioral health	Social work	Dietitian	Other services
Gender	0.0504	0.0274	-0.0233	-0.0635
Race	-0.0089	-0.0082	-0.0009	0.0056
Age	0.0334	-0.0320	-0.0139	0.0113

References

1. Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, et al. Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:181105577. 2018.
2. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:181001943. 2018.
3. Kasthurirathne S, Grannis S, Halverson P, Morea J, Menachemi N, Vest J. Use of Patient and Population-Level Datasets to Identify Need of Wraparound Social Services: A Precision Health Enabled Machine Learning Approach. JMIR Medical Informatics. 2020.
4. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, et al. The Indiana network for patient care: a working local health information infrastructure. Health affairs. 2005;24(5):1214-20.
5. Hall WJ, Chapman MV, Lee KM, Merino YM, Thomas TW, Payne BK, Eng E, Day SH, Coyne-Beasley T. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. American journal of public health. 2015 Dec;105(12):e60-76.
6. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R, editors. Fairness through awareness. Proceedings of the 3rd innovations in theoretical computer science conference; 2012.

Using Patient Counting Analytics to Enhance Data Quality in a Nationwide Clinical Research Network

Jeffrey G. Klann, PhD[1,2,3]; Michele Morris[4]; Vivian Gainer, MS[2]; Michael Mendis[2]; Elaina Sendro, MS, MBA[5]; Shawn N. Murphy, MD, PhD[1,2,3]

[1]Harvard Medical School; [2]Research Information Science and Computing, Mass General Brigham; [3]Laboratory of Computer Science, Massachusetts General Hospital – All in Boston, MA ; [4]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh PA; [5]The Chartis Group, Pittsburgh, PA

Introduction

Clinical research data networks are proliferating, with the promise that they will enable rapid discovery and feedback into the healthcare system. The clinical research pipeline beings with the preparatory-to-research stage, focused on defining and finding patient cohorts, and data networks increasingly provide tooling to support this phase of research (e.g., the i2b2 Query Tool, OHDSI's ATLAS, PCORnet's query builder, etc).

However, the prevalence of large-scale networks has highlighted the non-uniformity of clinical data across sites, increasing the urgency of ascertaining network-wide data quality prior to the cohort building stage. Warehouse-wide quality is often focused on detecting problems in data loading, through libraries of rule-based checks.

A novel approach to enhancing early-pipeline data quality would be to scale-up these libraries of checks to gather details on possible cohorts to support preparatory-to-research work. In particular, we propose that enabling analytics on all one-concept cohorts could help researchers and site administrators understand the quality of the data at different sites for specific use-cases. This could be done by simply counting the number of patients at every site, for every possible code and code-set in the network, at every data refresh. This would allow analytics on:

1) **Variation Across Sites.** Counting the number of patients with a concept, such as number of diabetics across a network, allows the computation of an average and standard deviation so that outlier sites can be flagged. This also allows comparison of coding differences across sites (e.g., different diagnoses codes for diabetes).
2) **Variation Across Time.** As the total data volume increases, patient counts would be expected to increase gradually over time (at each refresh) at each site. Quality issues can be seen when there is a divergence from a logical progression of patient counts at each refresh for each concept.

A new approach is needed to pre-compute these data-quality cohort counts. Modern clinical research networks have millions of concepts in their terminology dictionaries, with data on millions of patients. Even real-time cohort query tools are not designed to operate at such scale.

We are engaged in a 12-month pilot project to implement this new approach to data quality in a national network, by building distributed counting scripts, aggregating the results, and creating a web-based dashboard for researchers and site administrators to look for outlier concepts.

Methods

The Accrual to Clinical Trials (ACT) network is an NIH-NCATS sponsored network of 48 clinical sites across the United States that have adopted a common data warehousing platform (Informatics for Integrating Biology and the Bedside, i2b2) and set of shared medical concepts (an i2b2 ontology) [1]. This 2.5-million element ontology encodes a hierarchy of biomedical knowledge encompassing many common data domains such as medications (in RxNorm), diagnoses and procedures (in, e.g, International Classification of Diseases - ICD), and labs (in Logical Observation Identifiers Names and Codes - LOINC). Local sites map data to these ontology terms, which allows data to be traversed at varying granularity. For example, a high-level term like *diabetes* will include all the child terms, like *diabetes mellitus with ketoacidosis*, automatically. Network nodes are linked together through the Shared Health Research Informatics Network (SHRINE) platform, a distributed query system for i2b2 nodes. Any site can initiate a SHRINE query using a multi-site query tool that allows real-time querying of Boolean logic queries in a user-friendly interface.

We developed a high-performance method for each site to count the number of patients with each medical concept, designed entirely in SQL, leveraging the set operations and indexing that make relational databases so efficient. The SQL tool allows an ACT node to compute all of its single-item "cohorts" at one time, a scale at which SHRINE was never designed to handle. Our method also offers enhanced patient privacy protection by adding Gaussian noise and a low-patient threshold to the counts, following the same design as SHRINE. An aggregation script brings these exports together into a central repository database.

For this pilot phase, we are designing an analytics dashboard powered by a hierarchical ontology view similar to the i2b2 query tool. A user-driven component allows interactive visualization across elements of the ontology, and

the ontology browser is used to flag outlier elements in the ontology. We have identified the following four views to expose data quality information in the dashboard:

- **Explorer View:** As the user browses the ontology and selects items, the dashboard shows the patient counts at each site in the network (as a bar graph), or at each refresh at a single site (as a line graph).
- **Missingness View:** Items missing at the currently selected site are highlighted via the ontology .
- **Site Outlier View:** Instead of patient counts, this view shows the total *percent* of patients at each site with a given ontology item. The average and standard deviation are used to highlight outliers at the currently selected site in both the ontology and graphs.
- **Time Outlier View:** This view flags points where a given refresh has an unexpected change in the amount of data for an ontology item. For example, if the count of total patients with e.g., diabetes ever decreases, stays flat, or increases by more than a set threshold (usually two standard deviations), it is considered an outlier. This is inspired by Control Charts in manufacturing.

Results

We are presently implementing a pilot of this approach in a subgroup of ACT sites participating in COVID research.

We developed the counting scripts for three database platforms – Postgres, Microsoft SQL, and Oracle. Subsequently we undertook massive speed optimizations so that the counting would be efficient even on the 2.5-million concept ACT ontology. We are incorporating user feedback to support additional features, such as i2b2's recent support for multiple fact tables.

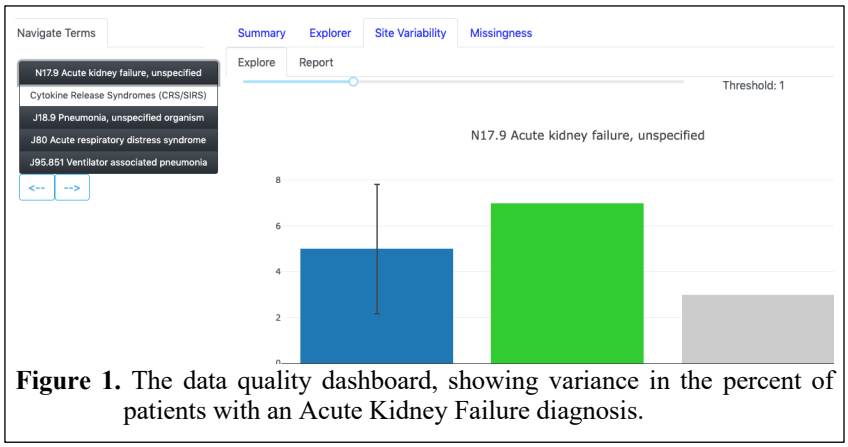


Figure 1. The data quality dashboard, showing variance in the percent of patients with an Acute Kidney Failure diagnosis.

We developed the interactive web-based dashboard using Python and the Dash framework, incorporating an interactive ontology view with bar graphs, line graphs, and text reports to support the four views described above. As users select their site and view, the ontology highlights outliers in the current view (i.e. missingness or site outliers).

ACT developed a specific ontology for studying COVID, with ~100,000 items. Sites participating in COVID research refresh their data approximately weekly. This provided an opportunity to initially test the platform on a rapidly evolving dataset. At the time of this writing, four sites have run the counting scripts on their COVID ontology, one of which has run the counting scripts a half-dozen times to support the time-outlier view. This provides preliminary data for dashboard development. By the March 2021, several sites will have run the counting scripts on their full ontology, and the dashboard will be deployed in an access-controlled environment for ACT researchers.

The pilot data has automatically flagged real data differences across sites: one site was missing all COVID-positive tests, a difference in coding at one site caused outliers in the number of patients having COVID-related diagnoses, and differences in RxNorm medication coding were found across the sites. As more data are collected, more results will become available. The presentation will present a detailed overview of these results.

The counting scripts will be part of the core i2b2 platform and the other components are available in GitHub (https://github.com/jklann/totalnum_tools).

Discussion

This method of pre-computing aggregated patient-centric concept counts across network sites extends the i2b2-SHRINE paradigm of live cohort-finding queries. It provides a scalable approach to collect information on millions of single element "cohorts" rapidly and then perform simple analytics on them to offer researchers and site administrators insights into the characteristics of the data in the network. The dashboard will help researchers develop queries for their preparatory-to-research work, and it will help site administrators identify anomalies in the data, all in a privacy-preserving way that uses only aggregate counts of patients. As a component of this pilot project (which ends in March 2021), we will develop a design to incorporate some of these features into a future version of SHRINE.

References

1 Visweswaran S, Becich MJ, D'Itri VS, *et al.* Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. *JAMIA Open* doi:10.1093/jamiaopen/ooy033

Classification of Chest Injury Severity Using Clinical Documents

Sujay Kulshrestha, MD[1,2], Dmitriy Dligach, PhD[3,4,5], Xin Su[5], Richard Gonzalez, MD[1,2], Cara Joyce, PhD[3,4], Matthew M. Churpek, MD, MPH, PhD[6], Majid Afshar, MD, MS[6]

[1]Burn and Shock Trauma Research Institute, Loyola University Chicago, Maywood, IL
[2]Department of Surgery, Loyola University Medical Center, Maywood, IL
[3]Center for Health Outcomes and Informatics Research, Health Sciences Division, Loyola University Chicago, Maywood, IL
[4]Department of Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, Maywood IL
[5]Department of Computer Science, Loyola University Chicago, Chicago, IL
[6]Department of Medicine, University of Wisconsin, Madison, WI

Introduction: Patients presenting after injury require providers to rapidly collect and process data to identify severity of injury and plan subsequent interventions. Formalized grading of injury severity currently is performed post-discharge by certified trauma coders manually annotating data to create trauma registries to create abbreviated injury scores (AIS) and injury severity scores (ISS) that allow for quality reporting and research. This method is costly and time intensive and is only employed by major certified trauma centers. Recent estimates suggest that 30-50% of patients with injury present to non-trauma centers, creating a significant blind spot to how injured patients are cared for in the United States. To our knowledge, no clinical decision support tools exist to automate estimation of injury scores from clinical documents to reproduce the scores manually calculated by trauma coders. A rich understanding of the data source and how its components evolve with time to affect model performance is essential to develop a parsimonious model. We aim to examine the performance of document type and time of document entry into the EHR for predicting chest injury severity in patients arriving to the emergency department (time 0) after trauma.

Methods: The Loyola University Medical Center (LUMC) Electronic Health Record (EHR) was queried for all patients presenting as a trauma activation. EHR data were linked by encounter identifiers to the internal trauma registry of patients maintained by the LUMC Department of Surgery and certified trauma coders. Severe chest injury was labeled using a thorax abbreviated injury score (AIS) cutoff for serious injury (AIS>2), which served as the binary outcome of interest for machine learning tasks. The AIS scores were labeled by trauma registry specialists credentialed and certified through the Registrar Certifying Board of the American Trauma Society, which is the gold standard for quality reporting. Free text clinical documents obtained from the EHR were converted to unigrams and concept unique identifiers (CUIs) obtained from linguistic processing using cTAKES version 4.0.0 to map named entities to the Unified Medical Language Systems (UMLS) Metathesaurus.

Unigrams and CUIs were used as bag of unigrams or bag of CUIs inputs to a logistic regression with elastic net regularization as both binary (unigram or CUI present versus absent) and normalized (term frequency-inverse document frequency) features at 1, 4, and 8 hours after emergency department arrival. Logistic regression was chosen for potential ease of implementation and for evaluation of interpretability for clinical relevance. Classifier hyperparameters were tuned to maximize area under the receiver operating characteristic curve (AUROC) using 10-fold cross validation on the training data (80%) with 15:1 case weighting. All model parameters including AUROC, positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity with 95% confidence interval (95% CI) results were reported from an independent, unweighted test data set (20%). Analysis of feature inputs, document type (i.e., admission notes, radiology reports, operative notes), and document time stamp were performed to compare AUROC by training separate classifiers with varied data inputs. The comparison between models was measured with net reclassification index test for nested models and the DeLong test for independent models.

Results: Between 2014 and 2018, 10,215 trauma patients were linked between the trauma registry and EHR, of which 6,891 had clinical document text available. Nearly 6.8% of trauma patients (n = 468) had at least a

serious chest injury (AIS >2). Increases in time interval from 1 to 8 hours increased the data corpus from 42,581 to 90,272 clinical documents. Models using CUIs in either binary or normalized fashion and unigrams in binary fashion performed similarly with an AUROC greater than 0.91 ($p > 0.05$ between model comparisons). In examining document types, the best performing were H&P, progress, and imaging documents with AUROC greater than 0.89. These documents had better performance in predicting chest injury severity than laboratory, nursing/ancillary staff, and operative documents ($p < 0.05$).

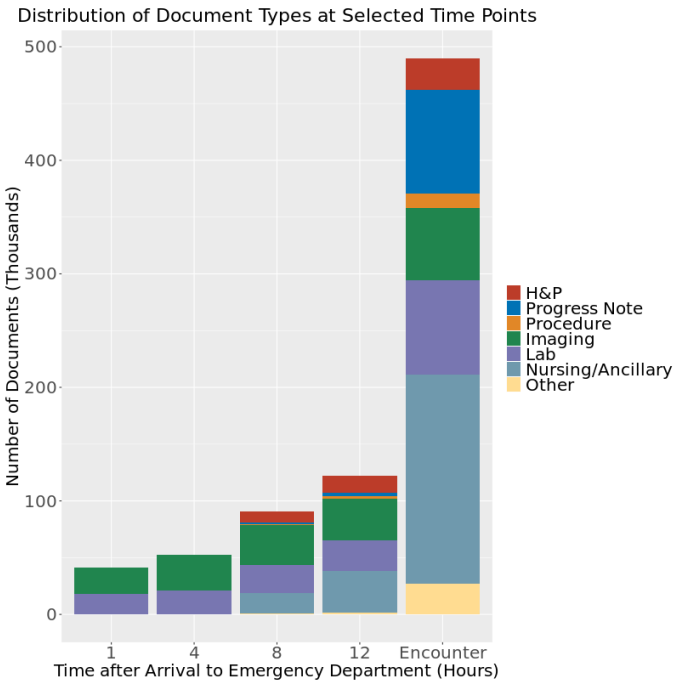
For models organized by time after presentation, the AUROC increased from 0.86 (95% CI: 0.82-0.91) at 1 hour to 0.94 (95% CI: 0.92-0.96) at 8 hours with an improved NRI of 0.18 (95% CI: 0.08-0.27, $p < 0.01$). Similarly, PPV increased from 0.28 (95% CI: 0.23-0.34) to 0.39 (95% CI: 0.33-0.46) and NPV remained stable between the 1- and 8-hour models (0.99 (95% CI: 0.98-0.99) versus 0.98 (95% CI: 0.97-0.99). Sensitivity and specificity of the 8-hour model was 0.86 (95% CI: 0.78-0.92) and 0.89 (95% CI 0.88-0.90) respectively, as compared with 0.78 (95% CI: 0.68-0.85) and 0.84 (95% CI: 0.82-0.86) for the 1-hour model. Inclusion of all 506,539 documents during the encounters did not improve model performance from the 8-hour mark with AUROC 0.95 (95% CI: 0.94-0.97) and NRI 0.06 (95% CI: -0.04-0.16, $p = 0.26$).

Discussion: Binary CUI, normalized CUI, and binary unigram models at 8 hours delivered the best overall performance for discriminating chest injury severity while minimizing data input. These were the most parsimonious models, as they required approximately 15% of the overall documents during a trauma encounter for accurate prediction. Overall, our model improved with time with a peak effect in AUROC at 8 hours, the time at which H&P documents enter the model, posing challenges for timely and accurate prediction at point of care until these types of notes are filed into the EHR. Additionally, further work and external institutional data is likely required to improve the precision of the model. These data will inform the ideal document time and document types to risk stratify patients in clinical decision support for prevention efforts of subsequent complications and to automate reporting of treatment of injury for centers without manual trauma coders.

References:

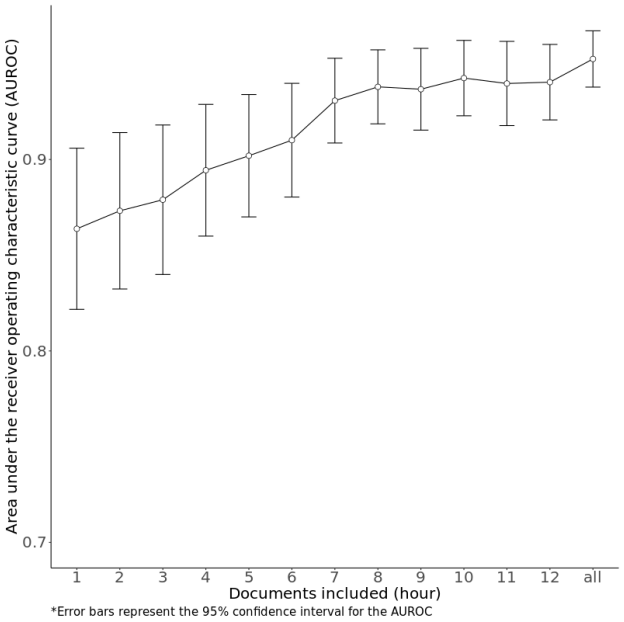
Kulshrestha S, Dligach D, Joyce C, Baker MS, Gonzalez R, O'Rourke AP, Glazer JM, Stey A, Kruser JM, Churpek MM, Afshar M. Prediction of severe chest injury using natural language processing from the electronic health record. Injury. 2020 Oct 25

Figure 1: Distribution of Document Types at Selected Time Points.



*Other notes included administrative documentation, death notes, consent notes, and discharge summaries.

Figure 2: Performance with Area Under the Receiving Operating Characteristic Curve across time intervals for predicting severe chest injury time intervals for predicting severe chest injury.



*Error bars represent the 95% confidence interval for the AUROC

Deriving family history scores from electronic health record data for prediction of disease risk

Nicholas B. Larson, MS PhD[1], Yiqing Zhao, Ph.D.[1], Jennifer St. Sauver, PhD, Véronique L. Roger, MD MPH[1], Hongfang Liu, PhD[1], Suzette J. Bielinski, MEd PhD[1]
[1]Mayo Clinic, Rochester, MN, USA

Introduction

Patient information extracted from electronic health records (EHRs) presents promising opportunities to develop risk prediction algorithms for various diseases(1). However, many features are stored in an unstructured free-text format, including family medical history. Family history (FHx) is routinely collected by clinicians to capture risk of various heritable complex diseases. The simplest representation of positive FHx is a binary indicator of an affected first degree relative (FDR; parent, sibling, offspring). More sophisticated weighted FHx scores additionally take into account the number of relatives and degree of relatedness, and can be more informative risk predictors for various conditions. For EHR-based data, however, lexical ambiguities in unstructured FHx information present challenges for scoring methods. Here, we extend our previous work(2) by outlining strategies to fully leverage available EHR data to quantify FHx of disease. We apply these methods to a large EHR-based patient population cohort to explore the distribution of FHx information for acute myocardial infarction (MI). Finally, we consider risk prediction models incorporating simple and complex representations of derived FHx, accounting for established clinical and demographic MI risk factors.

Methods

We derived an EHR-based population cohort using the Rochester Epidemiology Project (REP)(3, 4), including all individuals 45 years of age or older who resided in Olmsted County, Minnesota on January 1, 2006 and were free of cardiovascular disease. Incident MIs were collected from index through September 2017 from a long standing surveillance study.(5) Traditional MI risk factors used in the ACC/AHA Atherosclerosis Cardiovascular Disease (ASCVD) risk score(6) were derived for each patient at the index date. FHx content was retrieved from the "family history" section of unstructured clinical notes, as described previously(2). Briefly, MedTagger was used to extract mentions of family members. Disease mentions were extracted using the MetaMap API and Unified Medical Language System (UMLS) dictionary 2018AA.(7, 8) UMLS concepts were further mapped to Clinical Classifications Software (CCS) codes(9). Relationships between family member and disease were extracted using combined grammatical rules and distance-based rules. For MI, the CCS code 100 "Acute Myocardial Infarction" was used to capture relevant FHx of MI.

We considered two separate representations of FHx for comparative purposes. The first identified patients with any affected FDR (i.e., yes/no). We also defined a kinship-weighted FHx score based on all first degree and second degree relationships (i.e., grandparents, half-siblings, aunts/uncles, nephews/nieces), such that the score for subject i is defined as $S_i = x_{1i} + 0.5 \times x_{2i}$, where x_{1i} and x_{2i} are the numbers of unique affected first and second degree relatives, respectively. Given the lexical ambiguity of quantity (e.g., "brothers"), side of family (e.g., "paternal grandmother" vs. "grandmother"), gender (e.g., "sibling" vs "sister"), and redundancy (e.g., "grandparents" and "maternal grandfather" both listed), we developed a conservative and parsimonious scoring logic to consider any plural mentions of relations to indicate two affected relatives and collapsed any potential redundancies to a single relative with respect to the listed sources of ambiguity.

To evaluate the additional benefit of the FHx score vs. binary FDR representation of FHx, we fit Cox proportional hazards regression models adjusting for traditional risk factors to test associations with FHx predictors and estimate hazard ratios (HR) and 95% confidence intervals (CI). The concordance index (C-index) was used to compare risk prediction models, and missing covariate data were addressed via multiple imputation.

Results

The cohort consisted of N = 42,936 subjects, of which 55% were female and had mean age of 59 years (SD = 11.6). A total of 8431 subjects (19.6%) were positive for FDR family history of MI, while 9805 subjects (22.8%) had an FHS>0 (median = 1.00, range = [0.25,6.00]. There were 1199 incident MI events during approximately 11.7 years of follow-up. Kaplan-Meier plots of incident MI were stratified by FDR (panel A) and FHx score (panel B). FDR was strongly associated to MI occurrence after accounting for other ASCVD risk factors (HR = 2.53 [2.23,2.82], P<1e-16). The FHx score was significant (HR = 1.47 [1.56,1.63] per unit increase, P<1e-16). An FHx score >1

remained associated with MI occurrence after accounting for the presence of an affected FDR (P=0.001), indicating additional information gained through the FHx score. A Cox model using a smoothing spline relationship for FHx corresponded to a C-index of 0.766 vs. 0.722 using no family history data at all.

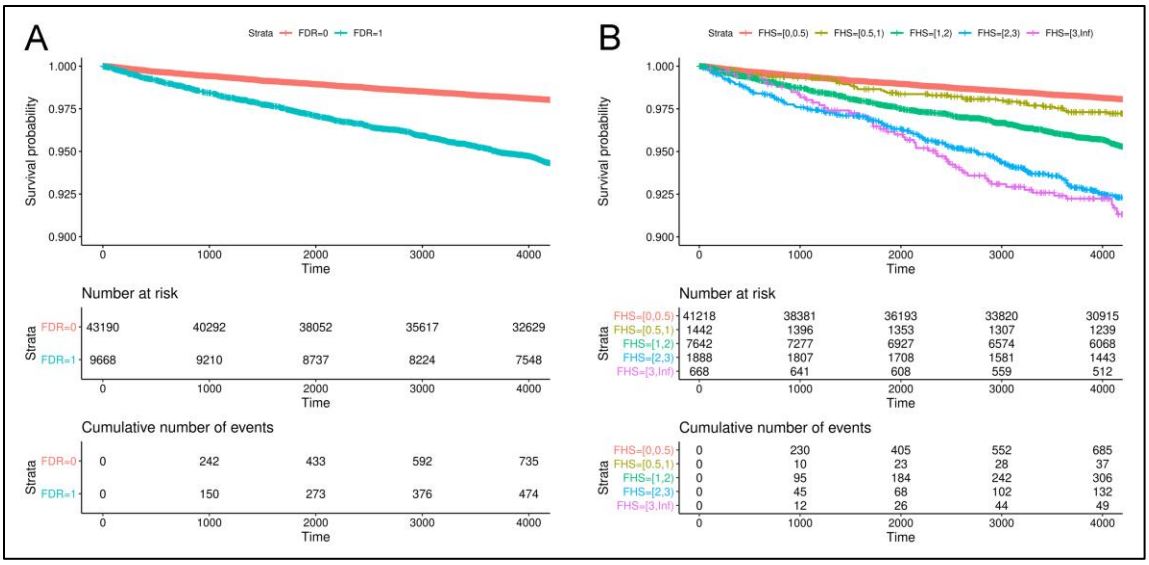


Figure 1. Kaplan-Meier plots of MI events stratified by (A) affected FDR at baseline and (B) ranges of the EHR-derived FHx score (FHS) for Acute MI.

Conclusion

We have outlined an algorithmic framework for translating family history data from EHRs into informative FHS values, and have illustrated the benefit of this approach using real data for risk of MI. These methods will allow for more sophisticated risk prediction models via machine learning incorporating other relevant conditions (e.g., hypertension, hyperlipidemia) as well as data-mine relationships between various disease FHx scores and coded conditions in the EHR to discover novel cross-disease relationships.

References

1. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198-208.
2. Wang Y, Wang L, Rastegar-Mojarad M, Liu S, Shen F, Liu H. Systematic Analysis of Free-Text Family History in Electronic Health Record. AMIA Jt Summits Transl Sci Proc. 2017;2017:104-13.
3. Rocca WA, Grossardt BR, Brue SM, Bock-Goodner CM, Chamberlain AM, Wilson PM, et al. Data Resource Profile: expansion of the Rochester Epidemiology Project medical records-linkage system (E-REP). Int J Epidemiol. 2018;47(2):368-j.
4. St. Sauver JL, Grossardt BR, Yawn BP, Melton LJ, 3rd, Pankratz JJ, Brue SM, et al. Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system. International journal of epidemiology. 2012;41(6):1614-24.
5. Roger VL, Weston SA, Gerber Y, Killian JM, Dunlay SM, Jaffe AS, et al. Trends in incidence, severity, and outcome of hospitalized myocardial infarction. Circulation. 2010;121(7):863-9.
6. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129(25 Suppl 2):S49-73.
7. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229-36.
8. U.S. National Library of Medicine. Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/ (accessed 06/09/2020).
9. Cohen JW, Cohen SB, Banthin JS. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. Med Care. 2009;47(7 Suppl 1):S44-50.

Differentiating between Hemorrhage and Sepsis for Hypotensive Subjects Using Arterial Pressure Data

Xinyu Li, MS[1], Ernest Pokropek[1,2], Michael R. Pinsky, MD[3], Artur Dubrawski, PhD[1]
[1]Auton Lab, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA;
[2]Institute of Telecommunications, Warsaw University of Technology, Warsaw, Poland;
[3]Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA USA

Introduction

Hemorrhage and sepsis are two critical medical conditions which share hypotension as the major early symptom. However, induced by different causes, these two conditions require different treatments and if not timely treated, both can lead to shock and, eventually, death. In current practice, the recognition of the specific condition that a hypotensive patient is exposed to highly depends on the vigilance of the clinical personnel. Previous studies[1,2] mainly focus on the prediction and analysis of sepsis using high-resolution physiological data instead of the distinction between sepsis and hemorrhage. In this work, we propose a data-driven machine learning approach to differentiate between these two conditions in hypotensive subjects by utilizing large amounts of routinely collected physiological time series. Our approach, demonstrated in laboratory animal experiments, confidently identifies the majority of septic subjects apart from hemorrhagic subjects at early stages of hypotension episodes. The proposed method has the potential to inform timely treatment decisions and facilitate favorable patient outcomes in critical care settings.

Methods

Data

We used laboratory animal data for our analysis. 22 healthy Yorkshire pigs were anesthetized and stabilized for one hour. Then 15 subjects were subject to induced bleeding at a constant rate of 5 mL/min until their mean arterial pressure (MAP) decreased to 30 mmHg, and the other 7 subjects were given two subsequent lipopolysaccharide (LPS) infusions to induce sepsis. The first LPS infusion was intended to weaken the subjects' immune mechanisms, and the second infusion was performed to induce sepsis-related symptoms. Arterial pressure was collected at 250 Hz.

We utilized data during the first episode of induced bleeding, and the second infusion of LPS, respectively, for the two groups, and extracted data from hypotension onset to the end of the experiments. Hypotension was defined as arterial pressure dropping below 60 mmHg. Statistical features including mean, median, standard deviation, and range, as well as spectral (Discrete Fourier Transform) features of different frequency bands (0.04-0.15Hz, 0.15-0.4Hz, 0.4-10Hz, 10-125Hz) were computed for 4-minute moving time windows updated at 2 Hz, extracted from arterial pressure waveform. These features were standardized using the first minute of induced bleeding or LPS infusion to mitigate possible subject-dependent biases introduced by the induced pathologies, so that the arterial pressures of two groups were at comparable levels.

Model and Evaluation Protocol

We trained a Random Forest machine learning model over the statistical features to differentiate between hemorrhage (negative class) and sepsis (positive class). To accommodate a relatively small cohort size, we evaluated the performance using leave-one-pair-out cross-validation: one negative subject and one positive subject, both randomly chosen, were held-out as the test set and the model was trained using the remaining subjects. This process was repeated 105 times (15 negatives x 7 positives), and the mean and 95% confidence intervals (CI) of performance metrics were reported.

Results

When evaluated on the entire period from the onset of hypotension to the end of induced bleeding or LPS infusion, our approach achieves AUROC 0.878 ± 0.042 (mean and 95% CI), and is able to confidently identify 64.6% septic patients while only giving 1 false alert out of 10,000 such predictions on average, as shown in Figure 1. We also evaluated the models on different truncated time intervals. As shown in Figure 2, during 5-10 minutes after the onset of hypotension, our approach achieves AUROC 0.852 ± 0.058, and identifies 80.4% septic patients at extremely low false positive rate, and during 10-15 minute interval, our approach improves to AUROC 0.910 ± 0.049, and identifies 87.8% septic patients with high confidence, as more discriminative evidence becomes available over time in arterial pressure waveforms.

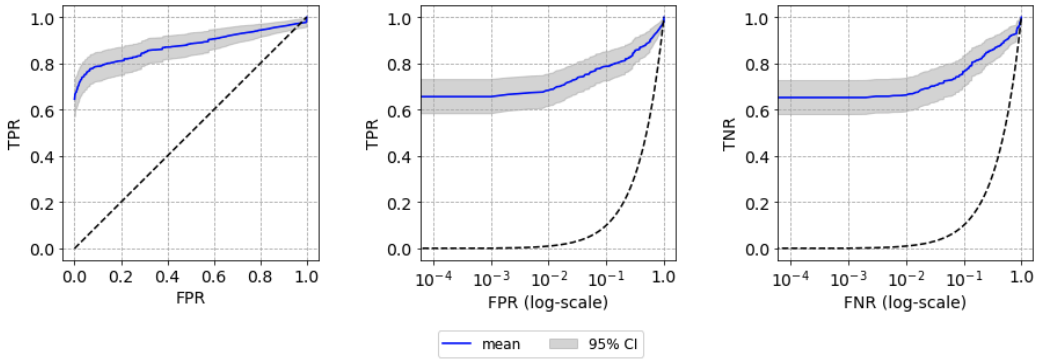


Figure 1: ROC curves when our approach is evaluated on the entire period of disease. False Positive Rate (FPR) and False Negative Rate (FNR) in the middle and right plots are shown in logarithmic scales to emphasize the performance at clinically relevant low FPR and low FNR settings. Dashed lines represent a random predictor for reference.

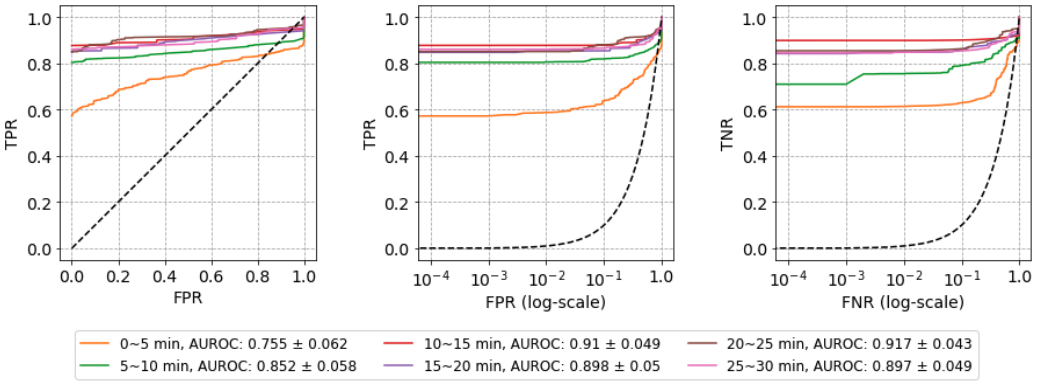


Figure 2: Mean ROC curves when our approach is evaluated at different truncated time intervals into the disease. The mean and 95% CI of AUROC are shown in the legend.

Conclusions and Discussion

Our results show that during a short time period after the onset of hypotension, our approach is capable of identifying the majority of septic subjects from hemorrhagic subjects using features derived from arterial blood pressure waveforms, despite both groups having similar mean arterial pressures. By utilizing the high-frequency physiological data collected from patients continuously monitored at the bedside, it should be possible to differentiate sepsis from hemorrhage within minutes of the onset of hypotension, and inform treatment that substantially differs between these two conditions. Our approach has the potential to be applied in critical care settings to support clinical decision making and resuscitation resource allocation. Future work includes the exploration of the utility of hemodynamic vital signs other than arterial pressure, and further investigation of the generalizability of the proposed approach on larger and more complicated datasets collected in bedside monitoring of human subjects.

Acknowledgements

This work has been partially supported by DoD (W81XWH-19C-0101), DARPA (FA8750-17-2-0130) and NIH (R01GM117622).

References

1. Hatib, F., Jansen, J. R., Pinsky, M. R. (2011). Peripheral vascular decoupling in porcine endotoxic shock. Journal of Applied Physiology, 111(3), 853-860.

2. Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., Buchman, T. G. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Critical Care Medicine, 46(4), 547-553.

Rapid Creation of a De-Identified COVID-19 Dataset for Clinical Research

Tanja Magoc, PhD[1], Yankuic Galvan, PhD[1], Richard Deason, MS[1], Jennifer N. Fishe, MD[2], Guillaume Labilloy, ME, MBA[2], Gloria Lipori, MT, MBA[3], Ian Tfirn, MPH[2], Christopher A. Harle, PhD[1]

[1]University of Florida College of Medicine, Gainesville, FL, [2]University of Florida College of Medicine-Jacksonville, FL, [3]UF Health, Gainesville, FL

Abstract

In addition to clinical trials, many researchers wish to analyze electronic health record (EHR) data to combat the COVID-19 pandemic. To support this new demand, we rapidly developed, disseminated, and iteratively updated a reusable de-identified research dataset of patients with both confirmed COVID-19 and COVID-19-like symptoms. The dataset was formatted using the Observational Medical Outcomes Partnership (OMOP) common data model, and is available to researchers across our institution without need for IRB approval.

Introduction

The COVID-19 pandemic has markedly increased demand for both retrospective and prospective research data, including electronic health record (EHR) data (1-5). In many academic health centers, research requests for EHR data typically require study-specific Institutional Review Board (IRB) approval, and an honest-broker process for specifying, extracting, and transmitting the dataset to study teams (6). That process is time-consuming and labor-intensive, and not well suited for the urgency that COVID-19 research demands. Therefore, to allow University of Florida (UF) researchers to rapidly analyze COVID-19-related EHR data, we built a reusable, de-identified COVID-19 dataset that is available to researchers without study-specific IRB approval, nor requires data specification or extraction. Herein we describe the features of the reusable COVID-19 dataset, and lessons learned in developing and disseminating the data to researchers.

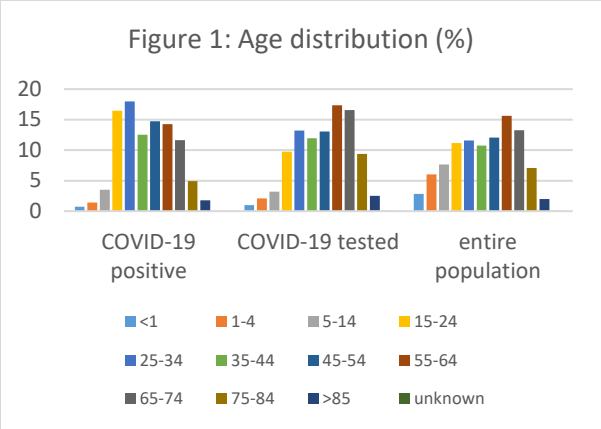
Methods

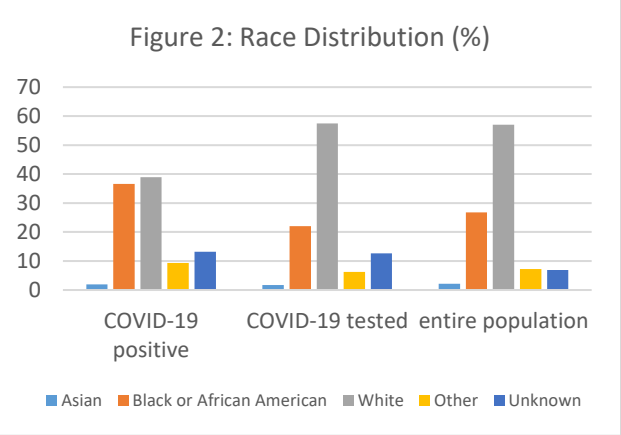
The Integrated Data Repository team at UF prepared a reusable de-identified dataset containing past and present medical history of patients of interest, including all patients with COVID-19-like symptoms such as respiratory illness, cough and fever, as well as all COVID-19-tested patients, and patients treated with Remdesivir since January 1, 2020. Note that many patients tested for COVID-19 were asymptomatic and tested pre-elective surgery or post-admission for infection control reasons. Because many research studies rely on past medical history, we also provided all available health data since January 1, 2012. Data also is prospectively provided up to the date of the dataset's release, for monitoring of outcomes and sequelae. As the dataset is refreshed regularly, the end date is a moving target.

All data elements are packaged into the reduced format of the OMOP common data model version 5.3 (7). We chose OMOP because of its global and increasing frequency of use. OMOP includes patient-level data such as demographics, diagnoses, procedures, labs, medications, and some vitals as well as encounter-level data such as start and end date, encounter type, and admit source and discharge status for inpatient visits. To de-identify the data consistent with HIPAA, we removed all Protected Health Information (PHI) from the dataset in two steps: (1) replaced HIPAA identifiers with mock-up identifiers, (2) shifted dates by randomly selecting a number between -30 and 30 for each patient and shifting any date associated with the patient that number of days.

Results

Data Summary: The dataset contains more than 100,000 patients as of its fourth release. We labeled three different cohorts and provide aggregate analysis based on them: COVID-19-positive patients, COVID-19-tested patients (regardless of the test results), and the entire sample population. The COVID-19-positive cohort is less than 3% while COVID-19-tested cohort is 43% of the sample population. At this time, the dataset is current through July 22, 2020, and shows the highest percent of COVID-19-positive patients in age groups 15-34 (Figure 1). However, the percent of tested patients is the highest in 55-74 age groups. Interestingly all three cohorts have average ages in 40s. Figure 2 shows that the COVID-19-positive population is relatively equal between African American and white races, although the percent tested is doubled for the white population. We also examined inpatient data and found mean length of stay (LOS) was slightly higher **for** COVID-19-positive patients (6.5 versus 6 days for other encounters), and the ICU LOS was also higher for COVID-19 patients (6.7 vs 5.8 days).





Lessons Learned: Our previous close collaboration with our institution's IRB was extremely valuable, allowing for rapid decisions regarding procedures for deidentifying the dataset and disseminating to researchers. Additionally, our team worked closely with UF Health IT to better understand the EHR data format and EHR changes due to the pandemic. We also learned it was necessary to educate researchers on OMOP and how to properly identify their cohort of interest. We have issued several versions to date, however the first dataset was initially disseminated to a few research collaborators who provided constructive feedback. Their suggestions ranged from desirable data elements that were not originally included in the dataset (e.g., interest in EKG readings for patients on hydroxychloroquine), to the format and necessary instructions for using a common data model.

Discussion

We created a reusable, OMOP-based dataset that allows researchers across our university to study patients with confirmed or suspected COVID-19, as well as all patients tested. That approach of creating, governing, and disseminating EHR data for research in the future can pave the way for wider and more efficient production and dissemination of enterprise EHR data for research (as opposed to fulfilling study-specific research data requests). Because of its de-identified nature, one limitation of this dataset is that researchers cannot use it to examine time trends across patients. Also, to date, this dataset includes only structured EHR data. Our future plans include updating the dataset regularly, advertising the dataset more widely across our institution, adding more data elements of interest, and creating new reusable datasets for other high-interest health conditions.

References

1. Grein J, Ohmagari N, Shin D, Diaz G, Asperges E, Castanga A et al. Compassionate use of Remdesivir for patients with severe Codiv-19. *N Engl J Med*. 2020l 382:2327-2336. Available from: DOI: 10.1056/NEJMoa2007016.
2. Kashyap S, Gombar S, Yadlowsky S, Callahan A, Fries J, Pinsky BA, Shah N. Measure what matters: counts of hospitalized patients are a better metric for health system capacity planning for a reopening. *Journal of the American Medical Informatics Association*. 2020; 27(7):1026-1131. Available from: https://doi.org/10.1093/jamia/ocaa076
3. Kuderer NM, Choueiri TK, Shah DP, Shyr Y, Rubinstein SM, Rivera DR et al. Clinical impact of Covid-19 on patients with cancer (CCC19): a cohort study. *The Lancet*. 2020;395(10241):1907-1918. Available from: https://doi.org/10.1016/S0140-6736(20)31187-9.
4. Lee LYW, Baptiste Cazier J, Starkey T, Turnbull CD, UK Coronavirus Cancer Monitoring Project Team, Kerr R et al. Covid-19 mortality in patients with cancer on chemotherapy or other anticancer treatments: a cohort study. *The Lancet*. 2020; 395(10241):1919-1926. Available from: https://doi.org/10.1016/S0140-6736(20)31173-9.
5. McMichael TM, Currie DW, Clark S, Pogosjans S, Key M, Schwartz NG et al. Epidemiology of Covid-19 in a long-term care facility in King County, Washington. *N Engl J Med*. 2020; 382:2005-2011. Available from: DOI: 10.1056/NEJMoa2005412.
6. Campion TR, Craven CK, Dorr DA, Knosp BM. Understanding enterprise data warehouses to support clinical and translational research. *Journal of the American Medical Informatics Association*. 2020. Available from: https://doi.org/10.1093/jamia/ocaa089.
7. Observational Health Data Sciences and Informatics. *Data Standardization*. Available from: https://www.ohdsi.org/data-standardization [Accessed August 8th, 2020].

Discovering Changes in the Neonatal Intensive Care Unit Structures Before and During the COVID-19 Pandemic: A Network Analysis

Hannah Mannering[1], Chao Yan, MS[2], Mhd Wael Alrifai, MD[3], Yang Gong, MD, PhD[4], Daniel France, PhD[3], You Chen, PhD[2,3]

[1]Loyola University, Baltimore, MD; [2]Vanderbilt University, Nashville, TN; [3]Vanderbilt University Medical Center, Nashville, TN, [4]University of Texas Health Science Center at Houston, Houston, TX

Introduction

Healthcare organizations (HCOs) change intensive care unit (ICU) staffing during the COVID-19 (C19) pandemic to protect healthcare workers and patients[1-3]. These changes can interfere with collaboration structures in the ICU, which may impact care quality and patient safety[4-6]. ICU staffing plans (e.g., team scheduling) are historically developed at a coarse-grained level, which increases the challenge for HCOs to assess the impact of staffing plan changes on teamwork structures. For instance, ICU staffing plans seldom consider connections among healthcare workers in a team due to complex clinical workflows and handovers[5-6]. Thus, it is hard to measure teamwork structures. There is a big gap between the staffing plans and clinical outcomes, and thus it is challenging to monitor how team structure and clinician interactions affect clinical performance and care quality in high-risk settings. Understanding how healthcare workers connect (e.g., exchanging health information) within their clinical teams when caring for patients can provide fine-grained evidence to potentially optimize existing staffing plans.

As mentioned above, one of the major challenges to measure connections among healthcare workers is the complexity of care in ICUs. Recent studies applied network analysis to electronic health record (EHR) utilization data to address this challenge[5-6]. In modern healthcare, an increasing number of healthcare works utilize EHR of diagnose and treat patients by exchanging all medical statuses[5-6]. Therefore, the volume of the EHR system utilization data has been increasing exponentially in recent years, providing abundant resources to identify connections among healthcare workers. In this study, we aim to leverage network analysis methods to learn structures of neonatal intensive care unit (NICU) in pre- and intra-C19 in terms of collaboration among healthcare workers. Patients hospitalized in the NICU include high-risk infants, who may suffer from, or at risk for having a variety of complex diseases or conditions. Management of NICU patients typically require a variety of various healthcare workers and highly specialized consultants. Investigating teamwork structures to reduce the gap between staffing plans and clinical outcomes can potentially inform actionable collaboration interventions to improve care quality and patient safety in the NICU[5-6]. Tertiary-level NICUs have a highly density of intense EHR utilization as well as heavy data sharing traffic per patient episode, making this environment ideal to investigate the connections among healthcare workers[5-6].

Methods

We use the utilization of EHRs for patients admitted to the NICU at Vanderbilt University Medical center (VUMC, Nashville, Tennessee, USA) between September 1, 2019, and June 30, 2020. We characterized pre-C19 as the months of September through December of 2019 and intra-C19 as the months of March through June of 2020. These two groups are compared using patients' clinical characteristics, including their age, sex, race, length of stay (LOS), and discharge dispositions.

We apply network analysis to the utilization of EHRs of 712 NICU patients (386 pre-C19 and 326 intra-C19) excluding those with C19 to learn healthcare worker networks to describe teamwork structures of pre-C19 and intra-C19. The healthcare worker actions stem from different tasks, including conditions (e.g., assessing a patient's condition), procedures (e.g., intubation), medications (e.g., prescription), notes (e.g., progress note writing), orders (e.g., laboratory test ordering), and measurements (e.g., measuring blood pressure). Prior research indicates that a threshold of one-day can capture meaningful collaborative healthcare worker relationships[5-7]. Thus, we characterize a teamwork relationship (network connection) between two health care workers (network nodes) as healthcare workers who performed actions to EHRs of the same patient on the same workday (7am – 7pm) or work night (7pm – 7am). These periods are based on the observation that most schedule shifts in the NICU occur around 7am or 7pm. The weight of a relation between two clinicians on a day/night is the number of patients they co-managed using EHRs. The relation's final weight is the cumulative number of patients the two clinicians co-managed across the days/nights in the four months. By doing so, we build the pre-C19 and intra-C19 networks.

We leverage sociometric measurements, including eigencentrality, betweenness, and eccentricity, to quantify the network structures. Eigencentrality indicates a healthcare worker's leadership in terms of collaboration, betweenness demonstrates a healthcare worker cares for a wide spectrum of patients, and eccentricity shows the difficulty for a healthcare worker to collaborate with others. To calculate these sociometric measurements, we utilized Gephi, a

network analysis and visualization software. We investigate if the differences in the healthcare worker leadership, care for wide spectrum of patients, and collaboration difficulty are statistically different between pre- and intra-C19 teamwork structures. We apply a Mann-Whitney U tests at the 95% confidence level to account for non-Guassian distribution of the sociometric measurements. Since the pre- and intra-C19 networks are made up of healthcare workers with different specialty (e.g., NICU registered nurses), we compare the differences at both network- (entire network) and specialty-level (each specialty). We apply a Bonferroni correction to account for multiple hypothesis testing (e.g., pairwise test for the specialty-level comparisons.

Results

The pre- and intra-C19 patient groups share similar distributions in sex (~0 difference), race (4% difference in White, and 3% difference in African American), LOS (IQR difference in 1.5 days), and discharge dispositions (~0 difference in home, 2% difference in expired, and 2% difference in others), which shows there are no significant changes in patient demographics and outcomes ($p < 0.0001$) between the two groups. Also, the pre- and intra-C19 networks have no significant differences in the number of nodes and edges (>1K nodes and >11K edges). There are several notable findings in network analysis to highlight. First, it was found that the intra-C19 teamwork structure has a higher collaboration difficulty (increased eccentricity) than the pre-C19 ($p = 2.2 \times 10^{-6}$). Second, NICU registered nurses had a reduced leadership responsibility (lower eigencentrality) in the intra-C19 structure than the pre-C19 ($p = 2.64 \times 10^{-15}$). Third, neonatology physicians care for a wider spectrum of patients (higher betweenness) during the C19 pandemic ($p = 5.43 \times 10^{-3}$).

Discussion

Our network analysis captures collaboration difficulty (increased eccentricity) and major shifts of neonatology physicians (betweenness). In addition, NICU nurses have reduced leadership responsibility (lower eigencentrality) in cooperation, suggesting that increased EHR use may reduce nurses' workload in the collaboration. The developed network methods and three sociometric measurements can be reapplied to EHR systems of other HCOs to assess teamwork structure differences in current and future disruptions in healthcare delivery (e.g. pandemics, etc.), which may inform actionable staffing interventions to reduce the collaboration difficulty in EHRs. However, it is important to note that the utilization of the learned connections among healthcare workers will be dependent on the validation and interpretation of those connections. Furthermore, the connection between two healthcare workers indicates the potential collaboration (information sharing) rather than actual collaboration and recruiting subject matter experts (e.g. clinicians, etc.) to evaluate the learned connections and team structures will be required to validate the results. Since this is a pilot study, we want to point out some limitations to guide the future work. First, we did not investigate the effectiveness of our network analysis approach by measuring reliability of the connections using sensitivity analysis. Second, we did not investigate the impact of the changes in the network structures on clinician workload and healthcare cost, which are directly related to care management. Third, we did not consider other factors such as season in our cohort studies. Fourth, the definitions of workday and worknight may be inappropriate for some types of healthcare workers whose schedule shifts are different from 7am and 7pm.

References

1. Black JR, Bailey C, Przewrocka J, Dijkstra KK, Swanton C. COVID-19: the case for health-care worker screening to prevent hospital transmission. The Lancet. 2020 May 2;395(10234):1418-20.

2. Abir M, Nelson C, Chan EW, Al-Ibrahim H, Cutter C, Patel K, Bogart A. Critical care surge response strategies for the 2020 COVID-19 outbreak in the United States. Retrieved from RAND Corporation: https://www. rand. org/content/dam/rand/pubs/research_reports/RRA100/RRA164-1/RAND_RRA164-1. pdf. 2020.

3. Bhardwaj R. Mitigating the Adverse Consequences of Pandemics: A Short Note with a Special Reference to COVID-19. Available at SSRN 3565460. 2020 Mar 31.

4. Kim MM, Barnato AE, Angus DC, Fleisher LF, Kahn JM. The effect of multidisciplinary care teams on intensive care unit mortality. Archives of internal medicine. 2010 Feb 22;170(4):369-76.

5. Kim C, Lehmann CU, Hatch D, Schildcrout JS, France DJ, Chen Y. Provider Networks in the Neonatal Intensive Care Unit Associate with Length of Stay. In2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC) 2019 Dec 12 (pp. 127-134). IEEE.

6. Chen Y, Patel MB, McNaughton CD, Malin BA. Interaction patterns of trauma providers are associated with length of stay. Journal of the American Medical Informatics Association. 2018 Jul;25(7):790-9.

An Integrative Network-based Analysis of Genetic Variants and Differential DNA Methylation in Brain Tissue of Multiple Sclerosis Patients

Astrid M. Manuel, BS[1], Yulin Dai, PhD[1], Saurav Mallik, PhD[1], Leorah A. Freeman, MD, PhD[3], Peilin Jia, PhD[1], Zhongming Zhao, PhD[1,3,4]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, USA; [2]Dell Medical School, The University of Texas at Austin, TX, USA [3]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, TX, USA; [4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

Multiple sclerosis (MS) is a chronic neurological disorder in which the immune system attacks the central nervous system. MS brain lesions typically manifest in normal appearing white matter (NAWM), however, the pathological changes that occur remain poorly understood. Investigators have postulated that MS is a genetically predisposed disease, onset after an unknown environmental trigger. Previously, we have linked genetic variants of MS to drug target genes by network-based analyses[1]. In this study, our objective is to expand on our network-based methods to incorporate DNA methylation in MS NAWM tissue. Although genome-wide associated studies (GWAS) have identified several genetic variants associated to MS, monozygotic twins are often discordant for MS[2]. Epigenome-wide studies of MS NAWM have also identified differentially methylated CpG sites associated to MS, which suggests that methylation may be acting as a key epigenetic mechanism in MS[3]. We hypothesize that epigenetic factors may contribute to environmental causes and dysregulate genes in MS. We use a new implementation of our network-based methods to integrate MS GWAS data with DNA methylation in MS NAWM tissue to better understand the biological underpinnings of MS mechanisms.

Methods

Network-based methods are valuable to readily connect signals from different experimental platforms (e.g. GWAS, epigenome microarrays). We perform a new implementation of the Edge-Weighted Dense Module Search of GWAS (EW_dmGWAS) tool to integrate MS GWAS with methylation and expression data in MS NAWM by using the human protein interactome as the reference network[4]. All integrated datasets pertained to a case and control study type with individuals of European decent. Datasets were collected from publicly available sources (Table 1).

Table 1. Descriptions of datasets used for integrative MS network-based analysis.

Dataset	Data Accession	Experimental Platform	Sample Size
GWAS Summary Statistics[2]	http://imsgc.net/	MS Chip	14,802 cases, 26,703 controls
NAWM Methylation[3]	Gene Expression Omnibus (GEO)	HumanMethylation450 BeadChip	28 cases, 19 controls
NAWM Expression[5]	Gene Expression Omnibus (GEO)	RNA-sequencing	10 cases, 11 controls

CpG sites are locations of the genome in which methylation commonly occurs. In cellular systems, methylation of promoter regions often leads to gene expression regulation. For instance, hypermethylation of the promoter region leads to gene silencing. Methylation microarray experiments assay the methylation levels of CpG probes. However, the effects of single CpG sites is difficult to interpret. In this study, we used annotated genome analysis to obtain gene-level methylation scores, as this allows consideration of only promoter regions[3]. After annotating CpG probes to respective promoter regions of genes, we use Stouffer's Z-score method to combine p-values. In this way, we obtained interpretable gene-level methylation scores, which were attributed to node weights of networks. In our network analysis, the nodes were weighted by a sum of GWAS-based and methylation-based gene-level scores. Gene expression data was included in this study to calculate differential co-expression between pairs of genes in MS NAWM and was subsequently used as edges weight. Modules were assessed by a permutation test. The resultant modules were

further assessed using functional gene set enrichment analysis, including Gene Ontology (GO) annotations and drug targets of MS FDA-approved medications. Our list of drug targets included 32 drug target genes of MS.

Results

The input for our network was 12,380 genes matched for 256,187 protein-protein interactions (edges). EW_dmGWAS yielded 4,942 gene network modules associated to MS. The top 100 modules were comprised of a gene set of 168 genes. The top enriched GO Biological Process function for this gene set was "response to type I interferon" (FDR = 3.62×10^{-5}). Amongst top 10 enriched functions, we also found the GO Biological Process terms "interleukin-27-mediated signaling pathway" (FDR = 6.00×10^{-4}). Notably, the top 100 modules were also enriched with four drug targets of MS FDA-approved medications: *IFNAR1* (target of interferon-beta-1a), *KEAP1* and *RELA* (targets of dimethyl fumarate), and *S1PR5* (target of fingolimod and siponimod). For instance, the figure below shows the network module containing *IFNAR1*.

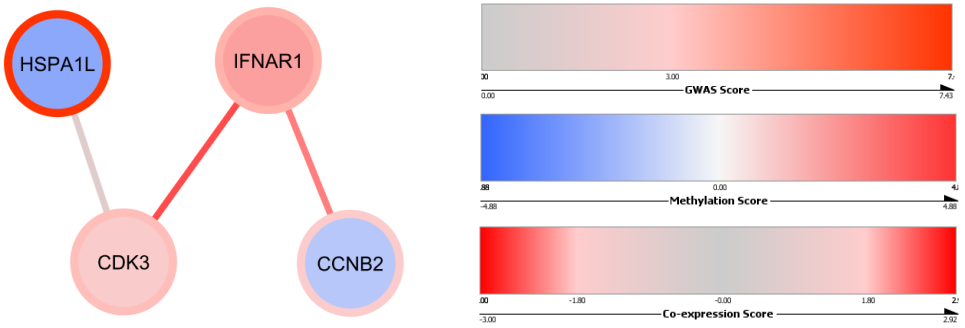


Figure 1. Top network module includes hypermethylated *IFNAR1*, drug target gene of the MS medication interferon-beta-1a. The color of node border indicates GWAS-based score (red signifies highly significant). The node-fill color represents methylation scores (blue indicates hypomethylated, red indicates hypermethylated). The edge color of red indicates significant differential co-expression in MS NAWM for the connected genes.

Conclusion

In conclusion, we performed an integrative network-based analysis of genetic variants and differential DNA methylation in brain tissue of MS NAWM, which yielded top network genes enriched for relevant MS mechanisms. Our top modules were enriched in functions including immune-related pathways, such as interferon and interleukin signaling pathways, providing insights on autoimmune reactions in MS. Importantly, our networks were also enriched with drug targets of MS FDA-approved medications, which provides insights about drug target mechanisms in MS. We reveal evidence of hypermethylated/hypomethylated drug target genes in MS, suggesting epigenetic mechanisms take part in drug target mechanisms. Genes present in our top networks are of importance to further investigate for deeper understanding of MS therapeutic mechanisms.

Acknowledgement

Astrid M. Manuel is supported by a training fellowship from the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics & Data Science (T15LM007093).

References

1. Manuel AM, Dai Y, Freeman LA, Jia P, Zhao Z. Dense module searching for gene networks associated with multiple sclerosis. BMC Med Genomics. 2020;13(Suppl 5):48.
2. Patsopoulos NA, Baranzini SE, Santaniello A, Shoostari P, Cotsapas C, Wong G, et al. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science. 2019;365(6460):eaav7188.
3. Huynh JL, Garg P, Thin TH, Yoo S, Dutta R, Trapp BD, et al. Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. Nat Neurosci. 2014;17(1), 121–130.
4. Wang Q, Yu H, Zhao Z, Jia P. EW-dmGWAS: Edge-weighted dense module search for genome-wide association studies and gene expression profiles. Bioinformatics. 2015;31(15):2591–4.
5. van der Poel M, Ulas T, Mizee MR, Hsiao CC, Miedema SSM, Adelia, et al. Transcriptional profiling of human microglia reveals grey–white matter heterogeneity and multiple sclerosis-associated changes. Nat Commun. 2019; 10(1):1139.

Drug-wide association studies of cancer using real-world health data

Rachel D. Melamed, PhD[1]
[1]University of Massachusetts, Lowell, Lowell, MA

Abstract

Certain common drugs have been shown to have a side effect of increasing or decreasing the user's risk of some cancers. But, discovering such effects in human populations requires either large scale clinical trials, or careful analysis of observational data. In this work, we develop a method that uses observational health claims data to test the effect of each common drug on cancer risk.

Introduction

Cancers result from mutations accumulated over a lifetime, but onset is influenced by factors including environment and medical history. For example, insulin dysregulation induced by diabetes is thought to promote tumor growth, and metformin, a diabetes drug, is under investigation for its antitumor properties. Discovering drugs that influence cancer onset is an impactful topic both for cancer prevention and possible therapy. In particular, repurposing drugs and discovering effective drug combinations is an intensive area of research investment. Since drugs taken for other purposes may impact pre-cancerous tissue, we hypothesize that this will manifest in altered rates of cancer among treated individuals. To estimate the effect of a drug on patient cancer outcomes, we re-use observational health claims data covering half of the USA population, with a method inspired by the cohort study design.

Cohort studies compare the disease rates in people exposed to the drug against rates in a similar cohort of people who never were exposed. However, one of these cohorts may be older, sicker, or otherwise at greater risk of cancer; these differences can confound effect estimates. Cohort studies typically are limited by the need for medical experts to curate confounding health factors that influence exposure or outcome. To conduct a drug-wide association study for all commonly used drugs, we develop a way to systematically substitute for expert knowledge and scale up the cohort study.

Methods

We use the IBM MarketScan claims data, containing histories of 150 million people, including coded prescriptions, diagnoses, and procedures. Our method automatically conducts many cohort studies to create multiple estimates of the effect of each common drug on each type of common cancer. Cohort studies typically start by defining two cohorts of patients, a treated and a comparator cohort. The comparator cohort is often defined as the set of people who took a similar drug. To determine the cancer effect of each common drug (hereafter referred to as "treatment"), our method automatically identifies a set of comparator drugs for each treatment. The comparator drugs have the same therapeutic use as the treatment, as defined in the MarketScan RED BOOK supplement, and they are additionally filtered on a previously described score of similarity to the treatment[1]. Each comparator drug defines one possible cohort study, and for each of these studies our method proceeds as outlined in Figure 1. In Figure 1, Step 2, all patients who took either the treatment or the comparator, are extracted. Any patients who have a history of cancer before receiving their drug are removed, to avoid reverse causal effects of cancer on rates of treatment. Thus, in Step 3, the potential cohorts consist of all patients receiving one of the two drugs, with no history of cancer. Patients are matched on their histories in the time before receiving their drug, using high-dimensional propensity scores, where covariates consist of all drugs, procedures, and diagnoses in the patient's history. Since this procedure accounts for confounding medical history, we

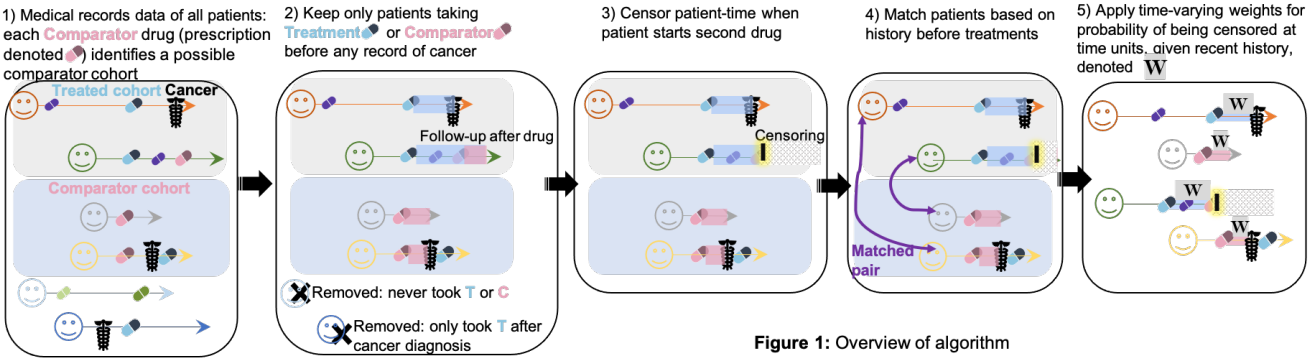


Figure 1: Overview of algorithm

could compare rates of cancer between the matched cohorts created in Step 3, but one complication is that people in the treated cohort may later initiate taking the comparator drug, or vice versa. An analysis that does not account for this is called an "intention to treat" analysis, and the effect estimates could be significantly biased toward the null[2]. Instead, we artificially censor patients at the time when they take the opposite drug (Figure 1, Step 4), which is called a "per protocol" analysis[2]. We perform a Cox regression in the matched populations to estimate risk of each common cancer due to the treatment, and we account for the artificial censoring by weighting by the inverse-probability of censoring[3].

The resulting effect estimates are, in isolation, vulnerable to bias from unmeasured confounders. However, if we find a similar drug effect across multiple types of cancer, or similar effects across drugs with the same active ingredient, we hypothesize that that effect is not likely due to confounding but rather is a true drug effect. We implement a Bayesian meta-analysis to share information across many effect estimates, similar to Shahn, et al[4]. To assess the success of our method, we use positive controls (drugs known to have an effect on cancer) and negative controls (most currently prescribed drugs should have no effect on cancer or on other common diseases).

Results

In support of our method, the distribution of drug effects on cancer is tightly centered around the null. This finding is in line with the expectation that most drugs do not impact cancer. This result suggests our method is generally robust to bias induced by confounding. As well, among the drugs with a preventive effect on cancer, we are able to reproduce the known effect of metformin, supporting the power of our method to identify repurposeable drugs.

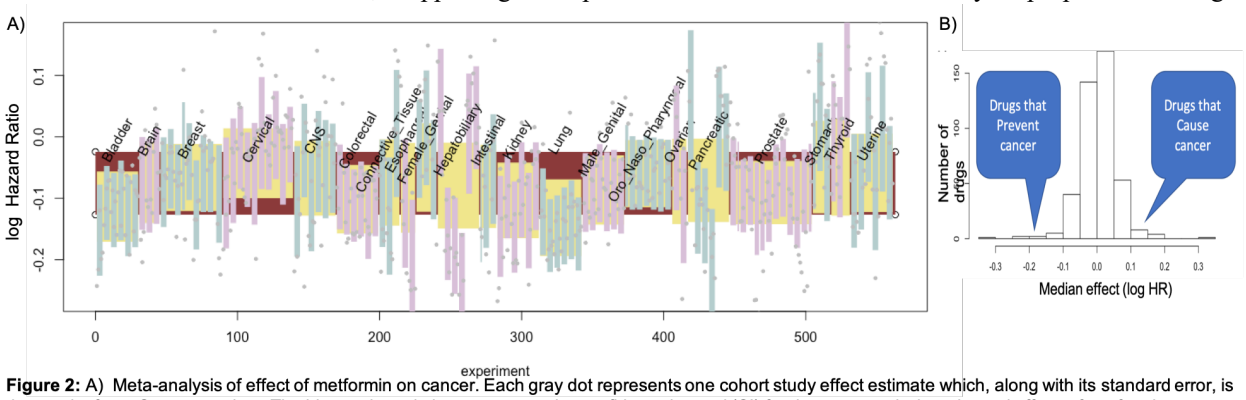


Figure 2: A) Meta-analysis of effect of metformin on cancer. Each gray dot represents one cohort study effect estimate which, along with its standard error, is the result of one Cox regression. The blue and purple bars represent the confidence interval (CI) for the meta-analysis estimated effect of metformin on one type of cancer, as measured using one particular control (ie,metformin vs glimeperide). The yellow bars represent the same effect, meta-analyzed across all controls. The red bar indicates the CI for the effect of metformin on risk of any cancer. B) Distribution of estimated effects on cancer for each common drug.

Conclusion

This study merges methods from epidemiology with approaches from data science, providing a unique big-data enabled survey of candidate cancer-relevant drugs. We are able to reproduce known drugs that affect cancer risk, and most drug effects are null. We are currently extending the method to perform a drug-combination wide survey of effects on cancer for thousands of drug combinations with a sizeable user population.

References

1. Melamed RD. Using indication embeddings to represent patient health for drug safety studies. JAMIA Open. In press. Available from https://www.biorxiv.org/content/10.1101/737049v3
2. Danaei G, Rodríguez LAG, Cantero OF, Logan R, Hernán MA. Observational data for comparative effectiveness research: An emulation of randomised trials of statins and primary prevention of coronary heart disease. Stat Methods Med Res [Internet]. 2013 Feb 1;22(1):70–96. Available from: http://dx.doi.org/10.1177/0962280211403603
3. Hernán MA, Robins JM. Causal Inference [Internet]. Available from: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2017/05/hernanrobins_v3.20.2.pdf
4. Shahn Z, Li Y, Sun Z, Mohan A, Sampaio C, Hu J. G-Computation and Hierarchical Models for Estimating Multiple Causal Effects From Observational Disease Registries With Irregular Visits. AMIA Jt Summits Transl Sci Proc [Internet]. 2019 May 6;2019:789–98. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6568089

COVID-19 Diagnostic Testing Prediction Using Natural Language Processing to Power a Data-Driven Symptom Checker

Stéphane M. Meystre, MD, PhD[1], Paul M. Heider, PhD[1], Youngjun Kim, PhD[1],
[1] Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC

Introduction: The Coronavirus Disease 2019 (COVID-19) is caused by the SARS-CoV-2 virus and was declared a public health emergency of international concern on January 30, 2020 and a pandemic on March 11, 2020. The first COVID-19 case in the U.S. was confirmed January 21, 2020. A rapid expansion to all 50 U.S. states followed, with about 5.2 million confirmed cases and more than 165,000 deaths as of August 12, 2020.[1] To help assess the extent of the pandemic, characteristics of the virus and the disease it is causing, discover and compare supportive or therapeutic approaches and population health measures applied at the patient level, detailed clinical information is needed. This detailed information is typically found in unstructured text notes from EHR systems or other ancillary systems. Extracting such information manually is costly, not scalable, and far too slow to address current needs.

At the Medical University of South Carolina (MUSC, Charleston, SC), a telehealth system (Zipnosis[2]) was advertised as the preferred option for local patients interested in COVID-19 testing. Patients would start a virtual visit, indicate their symptoms, COVID-19 exposure and travel history, and brief medical history. After this virtual visit was completed, the telehealth system would export a natural language generated summary text note based on the information entered by the patient.[3] This text note was the only information subsequently available in the EHR and other clinical systems. Care management based on some COVID-19 dashboard or decision support capabilities were required, but the unstructured text format of this note made them difficult if not impossible. As an effective and scalable approach to extract structured and coded information from unstructured text, natural language processing (NLP) has been used for many years now,[4] demonstrating successful applications to support clinical data reuse for research applications,[5,6] clinical care[7,8] and healthcare management[9] in general.

To help assess the local extent of the pandemic and support patient care as well as research, a new database was created at MUSC in March 2020, along with an NLP-based COVID-19 information extraction tool enriching this database. These efforts, as well as uses of this COVID-19 information for testing results prediction, are presented below.

Methods: A new database (*COVID data mart*) was created to collect clinical information from patients tested or treated for COVID-19 at MUSC. To enable access to structured and coded COVID-19 related information as documented by patients in the telehealth system, a new NLP application (*COVID-NLP tool*) was developed.[10] It focused on extracting information from the notes generated by the telehealth system. This application, the aforementioned data mart and all related data extraction, transfer and loading were developed, tested and made available for production in about ten days only, in March 2020. The COVID data mart was progressively enriched with information extracted from the telehealth system and combined with select clinical information from existing patient records at MUSC (demographics, coded diagnoses and procedures, observations, laboratory test results including SARS-CoV-2, medications, admission-discharge-transfer information). It included clinical information from 169,367 patients as of August 12, 2020. The clinical information collected in the COVID data mart has already been used to drive a real-time COVID-19 dashboard informing MUSC healthcare providers and for several research studies. This information included SARS-CoV-2 diagnostics test results and early success [11] with using information from the telehealth system to predict positive SARS-CoV-2 results encouraged further efforts to enhance the accuracy of these predictions and enable applications supporting patient care such as a novel data-driven COVID-19 symptom checker giving patients testing advice according to their predicted test result.

The initial features used for prediction included age, a selection of 23 symptoms, COVID-19 exposure and travel history, smoking status, pregnancy status and whether patients were healthcare workers. The outcome was the SARS-CoV-2 diagnostic test result (positive or negative). Information extracted from all telehealth system notes for a given patient with

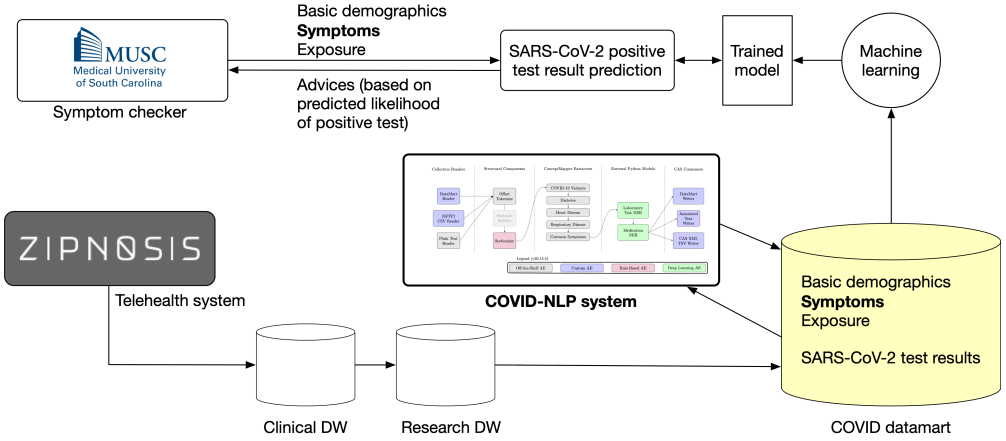


Figure 1: Overall COVID data mart, NLP tool and testing prediction architecture.

corresponding diagnostic test results was used (34,597 telehealth notes from 14,055 patients seen between April 19 and June 24, 2020, 1,101 testing positive and 12,954 negative). We used all positive cases and downsampled to ten percent of negative cases. We used 10-fold cross-validation to verify the integrety of training. We experimented with a variety of machine learning algorithms: decision trees, support vector machines (SVM), logistic regression, neural networks (multilayer perceptron), fastText[12] and deep neural networks (convolutional neural networks).
The final symptom checker was designed to return three possible levels of risk for an individual: low risk (recommending no action), medium risk (recommending caution), and high risk (recommending immediate medical action). We treated the cut-off determination for these thresholds as a seceondary parameter-tuning task after training the models. For evaluation, we treated the low-to-medium threshold as the boundary between true negative (TN) results and true positive (TP) results. Our objective was to maximize sensitivity (i.e., recall) with a value as close to 95% as possible. Results reported below are evaluated with this objective in mind.
As a second point of comparison, we evaluated the models against a curated dataset of 125 positive cases and 242 control (negative) cases who had been interviewed with regards to their symptoms and other COVID-19 related details at MUSC. This population was medically evaluated for different reasons than those seeing medical care through Zipnosis.

Results: When comparing the predicted SARS-CoV-2 diagnostic test result with the aforementioned reference standard, we found the simpler models tended to outperform the more complex models. Specifically, we focused on SVM and logistic regression models in the later stages of this work because decision trees, neural networks, and deep neural networks did not reliably beat the former models. The simpler models also had clearer explanatory power for the subject matter experts we conferred with during development. The best logistic regression evaluated with our telehealth data had a recall of 0.9514, specificity of 0.1244, and negative predictive value (NPV) of 0.9286. When evaluated against the curated dataset, recall and NPV reached 0.952 and 0.8286 respectively while specificity dropped slightly to 0.1198. The best SVM had a recall of 0.9509, specificity of 0.1108, and NPV of 0.9198. Against the curated dataset, recall and NPV rose to 0.952 and 0.9545 respectively while specificity rose to 0.5207.

Discussion: Our current production model has eliminated certain original features due to unintuitive performance or poor upstream data. For instance, ages were binned into three categories: <18, 18-64, and 65+. Young patients (<18) using Zipnosis were largely sicker than older patients (65+), who were more likely to seek routine care. This implicit sampling bias resulted in young patients receiving a higher risk score when all else was held equal. The curated dataset included only a subset of the features our models were trained on. This difference probably caused the accuracy variations observed with the curated dataset. We have so far ignored temporal considerations in our modeling. Future work will need to address changes in positivity rates at the population level over time and monitor model drift, for instance.

Acknowledgments: This work was supported in part by the SmartState endowment. We thank Katie Kirchoff for her work building and maintaining the COVID data mart, Matthew Davis and Rachel McNeely for their insights about modeling COVID-19 risks and Dr. Scott Curry for offering access to the curated dataset.

References

1. Johns Hopkins University Center for Systems Science and Engineering (CSSE). COVID-19 Dashboard. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
2. Zipnosis. https://www.zipnosis.com/our-solution/
3. Ford D, Harvey J, McElligott J, et al. Leveraging Health System Telehealth and Informatics Infrastructure to Create a Continuum of Services for COVID-19 Screening, Testing, and Treatment. *J Am Med Inform Assoc*. 2020.
4. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*. 2008:128–144.
5. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc*. 2016;23(5):1007-1015.
6. Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *J Am Med Inform Assoc*. 2013;20(e2):e341-348.
7. Meystre SM, Kim Y, Gobbel GT, et al. Congestive heart failure information extraction framework for automated treatment performance measures assessment. *J Am Med Inform Assoc*. 2017;24(e1):e40-e46.
8. Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. *International journal of medical informatics*. 2008;77(9):602–612.
9. Nguyen AN, Truran D, Kemp M, et al. Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. *AMIA Annu Symp Proc*. 2018:807-816.
10. Meystre S, Kim Y, Heider P. COVID-19 Information Extraction Rapid Deployment Using Natural Language Processing and Machine Learning. *AMIA NLP WG Pre-Symposium*. ; 2020 (submitted).
11. Obeid JS, Davis M, Turner M, Meystre SM, Heider P, Lenert L. An AI approach to COVID-19 infection risk assessment in virtual visits: a case report. *J Am Med Inform Assoc*. 2020.
12. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. *arXiv*. Published online August 9, 2016. http://arxiv.org/abs/1607.01759

Applying the Extension for Community Health Care Outcomes (ECHO) Model for Cancer Prevention and Survivorship Care: Formative Evaluation of Program Design

Zheng Z. Milgrom MD[1,2], Tyler S. Severance MD[3,4], Caitlin M. Scanlon MSW[3], Anyé T. Carson MPH[2], Andrea D. Janota MPH[2], Terry A. Vik MD[3,4], Joan M. Duwve MD[2], Brian E. Dixon MPA, PhD[1,2], Eneida A. Mendonca MD, PhD[1,4]

[1]Regenstrief Institute, Indianapolis, Indiana; [2]Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana; [3] Riley Hospital for Children, Indianapolis, Indiana. [4]Indiana University School of Medicine, Indianapolis, Indiana

Introduction

Extension for Community Health Care Outcomes (ECHO) is a tele-mentoring model for continuing medical education (CME) to connect experts (as the "Hub") at academic centers with health professionals (as "Spokes") in the community. "Spoke" primary care providers (PCP) from various locations participate at regularly scheduled times through a videoconferencing platform via internet or telephone. The "Hub" experts facilitate and guide the "Spokes" participants through two components in tele-ECHO clinics: Hub-led didactics on the curriculum topics, and Spoke-led case presentations on de-identified patient cases. The first ECHO program was launched at the University of New Mexico in 2004, focusing on Hepatitis C care. It successfully demonstrated that Hepatitis C care delivered by ECHO-trained rural physicians was equally effective as that given at the University of New Mexico.[1] Since then, the ECHO model has spread to 39 countries with 845 programs focusing on over 70 topics, among which 135 programs are cancer related.[2] However, little evidence demonstrates the effectiveness of the ECHO model in addressing a wide range of topics regarding cancer prevention and survivorship. In September 2019, the Cancer Prevention and Survivorship Care tele-ECHO program (Cancer ECHO), launched as a pilot program at Indiana University (IU) Fairbanks School of Public Health and in partnership with IU Health. The purpose of this study is to examine the ECHO model on cancer prevention and survivorship by conducting a formative assessment of the Cancer ECHO program.

Methods

This study employed a mixed-methods approach, including quantitative administrative data of IU ECHO programs and qualitative data from semi-structured interviews (N=21). Study participants were selected based on purposeful sampling (with approval of the IU Institutional Review Board) and recruited dynamically until data saturation. Data were collected across the Hub team (including IU ECHO leadership), Spokes members who have at least attended Cancer tele-ECHO clinics once, and Potential Spoke (PS) care providers from the attendees of other PCP-targeting ECHOs facilitated by IU. By including PS who participated in other IU ECHOs, we sought to explain PCP's adoption and explore the program design specifically on the context of cancer prevention and survivorship. We also sought to cross-validate our findings through different points of view. We assessed perspectives on the strengths and weaknesses of the program design and their participation decisions. PCP beliefs about the intervention and their suggestions were also examined. The interviews were recorded, then transcribed with NVivo machine-transcription services pairing with manual audits for accuracy. The themes of the transcribed interview recordings were coded with an iterative approach using NVivo 12. Two research team members (Z.M. and E.M.) met regularly to discuss the emerging themes, and the codes were reorganized until the code consistency was achieved.

Results

During the initial pilot year, Cancer ECHO possessed lower PCP participation (N=27 unique individual, 18% of all participating individuals) than other IU ECHOs that targeted PCPs. Three Hub team members, 11 Spoke members (6 care providers and 5 non-providers) and 7 Potential Spoke (PS) care providers were interviewed. The majority of the non-providers were health educators, care navigators, public health workers, or administrators. 6 Spoke members and all 7 of the PSs have experiences with other IU ECHOs. Table 1 summarizes the emerged themes in the interviews.

Discussion

The ECHO model was developed to democratize knowledge among health professionals in medically isolated communities. The pandemic-catalyzed rapid expansion of telehealth usage imposed the importance of virtual

Table 1. Salient Themes

	The components and strengths of Cancer ECHO	**Weaknesses of Cancer ECHO**	**Suggestions**
Format (within sessions of tele-ECHO clinics)	Livestreaming didactics and case presentations; Strength: case presentation is much more engaging than the didactics component.	The conflict of the benefits of livestreaming with their availability.	Decrease the length, change the time, provide asynchronous participation.
Content (across sessions)	Topics on both of cancer prevention and survivorship; Strength: discussions on real-life experiences.	Lack of consistency between the content covered in the didactics and case discussion; Topics in different sessions are too varied.	Adjust the format and curriculum: narrowing down the topics, merging the didactics with case discussion, focusing more on clinical workflow improvement.
Participant and community	Wide range of participants; Strengths: a nonjudgmental, safe learning environment, and support from an interdisciplinary community.	Attracting and involving appropriate participants, especially care providers in the Cancer ECHOs.	Assess care providers' needs, improve the participants' engagement and the targeted audiences' awareness.

collaboration in clinical care and education. Compared to the original Hepatitis C ECHO, the spectrum of cancer prevention and survivorship is broad, and prevention was part of PCP's daily work. Thus, the contents, participant types, and their values placed on the intervention are different. This study helps us understand the strengths and weaknesses of the ECHO model while applying it to the context of cancer prevention and survivorship. The Cancer ECHO reserves the components of a nonjudgmental, safe learning environment, discussions on real-life experiences, and support from an interdisciplinary community in the original ECHO model and demonstrates its potential. There are challenges maintaining the engagement of a wide range of participants and focusing on a broad range of topics.

Conclusion

While discussions in an interdisciplinary community are highly valued by the participants, the Cancer ECHO still needs to focus on the needs of the targeted audience, PCPs. Program adjustments accommodating care providers' needs and aligning with their attitudes towards the intervention are needed to improve participation and experiences. We suggest a care provider-driven or provider-focused approach by including participating providers, their supervisors, and other stakeholders early in the operational, content, and marketing designs.

Acknowledgement

This work was supported by a grant from the U.S. National Library of Medicine [Grant number: T15LM012502] and Indiana State Department of Health. The views expressed in this publication are those of the authors and do not necessarily reflect the position or policy of the ISDH, National Library of Medicine, or the United States government.

References

1. Arora S, Kalishman S, Thornton K, Dion D, Murata G, Deming P, et al. Expanding access to hepatitis C virus treatment--Extension for Community Healthcare Outcomes (ECHO) project: disruptive innovation in specialty care. Hepatology. 2010 Sep;52(3):1124–1133.
2. Arora S, Byers EL. Leveraging local expertise to improve rural cancer care outcomes using project ECHO: A response to levit et al. JCO Oncology Practice. 2020 Jul;16(7):399–403.

Bridging the Gaps between Evidence and Practice of Assessing Risks and Addressing Harms for Adults on Chronic Opioid Therapy

Meenakshi Mishra, M.Sc., MPH[1], Nicole G. Weiskopf, Ph.D.[1]
[1]Oregon Health & Science University, Portland, OR, USA

Introduction

Prescription opioids have contributed in part to the current opioid epidemic. The CDC guidelines for prescribing opioids for chronic pain recommends that providers weigh risks versus benefits when considering opioids for their patients. Primary Care Providers (PCPs) prescribe about half of all prescription opioids and manage most chronic pain patients.[1] Making treatment decisions for patients with chronic pain is complex, with goals to treat the pain but prevent addiction to opioids. Several biological, psychological, and social (biopsychosocial) factors contribute to opioid misuse and Opioid Use Disorder (OUD) in patients on Chronic Opioid Therapy (COT) for pain.[2] Clinical decision-making when prescribing opioids requires integrating disparate data sources, resulting in the cognitive burden of PCPs. We present a systems view of the prescription opioids in the primary care setting, limitations of current systems for assessing prescription opioid-related risks and make a case for developing a generalized approach to understanding and addressing the information needs of PCPs at the point of care.

Methods

A literature review was performed to develop a systems view of the prescription opioids for chronic pain in the primary care setting.

Discussion

Systems view of prescription opioids in the primary care setting

The PCP, when considering opioid prescription for chronic pain, pain lasting >90 days or beyond the time of normal tissue healing, must contextualize the evidence-based guidelines with patient-specific data to make evidence-based, patient-centered decisions. Figure-1 is a conceptual model of the opioid prescription for chronic pain in the primary care setting. However, there are several limitations to the various tools and information systems.

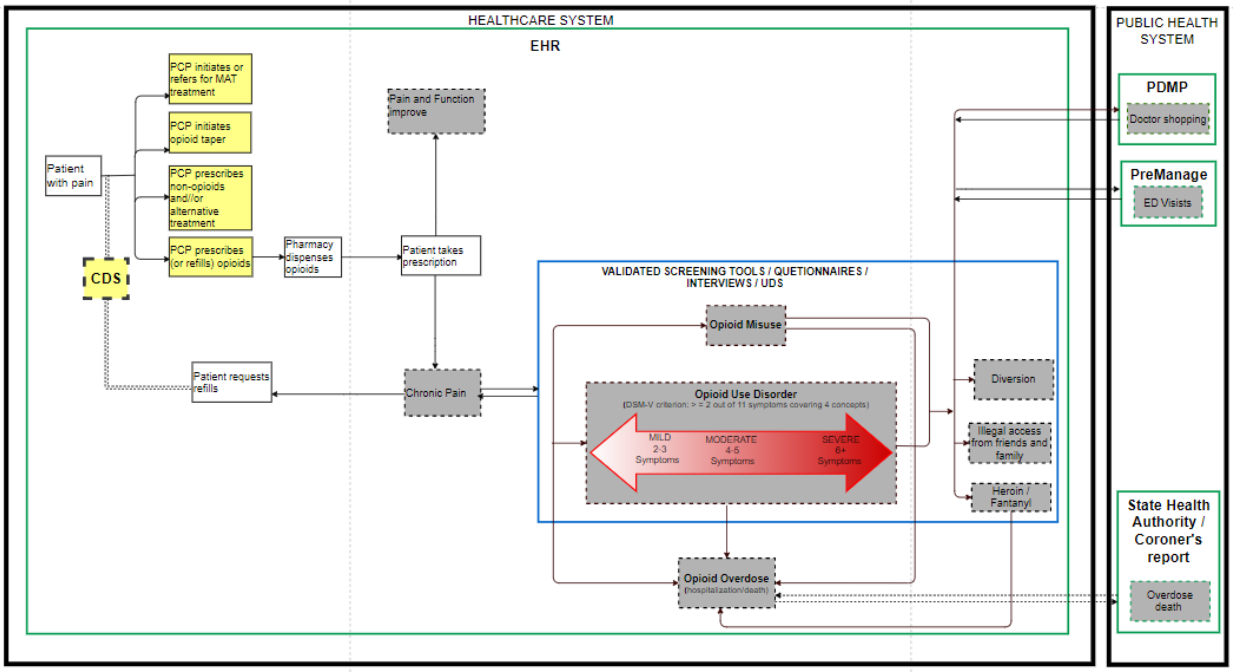


Figure-1. Systems view - Interplays between clinical tools, processes, and information systems in the primary care setting. (PDMP – Prescription Drug Monitoring Program, a state-run program to collect information of dispensing of controlled substances; PreManage – a collective ambulatory platform that can receive data from Emergency Department Information Exchange (EDIE); CDS – Clinical Decision Support; ED – Emergency Department; EHR –

Electronic Health Record). Greyed boxes represent outcomes; yellow boxes represent clinical decisions; Green outlines the various information systems; and blue outlines the multiple tools to determine opioid-related risks. The systems view spans both healthcare and public health systems.

Current limitations

Opioid guidelines: There are many reasons for non-adherence to CDC guidelines. A recent systematic review found many provider and guideline specific characteristics that make adherence to evidence-based practice challenging.[3] These include: non-familiarity with guidelines; guidelines being overly simplified and lacking credibility with difficulty in weighing risk vs. benefit; delay in delivery and intensification of treatment due to cognitive biases - overestimation of risk based on "gut feeling,"; belief that risk management practices are similar to policing patients and has no place in the delivery of healthcare; certain guideline expectation from providers, such as obtaining Opioid Treatment Agreement from patients, are viewed as unfavorable for patient-provider trust. While recent efforts to educate PCPs have led to increased awareness and willingness to identify and address risks from prescription opioids, there is a gap between the knowledge and actual practice.

EHR: EHR data offers a tremendous opportunity to identify patients at risk and guide patient management.[4, 5]. However, risk identification and management are challenging due to variations in documentation practices, difficulties identifying problem opioid use, and inconsistent use of problem opioid use terminologies.[6] Most providers are reluctant to clearly and unequivocally document problem opioid use in the patient's chart. The potential for patient stigmatization also hampers clear documentation of problem opioid use.

PDMP: There are limitations to the accuracy, accessibility, and interpretability of the PDMP data.[7] The states' PDMP programs operate under different regulatory agencies, collect different types of data, require data to be updated at different time intervals, and restrict access to people with specific roles. The objective criteria to determine drug-seeking behavior includes >= 4 opioid prescriptions from >= 4 providers. However, "many patients have multiple prescribers because of poor primary care access, visits to emergency departments for acute exacerbations of pain, and conditions requiring visits to multiple specialists." There is a need to determine what data values in PDMP should prompt intervention from the physician when considered alongside the patient's complete clinical encounter.[7]

OUD diagnostic criteria and validated screening tools: There are limitations to applying the DSM-V diagnostic criteria for OUD in chronic pain patients where most meet the symptoms of opioid dependence due to long term prescription opioid use. Moreover, there is no standard screening tool for determining misuse and OUD, and many providers are reluctant to use them for fear of losing patient trust.

Conclusion

There is a need for CDS that can contextualize evidence-based guidelines with patient-specific characteristics to promote evidence-based and patient-centered decisions. Present information systems and screening tools are fraught with many limitations. In the absence of standard clinical tools and processes and inadequate information systems to determine prescription opioid-related risks in this complex patient population in a busy primary care setting, a systematic and generalized approach is needed to determine and address the information needs of PCPs at the point of care. Such an approach will inform future CDS projects. Our future work will address this need to bridge the gap between evidence and practice

References

1. Levy B, Paulozzi L, Mack KA, Jones CM. Trends in Opioid Analgesic-Prescribing Rates by Specialty, U.S., 2007-2012. American journal of preventive medicine. 2015;49(3):409-13.
2. Wiss DA. A Biopsychosocial Overview of the Opioid Crisis: Considering Nutrition and Gastrointestinal Health. Front Public Health. 2019;7:193-.
3. Hossain MA, Asamoah-Boaheng M, Badejo OA, Bell LV, Buckley N, Busse JW, et al. Prescriber adherence to guidelines for chronic noncancer pain management with opioids: Systematic review and meta-analysis. Health Psychol. 2020;39(5):430-51.
4. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak. 2017;17(1):24.
5. Jonnagaddala J, Liaw S-T, Ray P, Kumar M, Chang N-W, Dai H-J. Coronary artery disease risk assessment from unstructured electronic health records using text mining. Journal of biomedical informatics. 2015;58 Suppl(Suppl):S203-S10.
6. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. International Journal of Medical Informatics. 2015;84:1057-64.
7. Griggs C, Weiner S, Feldman J. Prescription Drug Monitoring Programs: Examining Limitations and Future Approaches. Western Journal of Emergency Medicine. 2015;16(1):67-70.

Machine Learning Approaches for Predicting Diabetes-Related Long-Term Complications Using Real-World EHR Data

Abu S M Mosa, PhD[1,2,3,4], Chalermpon Thongmotai, BS[2,4], Humayera Islam, MS[3,4,1], KSM Tozammel Hossain, PhD[2,3], Vasanthi Mandhadi, MS[4,5]

[1]Department of Health Management and Informatics; [2]Department of Electrical Engineering and Computer Science; [3]Institute for Data Science and Informatics; [4]Center for Biomedical Informatics, [5]University of Missouri School of Medicine, Columbia, MO

Introduction

Diabetes is a major health issue that affects an increasing amount of people each year. By 2030, approximately 358 million people will have diabetes in the United States[1]. Diabetes also increases the risk for many long-term severe complications such as diabetic-related eye diseases, kidney diseases, and neuropathy. The biggest concern about diabetes-related complications is that they are unrecognized in the early stages. There are little to no symptoms in the early stages, but they can be immutable and devastating as people grow older. However, not all people develop long-term complications. As such, the prediction of diabetes-related diseases at an early stage can help intervene in preventative care. Many types of research were conducted to study diabetes-related complications. However, none of them used a data-driven approach to study diabetes-related complications. The objective of this abstract is to present a data-driven approach for predicting diabetes-related diseases using comorbid diagnosis data from real-world electronic health records (EHR) of 90 health systems across the United States. In this study, we created machine learning-based prediction models for three diabetes-related diseases: (a) eye diseases (ED), (b) kidney diseases (KD), and (c) neuropathy (NP).

Methods

Data Source, Inclusion, and Exclusion Criteria: We used Cerner's "Health Facts EMR Data", a de-identified EHR data from about 90 health systems across the United States. The database contains 69 million unique patients. We used the SUPREME-DM[2] algorithm for identifying the diabetes population from the EHR data. We created six criteria based on the lab results using HbA1c, fasting plasma glucose and random plasma glucose. Diabetes-related ICD9 and ICD10 diagnosis codes (250.x, 357.2, 366.41, 362.01–362.07, and E08.x-E13.x) were used with inpatient and outpatient encounters. In total, we had eight criteria, and we included only the adult population (age 18 or above at the first encounter of diabetes based on SUPRE-DM criteria).

We identified diabetes-related complications among the selected diabetes population by using ICD9 and ICD10 diagnosis codes, including diabetic ED(250.5x, E10.x, and E11.x), diabetic KD(250.4x, E10.x, and E11.x), and diabetic NP(250.6x, E10.x, and E11.x). We computed the duration and number of encounters between (a) the first encounter of diabetes and (b) the first encounter of diabetes-related diseases or the latest encounter (for people not having any diabetes-related diseases). **Figure *1*** presents our data collection window. Since we are modeling long-term complications, we set four additional population inclusion criteria to ensure the availability of longitudinal encounters: (a) for the diabetes patients having diabetes-related complications, having a minimum of 5 years of encounters starting from the first encounter of diabetes, (b) for the diabetes patients having no diabetes-related complications, having a minimum of 7 years of encounters from the first encounter of diabetes, (c) having at least 25 encounters during the study period, and (d) having at most 500 encounters during the study period to ensure removal of non-patient records.

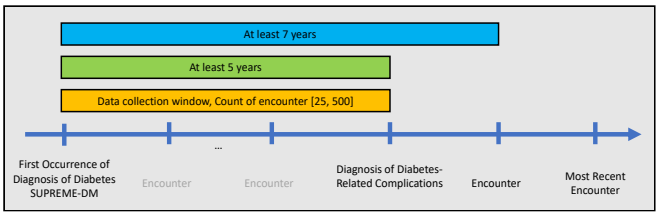


Figure 1. Longitudinal Inclusion Criteria

Data Preprocessing and Machine Learning Approach: ICD9 and ICD10 diagnosis codes were mapped into the 285 Clinical Classification Software (CCS) codes. The CCS groups individual diagnosis codes into clinically similar entities, which allows for analysis of broad categories of diagnoses and avoids model overfitting.

In this study, we conducted two experiments. First, the association between diabetes-related and potential comorbid diagnosis was assessed by odds ratios (ORs). Second, we created seven predictive models: (a) Decision Tree with Gini (DT), (b) Random Forest (RF), (c) Logistic Regression (LR), (d) Multilayer Perceptron with a single layer (MLP1), (e) Multilayer Perceptron with double layers (MLP2), (f) Multilayer Perceptron with triple layers (MLP3), and (g) Support Vector Machine (SVM). We also applied various techniques to handle imbalanced datasets such as

Random Undersampling, Random Oversampling, and Oversampling using SMOTE. We used 10-fold cross-validation and evaluated our models with sensitivity, accuracy, specificity, and area under the ROC curve (AUC). Python 3.6 with Pandas, NumPy, Scikit-learn, and Matplotlib were used to conduct all the experiments.

Results

This study takes advantage of the unique opportunity provided by our access to a large and rich EHR dataset. After applying the strict cohort selection criteria, a total of 102,876 patients with diabetes were satisfied, which contains highly imbalanced datasets for diabetic ED patients (4.36%), diabetic KD patients (9.04%), and diabetic NP patients (10.68%). The machine learning results, including sensitivity, accuracy, specificity, ROC curve, and AUC on each disease cohort studies are presented in six figures in **Figure 2**.

For the diabetic ED cohort, SVM attained the maximum accuracy of 94.9%, maximum specificity of 73.1%, and a maximum AUC of 93.2%, while RF achieved the maximum specificity of 98.7%. In the diabetic KD cohort, RF attained maximum accuracy of 89.1% and maximum specificity of 97%, SVM achieved maximum sensitivity of 86.1%, and maximum AUC was obtained from LR. For the diabetic NP cohort, LR attained the maximum accuracy of 83.3% and a maximum AUC of 89.4%, SVM attained maximum sensitivity of 84.8%, and RF attained 95.7% specificity. The results showed that the SVM approach achieved the best overall performance in each disease cohort. Considering the ROC curve and AUC values, the LR has the best measure of separability in this study.

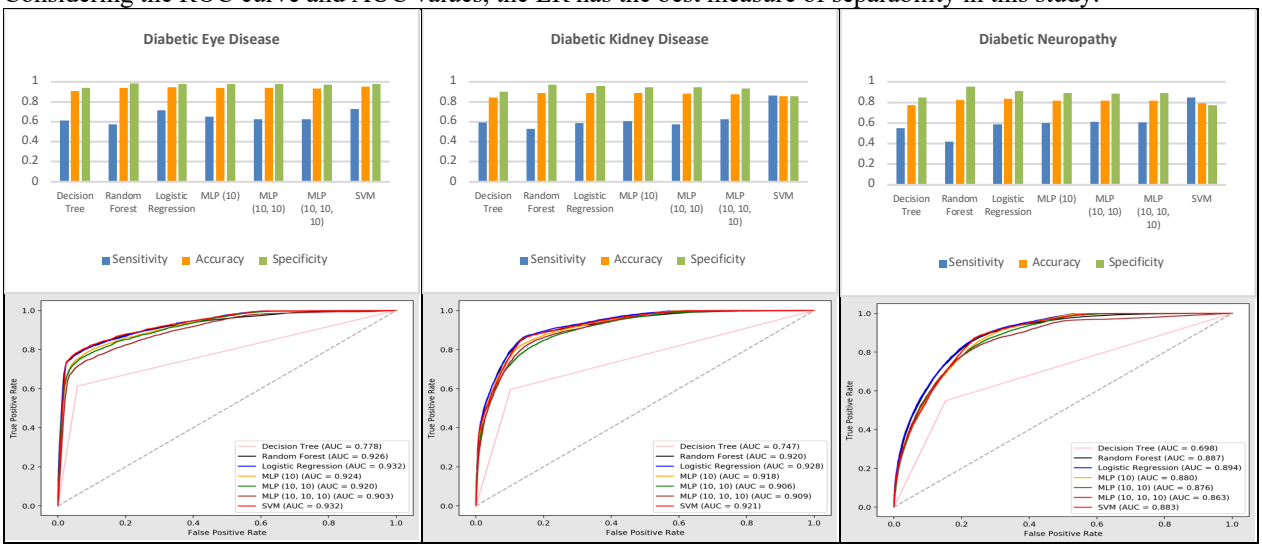


Figure 2. Performance of the Machine Learning Prediction Models

Discussion

In general, machine learning methods can play a vital role in developing clinical decision support systems. However, there are some challenges for using machine learning with large EHR data: data quality, incompleteness, and lack of standardizations may negatively affect building high accuracy predictive models. Despite the challenges, our methods demonstrated that the application of machine learning is promising for predicting long-term outcomes using EHR data. In our study, SVM performed the best with EHR data. However, our models were not designed to utilize the temporal component of EHR data. The longitudinal aspect of the EHR dataset can help predict long-term disease outcomes for diabetes patients. Therefore, our future work will utilize Long Short-Term Memory neural networks and advance time series models like ARIMAX and GARCH to classify, process, and make predictions based on time-series data. We expect that these models would produce better predictions with real-world modeling of the disease progression from its advantages.

References

1. Rowley WR, Bezold C, Arikan Y, Byrne E, Krohe S. Diabetes 2030: Insights from Yesterday, Today, and Future Trends. *Popul Health Manag*. 2017;20(1):6-12. doi:10.1089/pop.2015.0181
2. Nichols GA, Desai J, Lafata JE, et al. Construction of a multisite datalink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: The SUPREME-DM project. *Prev Chronic Dis*. 2012;9(6). doi:10.5888/pcd9.110311

Algorithmic Fairness in Risk Prediction Models for Patients with Opioid Use Disorder

Yoonyoung Park, ScD[1], Moninder Singh[2], PhD, Issa Sylla[1], Elaine Xiao[1], Jianying Hu[2], PhD, Amar Das, MD PhD[1]
[1]IBM Research, Cambridge, MA USA [2]IBM T. J. Watson Research Center, Yorktown Heights, NY USA

Introduction

Increasing use of machine learning (ML) algorithms to guide decision making in healthcare has drawn attention to algorithmic fairness[1]. A recent study demonstrated how an existing commercial algorithm for population health management in fact perpetuated racial disparities in healthcare.[2] Despite the advancement in fairness-aware ML research[3], few healthcare applications exist to date. In this work, we demonstrate an approach to examine data and models for potential bias and mitigate the observed bias in health outcome prediction. We hypothesize a setting in which algorithms are used to allocate treatment resources for patients with opioid use disorder (OUD). Utilization of evidence-based treatment for OUD called medication assisted treatment (MAT) is reported to be uneven across racial groups.[4] We show how bias in data, likely due to systemic inequality and human perceptions, can carry over to machine learning outcomes and how debiasing methods can reduce the bias measured by fairness metrics.

Methods

Using the IBM® MarketScan® Research Databases, we constructed a cohort of Medicaid patients with OUD (2013-2017). Eligible patients were over 10 years old, white or black, not dually eligible, continuously enrolled during one year prior to (baseline) and after the initial OUD diagnosis date (index date), and had one or more OUD diagnoses without cancer or hospice related codes for which use of opioid can be justified, and had not received any MAT prior to index date. We generated a patient-level data set including features such as age, gender, race, diagnoses, and number of emergency room visits. As shown in[2] the choice of outcome greatly impacts the magnitude and direction of bias. In the absence of perfectly measured outcome, we experimented with four binary targets to classify a 'high risk' subcohort with varying likelihood of the presence of bias. These targets were chosen for clinical and public health relevance based on prior literature and characterized placement in the top decile of 1) total healthcare cost, 2) emergency psychiatric inpatient admission cost, 3) outpatient ER visit cost, and 4) the number of psychiatric diagnosis in post-index period. After splitting data into train, validation, and test sets (5:3:2), we trained logistic regression (LR), random forest (RF), and extreme gradient boosted trees (XGB) for each of the four outcomes. We examined the cohort descriptively and assessed the associations between outcomes and race using generalized linear models adjusted for patient-level features.

In this study, positive label is a *favorable outcome* as it leads to receiving treatment. Race is the *protected attribute*, with white being the *privileged* value. We aim to achieve *group fairness*, meaning that white and black race groups will have equal model outcomes. For this definition of fairness we used disparate impact (DI) as a metric, the ratio of predicted favorable outcome prevalence between unprivileged and privileged groups. We applied both reweighing and Prejudice Remover debiasing methods using AIF360 Toolkit.[3] Reweighing modifies the training data by generating weights for (race, label) combinations, while Prejudice Remover adds a regularization term to the objective in logistic regression so that race features have less impact on the outcome prediction. We focus on the results from XGB models and reweighing for clarity and conciseness.

Results

Compared to whites, black patients were older, had greater physical disease burden measured by comorbidity index and ER visit rate, and similar mental disease burden measured by number of psychiatric diagnosis at baseline. White patients were significantly more likely to receive MAT (24% vs. 9%) and to receive it for longer periods. During the post-index period, the probability of having an outpatient ER visit and cost measures (total and ER cost) were higher among blacks. The probability of a recorded overdose event or emergency psychiatric admission was greater among whites.

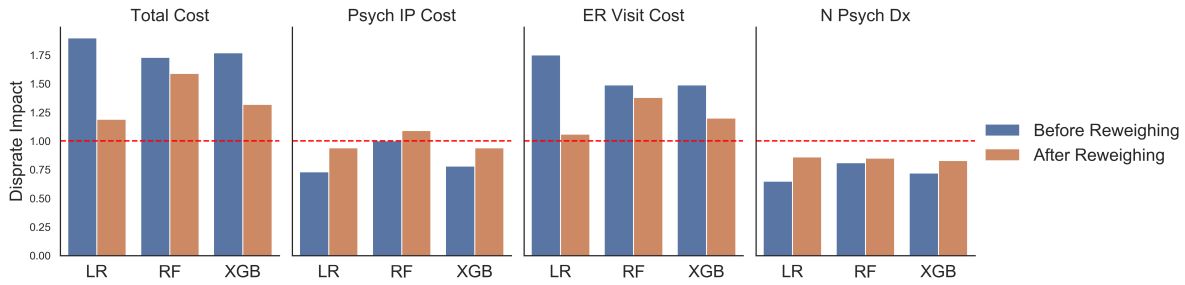


Figure 1: Bias metrics before and after reweighing
Psych IP: emergency psychiatric admission; ER: emergency room; N Psych Dx: number of psychiatric diagnosis

Figure 1 shows the DI values for each target. Before reweighing, the models for emergency psychiatric admission cost or number of psychiatric diagnosis favor white (DI < 1) while the models for total cost or outpatient ER cost favor black (DI > 1). The high risk subcohorts predicted by each model had greater disease burden both physically and mentally and utilized more health services compared to the base cohort as expected. With similar number of psychiatric diagnoses (black vs. white 3.1 vs. 3.1), high-risk black patients had a higher comorbidity index (mean 1.6 vs. 1.0) and incurred more cost (39.7K vs. 27.9K total cost, 2.8K vs. 1.2K ER cost, 9.6K vs. 6.2K psychiatric admission cost) but were less likely to receive MAT (3.5% vs. 15.5%). In the subcohorts based on total cost (and similarly for ER cost), blacks had higher chronic comorbidity (mean 3.4 vs. 2.6) but had a fewer number of recorded psychiatric diagnoses (mean 1.6 vs. 2.2)).

Implementing reweighing methods shifted the DI values closer to one. Debiasing did not negatively impact the balanced accuracy (0.71 0.76 before and 0.72 0.76 after debiasing for the four outcomes). Importantly, we observed that the discrepancies in proportion of patients classified as high risk (i.e. receiving favorable outcome) between whites and blacks decreased with debiasing.

Discussion

Identifying and mitigating biases in healthcare data rife with known and unknown sources of bias is challenging. In this work we demonstrated ways to assess and mitigate bias observed in the data using debiasing methods. That black patients with greater disease burden received less MAT suggests there may be underlying bias in diagnosis or treatment. Varying sources of bias and lack of understanding on true data generating process makes it difficult to have one-size-fits-all solutions. Depending on the label and subsequently on the DI values, follow-up actions to mitigate bias will have opposite impact on patients. We also showed that even with careful selection of targets, the lack of gold standard or unbiased surrogate outcome makes it very difficult to completely avoid bias. Our work highlights the importance of thorough evaluation of bias and efforts to mitigate the bias. The limitation of this work includes the use of DI as a fairness metric. In the future work we will examine other definitions of fairness and also additional debiasing algorithms.

References

1. Rajkomar A, Hardt M, Howell MD, Corrado GS, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2017;169:866–872.
2. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.
3. Bellamy R, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S, Ramamurthy K, Richards J, Saha D, Sattigeri P, Singh M, Varshney KR, Zhang Y. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. 2018 ArXiv, abs/1810.01943.
4. Volkow ND, Jones EB, Einstein EB, Wargo EM. Prevention and Treatment of Opioid Misuse and Addiction: A Review. JAMA Psychiatry. 2019;76(2):208-216.

Towards a Checklist for Data-driven Predictive Models

Panayiotis Petousis*, PhD[1,2], Anders O Garlid*, PhD[2], William Hsu, PhD[1,2], Alex AT Bui, PhD[1,2]
[1]Clinical and Translational Science Institute, UCLA, Los Angeles, CA;
[2]Medical & Imaging Informatics, Dept. of Radiological Sciences, UCLA, Los Angeles, CA

Introduction

In 1935, a deadly crash prompted Boeing to devise and implement flight checklists to reduce the inherent risk of aviation by establishing standard protocols for pilots and removing a degree of human error from the equation.[1] Similarly, developing and deploying data-driven predictive models in healthcare settings carries very real opportunities and risks that directly impact patients, but we lack a set of standard practices to mitigate these risks.

Reproducibility presents a significant challenge in developing, applying, and deploying machine learning (ML) models in healthcare.[2] Manuscripts are frequently published with insufficient information on data preprocessing, model training, and model evaluation. Many of the assumptions and decisions that go into an ML model often go unstated, making it difficult to fully understand its strengths and weaknesses, leaving the context for performance measures unclear. Researchers who have attempted to reproduce a published model can appreciate that data and code sharing is only the start of the process towards deployment in the clinical setting. In the absence of standardized and adopted approaches and processes for model development, implementation, and reporting, these issues will continue to hamper our efforts to reproduce and deploy these models more broadly.[3]

In this work, we review existing studies that propose the use of a model checklist as a means to facilitate reproducibility. We highlight the strengths and limitations of each framework, and propose a community-driven effort to generate a more comprehensive checklist that includes the information needed for reproducing and applying biomedical ML models within real-world healthcare environments.

Methods and Results

We conducted a search for ML model checklists by surveying the research, industry, and biomedical domains; the FDA; and Google Scholar. We reviewed 13 existing model checklists focused on artificial intelligence (AI) and ML modeling research; industry ML modeling; as well as clinical research and deployment of ML models. We evaluated each checklist for inclusion of key components of model development, including defining the problem, scope, and goals; data preprocessing and feature selection; model training, tuning, performance evaluation, comparison, implementation, and monitoring; as well as special considerations when involving patients.

The majority of the model checklists emphasize model description, analytical/technical evaluation, and data engineering (**Figure 1**) with a focus on industry standards for coding and infrastructure, though many neglected to report the target population or to define the problem of interest. Dealing with missing data and their effects on model training were partially addressed by most studies as part of the data engineering processes. However, employing imputation methods or filtering out cases with missing values is a critical step that can introduce bias (e.g., using a dataset derived

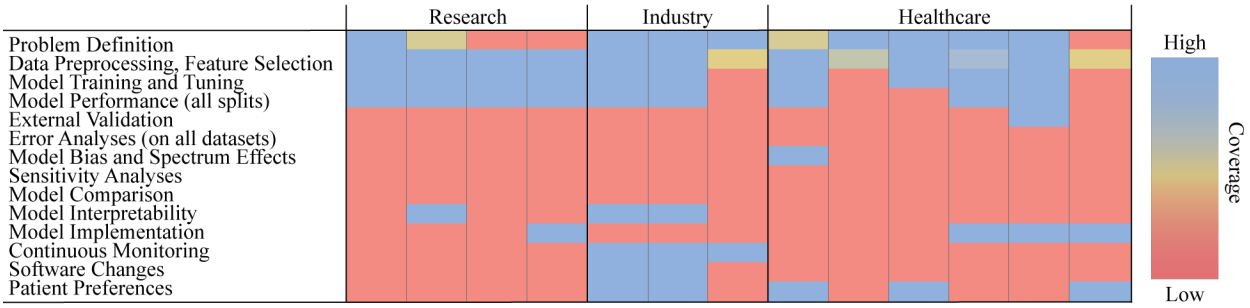


Figure 1: Degree of coverage by category for ML model development in each of 13 reviewed checklists, ordered by primary domain (top row). Gradient: blue/red = high/low coverage (see details at https://aogarlid.github.io/ML_model_checklists/).

from a specific demographic profile to impute missing values) and changes model performance, especially if all the imputation steps cannot be accurately reproduced. Only one checklist mentioned the evaluation of bias, but none investigated bias across population minorities. Model explainability is a requirement for deploying ML models in the clinical setting, but it is mentioned in only two checklists. Several of the frameworks provide details on code sharing and versioning, both of which are necessary to accurately reproduce models across institutions. Additionally, a small number of checklists addressed policies on updating and re-evaluating updated models and software.

Discussion

We systematically evaluated current ML model development checklists reported in the literature to identify an approach or protocol that could be adopted broadly across the modeling communities to facilitate reproducibility and achieve the promise of ML in healthcare and otherwise. Our review reveals significant gaps in the existing checklists and highlights the need for a consensus set of standards for ML model development. Each of the checklists include sections for model descriptions and general information around model evaluation, providing guidance on some of the steps needed to reproduce a published model. Unfortunately, none is comprehensive enough to facilitate successful, end-to-end deployment of such models in real-world clinical settings.

Comprehensive model evaluation entails detailed analyses on cases of misclassification, error tolerance limits, sensitivity analyses, and external validation. Misclassification analysis can provide insights on erroneously labeled cases by models, and, in combination with error tolerance (e.g., confidence intervals), a detailed approach to using ML models for decision making. Sensitivity analyses should be performed on the training, validation, test, and external test sets, and they should be performed dynamically while the model is used in the clinical setting. External model validation (i.e., performance on an external dataset), in combination with test set results, demonstrate true model performance in addition to model transferability. Most checklists lacked sufficient coverage of these evaluation procedures.

The US Federal Drug Agency (FDA) guidance on AI and ML in clinical applications[2] highlights the necessary steps when software changes or updates are applied on an existing model. Notably, such software updates are equivalent to the creation of a new model altogether. Dataset shift must also be considered if a change in the clinical workflow is introduced (e.g., a new diagnostic method). In either case, all evaluation procedures should be repeated whenever changes in the data, environment, and/or model are introduced. Furthermore, models applied in clinical settings should be blindly evaluated against human annotators over random samples to ensure quality of model prediction over time.

We envision a comprehensive checklist arrived at by community consensus that will serve as a guide for model development, evaluation, and detailed reporting. Such a strategy will lead to greater model reproducibility and enable comparison between existing and novel models by establishing a set of standardized evaluation processes and measures. Guidelines on proper approaches to model development and reporting will empower the translation and calibration of models across institutions with different populations and enable greater impact in healthcare. A core checklist will cater to the broadest possible audience and establish the base standards and protocols necessary for any work and reporting on ML models. Given the heterogeneity of biomedical ML model development, we envision an extensible checklist design with modules for specific sub-fields (i.e., mass spectrometry proteomic analysis), allowing customization with different reporting items based on end user application (i.e., model developer vs. clinical user).

Creating a universal Checklist for Data-driven Predictive Models for facilitating model reproducibility in healthcare may best be accomplished with a web-based platform where model developers and end users can share and interrogate models along the criteria outlined in these checklists. Ultimately, this should be considered a community challenge that invites members to extend and use it to report model development, application, and deployment across disciplines.

References

1. Gawande A. The checklist manifesto: how to get things right. New York: Picadur. 2010.
2. US Food and Drug Administration. Proposed regulatory framework for modifications to AI/ML-based software as a medical device (SaMD). https://www.fda.gov/media/122535/download; 2019.
3. National Academies of Sciences, Engineering, and Medicine. Reproducibility and replicability in science. National Academies Press; 2019.

A knowledge graph strategy for integrating clinical and experimental COVID-19 data

Justin T. Reese, PhD[1], Deepak Unni, MS[1], Tiffany J. Callahan, PhD[2], Luca Cappelletti, MS[3], Vida Ravanmehr, PhD[4], Seth Carbon, BA[1], Tommaso Fontana, BS[5], Hannah Blau, PhD[4], Nicolas Matentzoglu, PhD[6], Nomi L. Harris, MS[1], Monica C. Munoz-Torres, PhD[7], Melissa Haendel, PhD[7], Kent Shefchek, MS[7], Peter N. Robinson, MD, PhD[4], Marcin P. Joachimiak, PhD[1], Christopher J. Mungall, PhD[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [2]University of Colorado, Anschutz Medical Campus, Aurora, CO, USA; [3]University of Milano, Milan, Italy; [4]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA; [5]Politecnico di Milano, Milan, Italy; [6]Independent Contractor, London, UK; [7] Oregon State University, Corvallis, OR, USA

Introduction

COVID-19 is a complex disease involving many biological processes and pathways, each of which involves many genes. The research community is still learning about COVID-19; its symptoms and underlying pathological mechanisms are being gradually revealed. Integrated, up-to-date data about SARS-CoV-2 and COVID-19 is crucial for expediting the response to the COVID-19 pandemic by the biomedical research community. Rich biological knowledge exists for SARS-CoV-2 and related viruses (SARS-CoV, MERS-CoV), but integrating this knowledge is difficult, since much of it is in siloed databases or in text format. Furthermore, the data required by the research community varies drastically for different tasks - the optimal data for a machine learning task, for example, is much different from the data used to populate a browsable user interface for clinicians.

To address these challenges, we created KG-COVID-19, a flexible framework that ingests and integrates biomedical data to produce knowledge graphs (KGs) for COVID-19 response. KGs provide a mechanism to integrate heterogeneous data and represent their interrelationships. In a KG, discrete pieces of information form distinct nodes interconnected by edges; In our KG, both nodes and edges are typed using the Biolink Model[1]. The KG-COVID-19 framework enables the creation of customized KGs containing COVID-19 knowledge for different applications. For example, a drug repurposing application would make use of protein data linked with approved drugs, while a biomarker application could utilize data on gene expression linked with pathways.

Methods

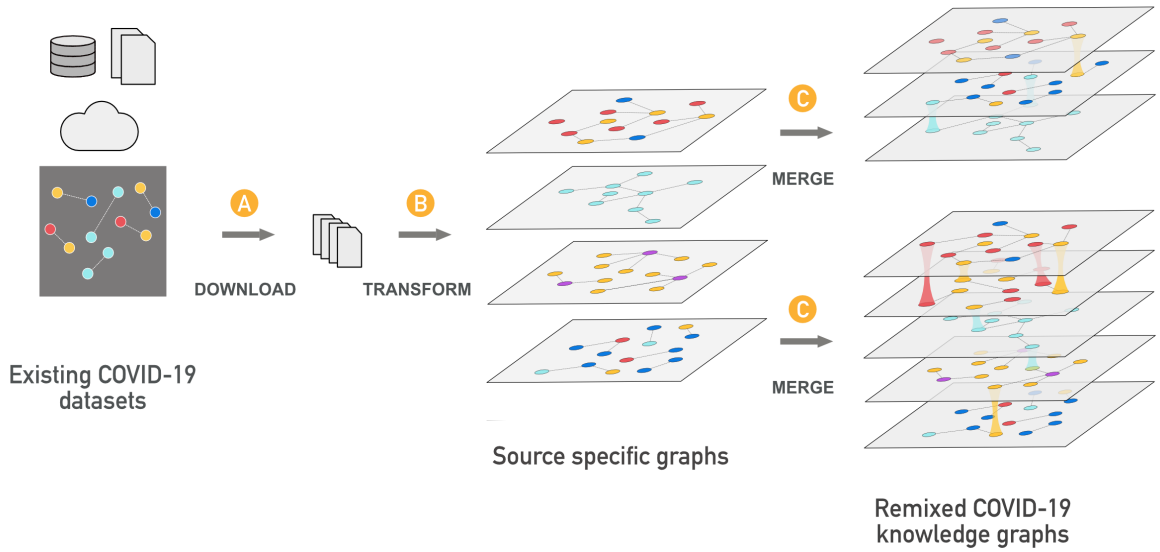


Figure 1. The KG-COVID-19 framework for integrating and remixing data to produce knowledge graphs.

Our process for generating the KG was designed to support interoperability, preserve provenance, and provide the ability to flexibly mix and match data from different sources. The workflow is divided into three steps: data download (fetch the input data), transform (convert the input data to KGX interchange format, normalize identifiers), and merge (combine all transformed sources) (Figure 1). The data we ingest are focused on sources relevant to drug repurposing for our downstream querying and machine learning applications, including drug databases, protein interaction databases, protein function annotations, COVID-19 literature, and related ontologies. We use a core set of standardized ontologies and the Biolink Model, a biological data model for categorizing nodes and edges, to facilitate interoperability and data summarization.

Results

Our KG integrates drug and chemical compound data, functional information for coronavirus and human genes and proteins, protein interaction data, data about the occurrence of concepts (such as ontology terms and proteins) in COVID-19 scientific publications, as well as phenotype, protein pathway, and disease data.

The KG can be applied to several use cases. We provide tooling that allows for sophisticated querying for applications such as drug repurposing. For example, one might query for druggable human proteins that interact directly or indirectly with SARS-CoV-2 proteins. We also integrate tightly with Embiggen, a machine learning (ML) package, for applications such as automatic ranking of antiviral compounds. The KG is also used in the National COVID Cohort Collaborative (N3C)[2] as a source of knowledge about COVID-19, where it can be queried in combination with a national comprehensive patient-level COVID-19 clinical dataset. This large-scale translational integration opens the door to revealing and validating mechanisms, repurposing drugs, and improving care practices as N3C matures. Our KG is also used by the National Virtual Biotechnology Laboratory (NVBL)[3] to supply researchers with information about possible drug repurposing candidates and drug targets, which can also be fed back to N3C for further interrogation

There have been other efforts to construct COVID-19 KGs, each integrating different data sources for different purposes. Our framework adopts a strategy that first integrates the data, and then allows the user to extract subsets of the data for specific use cases to address specific questions. Other advantages of our framework are that it integrates more tightly with ontologies (Human Phenotype Ontology, Mondo disease ontology, and the Gene Ontology) and with downstream machine learning tools; offers a more detailed summary of the contents of its KG; covers a wider range of data sources, and automatically incorporates new and updated data.

Discussion and Conclusions

KG-COVID- 19 enables complex queries over relevant biological entities as well as machine learning to generate graph embeddings for making predictions. The lightweight framework we have developed provides a rapid route for bringing together new sources of data and knowledge, including KGs from several different sources, to form a "hub" to support COVID response efforts across the translational spectrum.

Acknowledgements

This work was supported in part by grants from the Director, Office of Science, Office of Basic Energy Sciences and Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231, as well as NIH NCATS U24TR002306.

References

1. Biolink Model. Available from: https://biolink.github.io/biolink-model/
2. Haendel M, Chute C, Gersing K. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. J Am Med Inform Assoc 2020 Aug 17. http://dx.doi.org/10.1093/jamia/ocaa196
3. US Department of Energy, Office of Science. National Virtual Biotechnology Laboratory, 2020. Available from: https://science.osti.gov/nvbl

Mine the Gap: Utilizing Process Mining to Identify Actual Stroke Care Versus the Guidelines

Christian C. Rose[*,1], Morteza Noshad[*,1], Vincent X. Liu[3], Julia Adler-Milstein[4], Jonathan H. Chen[1,2]
[1]**Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA**
[2]**Division of Hospital Medicine, Stanford University, Stanford, CA**
[3]**Kaiser Permanente Division of Research, Oakland, CA**
[4] **University of California San Francisco, School of Medicine, San Francisco, CA**
[*]**Both authors contributed equally**

Introduction

Many life-threatening emergent medical conditions require time-sensitive progression through multiple points of evaluation and management to deliver definitive life-saving interventions. In the case of an acute stroke, blood samples must be collected by a nurse, evaluation must be performed by a physician, the CT scanner must be prepared by a technician and a pharmacist must prepare the medication for delivery as soon as a stroke has been identified by a radiologist. Studying and improving care processes for these conditions underlies the vision of a learning healthcare system.[1]

Given that these conditions are often complex, requiring multiple providers with differing responsibilities in multiple care settings, it can be difficult to obtain granular detail about how they occur in practice. However, understanding the current process is critical for determining how much variation there is in the process across the organization, where performance problems exist, and where to invest in improvements.

Currently, defining current state processes often relies on expert opinion and recall instead of direct observation. When it is directly observed, this tends to occur in-person or via video surveillance. This can be time-consuming, miss events that do not occur directly at the bedside (such as telemedicine consultation), only represents the time period of that observation and can bias the observed practice, limiting longitudinal evaluation and generalizability.

Process mining, the method of determining the order of events from an event log, may address these issues.[2] Process mining can help organizations easily capture workflow information from enterprise systems and provides detailed, data-driven insights about how key processes are being performed. These logs may illuminate how computer-mediated work, where most of the clinical time may be spent,[3] is really happening - including who did it, how long it takes, and how it deviates from best practice guidelines.

In this exploratory data analysis, we utilize process mining of the event log from a common EHR vendor, Epic (Build ?), at our academic medical center to build a model for tPA management of stroke care - which could then be compared to the medical center's stroke guidelines. Specifically, we sough to assess whether this method helps us understand how the stroke process is currently performed at our institution, providing insight that can then be used to close the gap between actual and ideal care.

Methods

Process mining is based on a set of simple rules to create a graph. In this graph the nodes represent the clinical events, edges represent the subsequent events, edge labels show the relative lag between the events. The graph nodes are time-ordered from top to the bottom. The process, when applied to a cohort of stroke patients, starts with the hospital admission and ends with the tPA-administration.

The steps for the process mining graph are as follows:

1- Sort all clinical events for each patient according to the event times. 2- Choose the top n most prevalent clinical events among all patients. 3- Create nodes according to the events from step 2 and set a rank for each node according to their average times after the patient's admission. 4 - Among all patients and based on the events chosen at step 2, all subsequent event pairs are counted 5- Sort the event pairs according to their prevalence and choose the top m 6- Assign directed edges to node pairs according to step 5. 7- Remove the edges which connect higher rank nodes to

lower rank nodes 8- If the ranks of a set of nodes are within a threshold eta from each other, merge all of them into a single node 9- Plot the graph with the nodes with their location according to their rank from top to bottom.

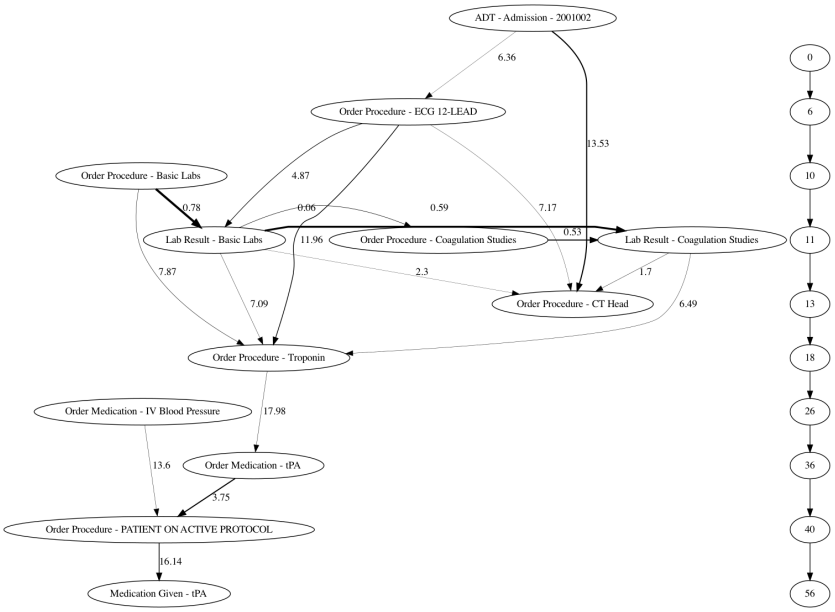


Figure 1: A sample time-ranked process mining graph of the tPA-treated stroke patients. The side timeline graph shows the relative average times (min) after the patient's hospital admission. The widths and labels of the edges respectively show the frequency of the sequence and the lag between events.

Figure 1 shows a sample time-ranked process mining graph of top 12 clinical events for the tPA-treated stroke patients.

Discussion

We were able to automatically create a workflow model for the care of acute stroke patients receiving tPA in our hospital by utilizing the event log data captured in our EHR. Notably, we were able to determine the timing of key events like CT scan ordering and delivery of alteplase (tPA). Of note, however, some procedures, like coagulation studies or troponin testing appear to occur later than other lab testing studies, which might lead to delays in care or suggest a possible future intervention to group these studies in the future.

This first step in automating the process mining for a major EHR vendor sets the stage for future work to identify points in the workflow that may be bottlenecks or where work is happening in ways that deviate from best practice guidelines.

Given our successful proof of concept identifying the care process for a condition like stroke from the event log - which has a known recommended pathway - it should be feasible to produce similar results for conditions like myocardial infarction or sepsis management. However, this method can likely also be applied to more heterogeneous conditions that may currently lack guidelines, such as neonatal fever or even COVID-19. It is hoped that with a clearer understanding of how care is provided, we can then begin to describe and thus improve these processes and the care associated with them.

References

1. Leigh Olsen, Dara Aisner, and J McGinnis. The learning healthcare system: workshop summary. 2007.
2. Wil Van Der Aalst. Process mining. *Communications of the ACM*, 55(8):76–83, 2012.
3. Lena Mamykina, David Vawdrey, and George Hripcsak. How do residents spend their shift time? a time and motion study with a particular focus on the use of computers. *Journal of the Assoc of American Med Colleges*, 2016.

Temporal Phenotypic Progression in Neurofibromatosis Type 1

David J. Schlueter, Ph.D.[1,2], Lisa Bastarache, M.S.[1], Joshua C. Smith, Ph.D.[1], Joshua C. Denny, M.D., M.S.[1,2]

[1] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN;

[2]National Institutes of Health, Bethesda, MD

Introduction

Neurofibromatosis Type 1 (NF1) is a Mendelian disorder caused by mutations to the *NF1* gene, located at 17q11.2, that causes heterogeneous clinical manifestations including neurofibromas, cafe-au-lait macules, and bony abnormalities. NF1 is inherited in an autosomal dominant pattern but about half of cases are thought to result from new mutations[1]. Disease severity can vary significantly, even within families[2], and, with few exceptions, have not been linked to specific mutations. This heterogeneity among NF1 patients motivated us to explore associated conditions and phenotypic temporal progression using a large-scale electronic health record (EHR).

Methods

To assess the association of NF1 on potential observed phenotypes, we first performed a Phenome-Wide Association Study (PheWAS)[3] adjusting for sex, race, EHR length, age at last code, and number of unique dates using phenotypes derived from ICD billing codes. The independent variable for the PheWAS was NF1 status, which was defined using a phenotype algorithm. For each patient with at least one NF-related EHR code, we developed a logistic model based on NF-related codes adjusted for confounding factors such as sex and splined age at first code using previously chart reviewed individuals as labeled records (AUROC: 0.935, 95% CI: 0.894-0.977). Patients classified as not NF1 by the algorithm, as well as other non-NF-coded individuals with any mention of NF in their notes were excluded from the PheWAS. Next, we used temporal EHR information to develop a natural history study within NF1 patients. For strongly associated phenotypes, we asked the following question: "If a patient with NF1 lives up to a certain age without experiencing the phenotype, what is their future prognosis?". To answer this, we found the restricted mean survival time (RMST)[4] up to 20 years of follow-up (equal to area under the survival curve for the next 20 years of follow up) at ages 0, 20, and 50 using Kaplan-Meier survival estimates among NF1 classified individuals. This quantity has the following interpretation for age 20: "If we follow NF1 patients who have survived without the condition up to age 20, what is the average number of years of survival without the condition during the next 20 years?" All analyses were conducted in R[5].

Results

The logistic regression model identified 1025 identified NF1 cases. The PheWAS identified 231 phenome-wide significant ($p<2.68 \times 10^{-5}$) associations with NF1 status (Figure 1). The top parent phenotypes (e.g. Depression instead of Major Depressive Disorder), ranked by smallest p-value and without numerical warning, from the PheWAS were: Neoplasm of Uncertain Behavior, Other benign neoplasm of connective/soft tissue, Malignant/unknown neoplasms of brain/nervous system, Dyschromia/Vitiligo, Cancer of bone/connective tissue, and Benign neoplasm of brain/nervous system. Among all patients at VUMC with these phenotypes, we see that the age distributions among NF1 classified individuals tend to be younger (Figure 1). Furthermore, we see that conditioning on previous survival does not appear to affect 20 year RMST in most of these phenotypes (Figure 2). However, in Other benign neoplasm of connective/soft tissue, we see a decrease in RMST comparing followup ages 0 and 50: 19.2 years, 95% CI: (19.0, 19.4) and 16.7 years, 95% CI: (15.1, 18.3), respectively. Alternatively, we see an apparent increase in RMST for dyschromia: 18.2 years, 95% CI: (17.9, 18.5) and 20 years, 95% CI: (20, 20).

Discussion

We present the use of EHRs to study the natural history of a rare genetic disease. Using PheWAS, we identified 231 associations with NF1. We also demonstrated that the risk of developing some of these conditions varies by age, while others do not. Limitations of this study include lack of granularity among some specific manifestations of

NF1, lack of adjustment of the survival analyses for confounding variables, potential for underestimation of uncertainty due to Normal approximation under small events as well as conditioning on a point estimate of the survival curve, and a focus on select top associations from the PheWAS. Nonetheless, this study illustrates the general feasibility of such an approach to rapidly profile diseases that could be transportable to many EHRs.

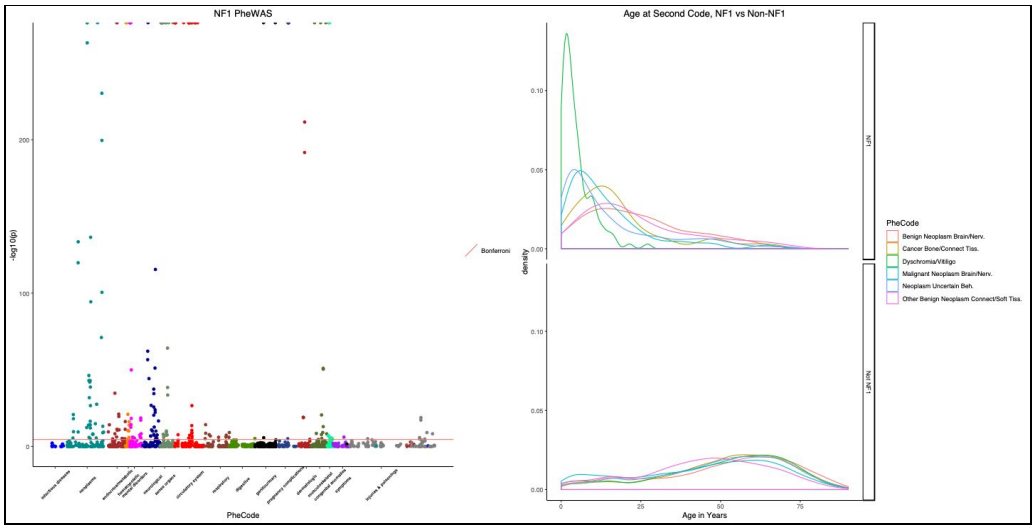


Figure 1. Manhattan Plot of NF1 PheWAS and age distribution of top phenotypes (NF1 vs. Not NF1)

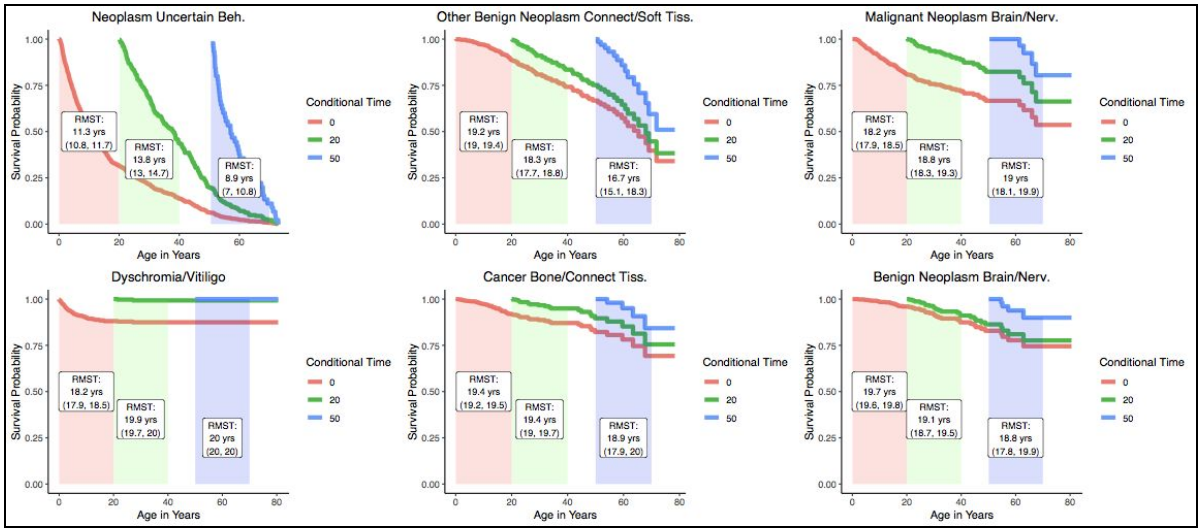


Figure 2. Stratified conditional survival curves for top parent phenotypes from PheWAS, starting at 3 different ages.

References

1. Rasmussen SA, Friedman JM. NF1 gene and neurofibromatosis 1. American journal of epidemiology. 2000 Jan 1;151(1):33-40.
2. Ferner RE, Gutmann DH. Neurofibromatosis type 1 (NF1): diagnosis and management. InHandbook of clinical neurology 2013 Jan 1 (Vol. 115, pp. 939-955). Elsevier.
3. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010 May 1;26(9):1205-10.
4. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC medical research methodology. 2013 Dec 1;13(1):152.
5. Team RC. R: A language and environment for statistical computing.

Interoperable and Computable Evidence-Based Medicine (EBM) on FHIR

Andrey Soares, PhD[1], Lisa M. Schilling, MD, MSPH[1], Brian S. Alper, MD, MSPH, FAAFP, FAMIA[2]
[1]School of Medicine, University of Colorado, Aurora, CO; [2]Computable Publishing LLC., Ipswich, MA

Introduction

This presentation will introduce a set of Fast Healthcare Interoperability Resources (FHIR) resources for expressing evidence in a computable (machine-readable) form. Expressing evidence in machine-interpretable format will make all aspects of evidence identification, evaluation, and reporting more efficient by orders of magnitude.

Methods

Since 2018, the Health Level Seven International (HL7) EBMonFHIR working group[1] has been extending the FHIR infrastructure to provide standards for interoperable data exchange to express biomedical evidence and statistics in a machine-readable format. Via weekly web meetings and five Connectathons, the group created FHIR Resources to represent evidence from clinical studies and tools to assist with the creation and visualization of FHIR Resources. In light of the COVID-19 pandemic, where scientists and physicians need timely results and evidence, and with the difficulties and challenges of disseminating evidence, the working group has expanded to focus on COVID-19 with the creation of the COVID-19 Knowledge Accelerator (COKA) project[2]. COKA is a virtual organization with collaborators from more than 25 organizations in 7 countries working across 10 active working groups to develop and advance interoperability standards for COVID-19 knowledge.

Results

The EBMonFHIR working group has outlined FHIR Resources (Figure 1) for Evidence (http://build.fhir.org/evidence.html), Evidence Variables (http://build.fhir.org/evidencevariable.html), Statistics (http://build.fhir.org/statistic.html) and Ordered Distribution of Statistics (http://build.fhir.org/ordereddistribution.html). The COKA group has further developed these and additional FHIR Resources (Figure 1) to express Evidence Reports (http://build.fhir.org/evidencereport.html) and Citations (http://build.fhir.org/citation.html). Examples of clinical outcomes results extracted from articles in the CORD-19 Open Research Dataset[3] and represented with FHIR Resources can be found at https://www.gps.health/COVID19TrialResults in both human and machine-readable formats.

Discussion and Conclusion

Computable evidence can support relaying EBM components in a manner that is interoperable and consumable by downstream tools and health IT systems to support evidence users (i.e., creators of Biomedical Knowledge Bases, Clinical Practice Guidelines, Clinical Decision Support tools and Systematic Reviews). With the FHIR Resources we can represent evidence knowledge from study results in a way that is readily available, shareable and machine readable. A global standard for data exchange has not been previously developed to bring scientific results to the interconnected computable era. Accelerating the dissemination of scientific knowledge has profound social benefits, starting with more quickly reducing the adverse consequences of COVID-19 and extending to all diseases where diagnostic, management, and preventative evidence exists.

Limitations

The underlying model while in active development, is currently advanced due to global multidisciplinary input. The systems for functional, scalable, and sustainable implementation are the next step for future development.

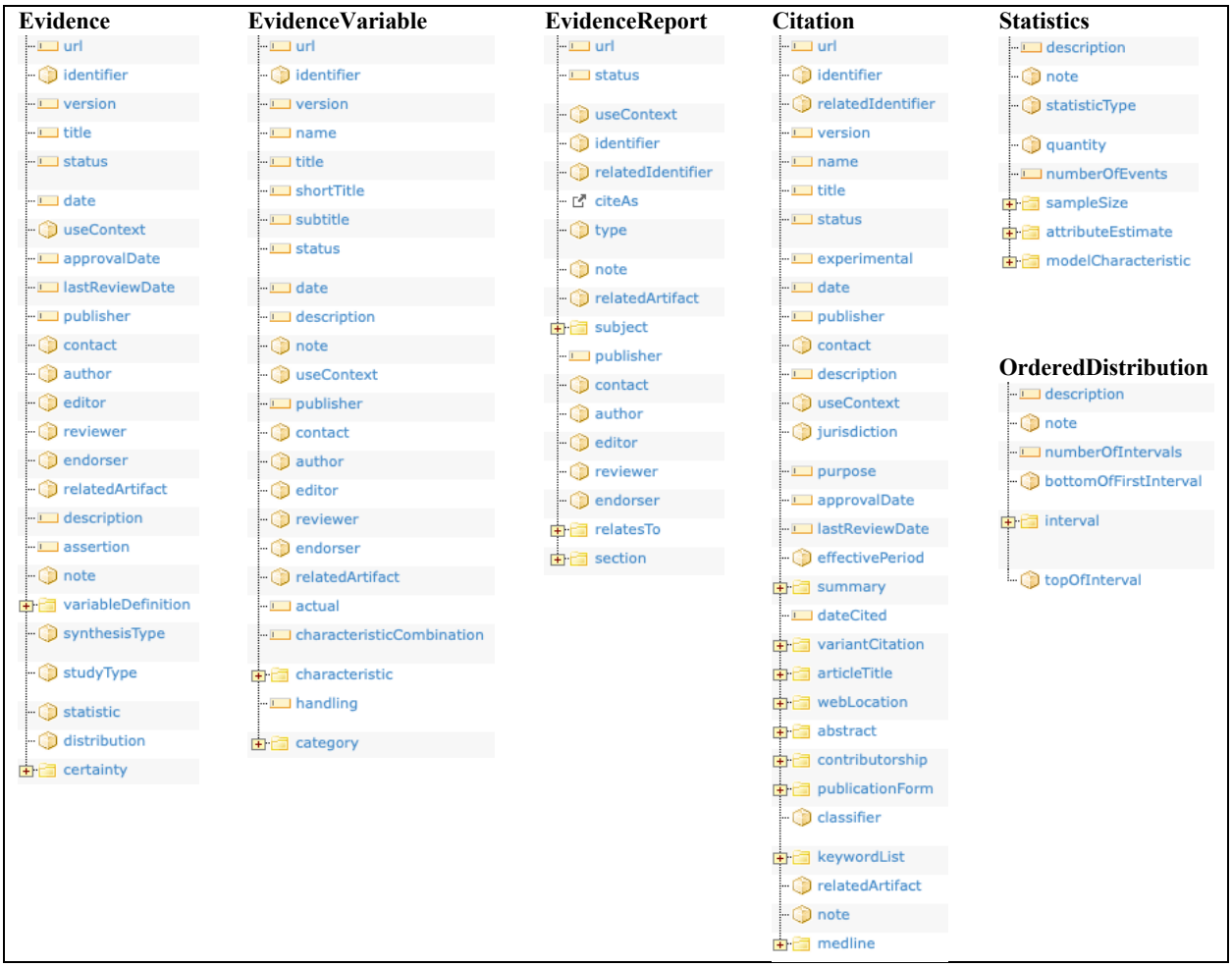


Figure 1. Sample structure of the FHIR Resources to support expressing evidence. Source: http://build.fhir.org

Reference

1. EBMonFHIR - Clinical Decision Support - Confluence. Accessed December 21, 2020. https://confluence.hl7.org/display/cds/ebmonfhir
2. COVID-19 Knowledge Accelerator (COKA) - Clinical Decision Support - Confluence. Accessed December 21, 2020. https://confluence.hl7.org/pages/viewpage.action?pageId=97468919
3. Lu Wang L, Lo K, Chandrasekhar Y, et al. CORD-19: The Covid-19 Open Research Dataset. *ArXiv*. Published online April 22, 2020.

Predictive Modeling Using Transcriptomic Signatures of COVID-19 and Other Infectious Diseases

Harshavardhan Srijay[1], Florica Constantine, MS[1], Micah T. McClain, MD, PhD[1], Christopher W. Woods, MD[1], Ricardo Henao, PhD[1]; [1]Duke University, Durham, NC, USA

Introduction

Given the current climate of global disease, the accurate and specific prediction of acute respiratory infectious disease state, especially COVID-19, can prove crucial in limiting transmission and enabling targeted therapy. As such, we are attempting to use host transcriptomic signatures, a relatively responsive and precise modality in terms of pathogen exposure response, to predict disease state using feature engineering to obtain multiple representations of our expression data that are used to train supervised classifiers. Diagnostic tools such as RT-PCR tests or rapid molecular assays can detect the presence of specific pathogens[1]. However, given that clinical signs and symptoms of such infections are not pathogen-specific, we are interested in building robust multi-disease predictive models that can enable better understanding of the differentiated host transcriptomic signatures between relevant respiratory infections.

Methods

Our whole-blood RNA-Seq data was obtained from the Duke University Center for Applied Genomics and Precision Medicine and contains read count information from patients with one of the five following infections/conditions: COVID-19, other seasonal coronavirus, bacterial, viral (influenza), or the healthy control group. We perform quality control on the data, and after filtration and trimmed-mean (TMM) normalization, the data contains: 13,569 genes and their read counts for 77 PCR-proven symptomatic Sars-CoV-2 (COVID-19) samples from multiple timepoints ranging from 1-35 days after symptom onset, 59 seasonal coronavirus samples, 17 influenza samples, 23 bacterial pneumonia samples, and 19 healthy samples (total of 195 samples).

We first build a L1 regularized multinomial logistic regression model, which minimizes the negative log-likelihood objective function and applies the softmax function to the linear combination of training samples scaled by the learned 13,569-dimensional regression coefficient vector. We conduct nested leave-one-out cross validation to optimize models trained on 194 samples and calculate the predicted probabilities of class membership of each corresponding holdout sample. We train this model on the original data, to establish a baseline classification performance (Table 1).

We then use the Gene2vec model to map each gene to an n-dimensional embedding, such that co-expressed genes are spatially closer in this n-dimensional space. This generates gene embeddings that represent hidden functional gene-gene interactions[2]. We simply use the pre-trained 200-dimensional gene embeddings generated from the Gene2vec model – these have 11,751 genes in common with our data, which we use for the remainder of relevant analysis.

Using these pre-trained embeddings, we train a multilayer perceptron classifier. The first layer is a fully connected dense layer, whose input is the original expression matrix, in which the expression values are converted to proportions that represent the relative expression levels of all genes for a given sample as a weight. We initialize the weight matrix for this layer using the 11,751 x 200-dimensional pre-trained gene embedding matrix. Initially, we train this model with this weight matrix frozen, but then we compare it to the same model where we instead unfreeze these weights and allow the pre-trained embeddings to be tuned during learning. The output of this first hidden layer is a 200-dimensional vector for each sample obtained by computing the weighted average of all 11,751 gene embeddings, where each embedding corresponding to a given gene is weighted by the expression proportion level of that gene. Then, this is passed into a fully connected softmax layer, whose output is the predicted probabilities of class membership for each sample. We conduct leave-one-out cross validation to make predictions for each sample as shown in Table 1. We conduct backpropagation by minimizing the categorical cross-entropy loss using the Adam learning rate method, with 75 epochs, batch sizes of 10, and a learning rate of 0.0026.

Lastly, we conduct non-negative matrix factorization (NMF), which extracts and constructs a reduced representation of our original (non-negative) expression data into a factorization matrix with a specified number of metagenes (rank) that capture subtle, context-dependent inter-gene dependencies not captured in the original expression data[3]. We tested this matrix with ranks ranging from 10-50 and found that rank 33 NMF factorization matrices had the optimal combination of minimal model complexity and maximal classification strength. We run the logistic regression classifier on this rank 33 NMF data using leave one out cross validation. Similar to the logistic regression model trained on the original data above, this model feeds the 33-degree normalized feature vector $\mathbf{x}$ for a given sample into a linear predictor of 33 learned regression coefficients which scales each covariate to represent the log-odds ratio of a given disease group. This is fed into a softmax function to convert the logit scores to conditional probabilities p_i (Figure 1).

Results

Table 1. Classification metrics for all 4 multinomial classifiers: Logistic regression models trained on original and NMF data, and MLP model using frozen pre-trained embeddings and using tuned embeddings.

	COVID -19	CoV Other	Bacterial	Influenza	Healthy	Mean AUC	Accuracy	Kappa
Original Data	0.950	0.933	**0.998**	0.988	**0.945**	0.963	0.841	0.778
33 NMF Features	0.968	**0.968**	0.972	**0.997**	**0.945**	**0.970**	**0.882**	**0.836**
Pre-Trained Embeddings	0.765	0.775	0.985	0.870	0.909	0.861	0.585	0.381
Trained Embeddings	**0.969**	0.955	0.994	0.991	0.907	0.963	0.862	0.805

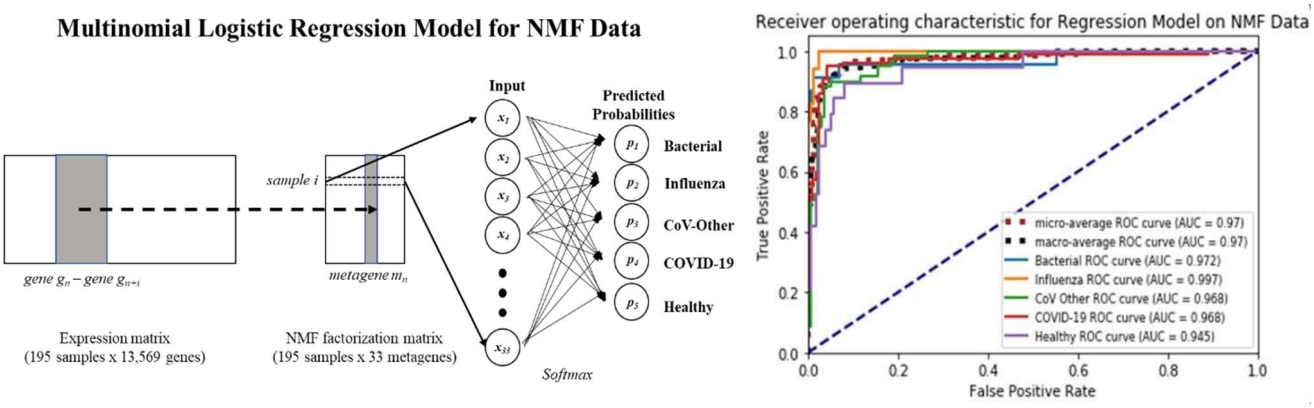


Figure 1. (Left) - architecture for prediction of given sample *i* in the strongest model – multinomial model trained on NMF data; (Right) – corresponding AUC/ROC curves for the NMF model

Discussion and Conclusion

Our results overall support the potential for a more robust and precise diagnostic tool for disease state classification and prediction, using both feature extraction and feature learning methods of NMF and the Gene2vec algorithm. The logistic regression model trained on the NMF data was the strongest. Also, the MLP model using tuned embeddings was as strong as the baseline logistic regression model trained on the original data. This suggests that models trained on learned/extracted features can be less complex (and thereby more generalizable), while maintaining at or above the strength of the most complex model's performance. We intentionally designed the models to be simple, with as few trainable parameters as necessary, in order to reduce complexity and minimize the likelihood of overfitting.

This shows the potential for a diagnostic strategy for acute respiratory infections that can reliably and seamlessly discriminate between infections, which can help enable simultaneous multi-disease testing and illuminate key differences in the host transcriptomic signatures between infections that can improve treatment for patients. We do this by constructing simple, generalizable, and powerful models using gene expression. While the work done here was primarily in model building and optimization, further work also must be conducted to translate these findings to understand the biological relevance of these results, possibly by incorporating an attention mechanism into the models that identifies genes responsible for the differentiated host responses between infections or by conducting gene set enrichment analysis for the NMF metagenes, for instance. Future work could also identify other quantifiable biomarkers outside the transcriptomic modality that could lead to stronger class discrimination and predictive ability.

References

1. Islam K, Iqbal J. An Update on Molecular Diagnostics for COVID-19. Front Cell Infect Microbiol [Internet]. 2020 Nov [cited 2020 Dec 12]; 10. Available from https://doi.org/10.3389/fcimb.2020.560616.
2. Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. Gene2vec: distributed representation of genes based on coexpression. BMC Genomics [Internet]. 2019 Feb [cited 2020 Aug 8]; 20(82). Available from https://doi.org/10.1186/s12864-018-5370-x.
3. Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput Biol [Internet]. 2008 July [cited 2020 Aug 3]; 4(7). Available from https://doi.org/10.1371/journal.pcbi.1000029.

Association of a history of pneumonia with mortality for Coronavirus Disease 2019 (COVID-2019)

Zachary H Strasser MD,[1,2] Hossein Estiri PhD, [1,2,3] Shawn N. Murphy MD, PhD[1,2,3]

[1]Harvard Medical School; [2]Mass General Hospital; [3]Mass General Brigham, Boston, MA

Abstract:

A history of pneumonia has not been assessed as a risk factor for COVID-19 severity. This study leverages the electronic health record to evaluate this relationship. A nearest neighbor propensity score matching algorithm was performed to control for demographics, comorbidities, and healthcare utilization. After controlling for these confounders, a history of pneumonia was still associated with a 5.4 % increased absolute risk of COVID-19 mortality. Physicians should consider asking about a pneumonia history when assessing patients with COVID-19.

Introduction:

Chronic respiratory disease has repeatedly been identified as a risk factor associated with mortality in Coronavirus disease 2019 (COVID-19).[1-2] However, the term represents a collection of respiratory diseases that have a broad spectrum of etiologies, trajectories, and outcomes. This makes it challenging for a physician to assess how the underlying chronic disease will affect the outcome for patients with COVID-19. Additionally, many of the epidemiology studies that documented an association between chronic respiratory disease and mortality did not control for other comorbidities and demographics which may have been highly correlated with one another.[1-2] This study examines how a previous diagnosis of pneumonia alone, which is a simple and easy-to-assess historical event, is independently associated with increased mortality in COVID-19 patients.

Methods

This is a retrospective cohort study based on the electronic health records of COVID-19 confirmed cases between March 3rd and May 24th, 2020 across the Mass General Brigham (MGB) network. Outcome records were collected up until June 24th, 2020. Diagnosis codes (International Classification of Diseases, Ninth and Tenth Revision) from the MGB Research Patient Data Registry, up until 14 days prior to the positive COVID-19 test date, were included. The diagnostic codes were curated into 46 clinical conditions by a physician which were then used in the model. The primary analysis was the association of a previous diagnosis of pneumonia with death. A 2:1 nearest neighbor propensity score matching algorithm was performed to identify two cohorts of patients with and without a pneumonia diagnosis while controlling for confounders. The groups were matched twice. The first matching was only adjusted for gender and age. Then the cohort was fully adjusted to account for gender, age, comorbidities, and the number of medical encounters. The comorbidities included hypertension, hyperlipidemia, diabetes mellitus, coronary artery disease, heart failure, cerebrovascular disease, chronic kidney disease, COPD, asthma and smoking history. In order to control for patients who repeatedly visit a health care provider, all patients were divided into one of four groups based on the frequency of their encounters with a health care provider. For the fully adjusted match, calipers were set to 0.15. All statistical analyses were performed using R version 3.5. The use of data for this study was approved by the Mass General Brigham Institutional Review Board (2020P001063).

Results

Table 1 shows basic characteristics of the survivors versus the non survivors. The original cohort included 12,224 COVID-19 polymerase chain reaction (PCR) confirmed patients. The average age was 51.6±20.7. Of this cohort, 56.5% of the patients were female. 14.1% of the patients had a previous diagnosis of pneumonia.

Table 1: Demographics and characteristics of COVID-19 positive cohort

	Survivors (n = 11,594)	**Non-survivors (n = 630)**	**Total (n = 12,224)**
Age, mean (SD), y	50.1 (13.5)	77.9 (20)	51.6 (20.7)
Female, No. (%)	6620 (57.1)	286 (45.4)	6906 (56.5)
Pneumonia Diagnosis, No. (%)	1464 (12.6)	263 (41.7)	1727 (14.1)

Table 2 shows a two-sample proportion z-test of the absolute risk for patients with a pneumonia diagnosis and those without the diagnosis with a 95% confidence interval. After matching based on sex and age and repeating the comparison, there is still a significant difference of 8.2% between those with a diagnosis and without. After adjusting for age, gender, comorbidities (including asthma, smoking, and COPD), and medical encounters, a diagnosis of pneumonia is still associated with a 5.4% increase in mortality compared to those without the diagnosis.

Table 2: Standardized Absolute Risks for Death in COVID-19

	Mortality in patients with no pneumonia diagnosis, No./Total No. (%)	**Mortality in patients with a pneumonia diagnosis, No./Total No. (%)**	**Mean difference, % (95% CI)**	**P Value**
Unadjusted	367/10497 (3.4)	263/1727 (15.2)	-11.2 (-13.5 to -10.0)	<0.001
Age- and sex-adjusted	243/3437 (7.1)	263/1727 (15.2)	-8.2 (-10.1 to -6.2)	<0.001
Fully adjusted	166/2410 (6.9)	174/1412 (12.3)	-5.4 (-7.4 to -3.4)	<0.001

Discussion:

The findings suggest the need for prospective studies to go beyond just assessing for chronic lung disease but to consider specifically documenting whether the patient has a history of pneumonia. Even after accounting for chronic lung diseases like COPD, asthma, and a smoking history, pneumonia was still found to be strongly associated with COVID-19 mortality. The strong association seen between pneumonia and COVID-19 mortality suggests that physicians caring for COVID-19 patients should make a point of inquiring about their patients' history of pneumonia when considering their prognosis. There are important limitations to this study as it relies on retrospective data drawn from diagnosis codes and demographics in the electronic health record. The diagnosis codes' presence or absence from a chart does not guarantee that the patient had or did not have the disease. Future studies, which can include chart reviews and patient surveys, could further confirm this relationship between a previous pneumonia diagnosis and COVID-19 severity.

References

1. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020;323(13):1239–1242. doi:10.1001/jama.2020.2648

2. Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. Int J Infect Dis. 2020;94:91-95. doi:10.1016/j.ijid.2020.03.017

Telemedicine Use among Geriatric Outpatients during the COVID Pandemic

Anita Szerszen, DO[1], Yulia Kogan, MPH, MBA[2], Yuval Romm, Bsc[2], Raman Vig, MBA[3], Simita Mishra PhD[2], Edith Burns,MD[4,5]

[1]Department of Medicine, Division of Geriatrics, Northwell Health, Staten Island, New York; [2]Department of Medical Informatics, Northwell Health, New Hyde Park, New York; [3]Department of Telehealth Services, Northwell Health, Syosset, New York, [4]Institute of Health Innovations and Outcomes Research, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, New York; [5]Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Northwell Health, Hempstead, New York

Introduction: In an effort to provide continuity of care and avoid complications due to lapse in monitoring active chronic conditions during the Covid-19 pandemic, health systems have attempted to rapidly expand telemedicine services[1].However, implementing synchronous audio-video conferencing has met with variable success among older patients, who are the highest utilizers of medical care[2]. Over 40% of adults 65 years and older do not subscribe to high speed internet service and about 50% do not own smartphones[3]. While telemedicine health care at home is technologically feasible for seniors, many might have to rely on their family members for access to such services. Little is known about the patterns of use or effect on outcomes of video-enabled telemedicine models of care in geriatric patients. Here we describe telemedicine utilization trends, challenges and potential barriers to uptake and follow-through for older adults and their caregivers during the first seven months of the pandemic in New York State.

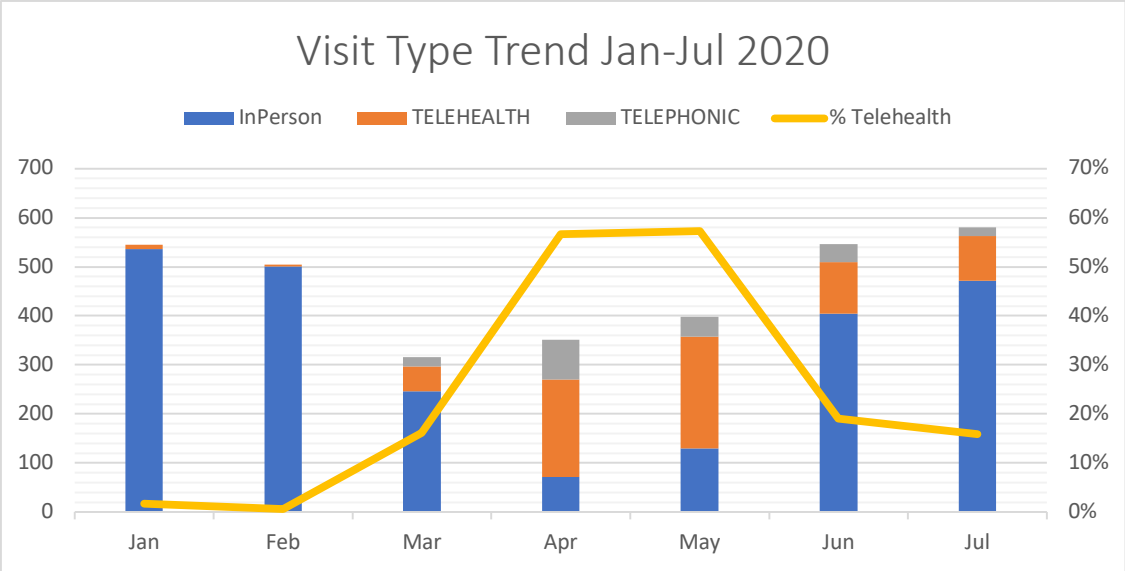
Study Objectives & Methods: This was a retrospective observational study examining patterns of telemedicine use among older adults receiving care from outpatient Geriatrics practices at Northwell Health. Electronic medical records (EMR) were examined for patients $\geq$65-years who had a visit during the first six months of 2020. Demographics and details of engagement in the telemedicine process are described and compared to those who did not access telemedicine during the same time period.

Preliminary Results: A total of 1561 unique patients completed 3139 encounters during the study period; 1090 patients (70% of patients) completed 2304 in-person visits (73.4% all encounters) and 471 patients (30%) completed 835 telemedicine visits (either audio-visual or telephonic, 26.6% of all encounters) (Table 1). Of the virtual visits, 81% had one visit, 14.4% had 2 visits, 2.9% 3 visits, and 1.7% had 4 or more visits. Demographic characteristics were similar among those engaging in either type of encounter (Table 1).

Table 1. Characteristics of Geriatric Patients Engaging in Telemedicine vs. In-person Health Visits during Jan-July 2020.

	All (Total visits=3077)	In Person Visits (n=2267)	Audio-Video Visits (n=628)	Telephone Visits (n=182)
Age				
65-74	439 (14%)	331 (15%)	84 (13%)	24 (13%)
75-84	1164 (38%)	860 (38%)	238 (38%)	66 (36%)
$\geq$ 85	1474 (48%)	1076 (47%)	306 (49%)	92 (51%)
Gender				
Female	2202 (72%)	1662 (72%)	442 (70%)	138 (76%)
Male	875 (28%)	645 (28%)	186 (30%)	44 (24%)
Race				
White	2088 (67.9%)	1563 (68.9%)	410 (65.3%)	115 (63.2%)
Black	330 (10.7%)	223 (9.8%)	80 (12.7%)	27 (14.8%)
Other	500 (16.2%)	365 (16.1%)	109 (17.4%)	26 (14.3%)
Unknown	159 (5.2%)	116 (5.1%)	29 (4.6%)	14 (7.7%)
# Medical Diagnoses	12.6 $\pm$ 6.6	12.6 $\pm$ 6.7	13.6 $\pm$ 6.5	16.0 $\pm$ 8.0
# Medications	8.4 $\pm$ 4.6	8.3 $\pm$ 4.6	9.0 $\pm$ 4.6	10.3 $\pm$ 4.9

Fig 1. Number and percent of Telemedicine Visits vs. In-person Health Visits per month, during Jan-July 2020.



Discussion: About 30% of older adults followed at 2 geriatrics outpatient practices engaged in an audio, telephonic or A-V telemedicine visit during the first 6 months of the COVID pandemic. There was little variation by age, with similar proportions of adults 65-74, 75-84 and above 85 engaging with the technology. There was no apparent difference by gender, or primary language. In our study Black older patients were equally represented among patients who utilized telemedicine service and among Black patients between 75 and 84 years old; there was a trend suggestive of preference for telemedicine. Those engaging in telehealth contacts may have had more medical problems and were talking more medications than those who waited for an in-office visit.

The COVID pandemic resulted in rapid implementation of telemedicine as a modality for delivering care to geriatric patients; at the height of the pandemic in New York State, almost 60% of ambulatory care for our older patients was provided using virtual platform. Utilization patterns mirrored the trends of the pandemic – rising steeply from March to April, falling rapidly in mid-May-July, and settling above baseline and remaining stable (about 13%) by June. Limitations to the present study include reliance on accuracy of retrospective data and is a subject to number of cofounding factors. The data is also based upon a convenience sample of patients who wished to engage in a medical encounter during the time period under study.

At the time of data collection for this proposal, Northwell Health utilized a telehealth platform where a provider had to manually send a link to a patient to conduct a telehealth visit on an ad-hoc basis. Patients were advised to download an app to access a telehealth visit via mobile device. However, the visit could be accessible without downloading the app as well. Northwell is currently implementing an updated version of the telehealth platform that integrates the scheduling system, the telehealth platform, and Northwell's patient portal. Both platforms support translation services for non-English speakers.

Although telemedicine has demonstrated potential for provision of services in the state of emergency, it is important to identify barriers affecting its sustained use in care of older patients. Our next goal is to understand if these visits required additional support (i.e. 3rd party) whether technical issues were involved (e.g. tele-video visits completed or switched to telephone); and role of illness burden – were more visits for those with more medical issues? Impact of COVID – what proportion of visits were specifically attributed to COVID, concerns about COVID? Finally, we plan to look at relationship of telemedicine participation and other health care utilization – i.e. ED and hospitalization visits. Data to be analyzed.

References:

1. Hollander JE, Carr BG. Virtually perfect? Telemedicine for Covid-19. N Engl J Med. 2020;382:1679-1681.
2. Almathami HKY, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. J Med Internet Res 2020;22:e16407
3. Pew Research Center: Internet, Science & Technology. Mobile Technology and Home Broadband 2019; https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019. Accessed, 03/14/2020

Acknowledgments: Iris Berman, RN, MSN; Patrick McCarthy, MBA; Theodore Maniatis, MD; Andrew Tucci, MHA; Jeffrey Paul, MBA; Donna Seminara, MD; Shaun Allicock MS.

Deep Clinical Phenotyping and Network Analysis of Alzheimer's Disease Patients Leveraging Electronic Medical Records Data

Alice Tang[1], Tomiko Oskotsky, MD[1], Marina Sirota, PhD[1]

[1] Bakar Computational Health Sciences Institute, UCSF, San Francisco, California, USA

Introduction

Alzheimer's disease (AD) is the most common type of dementia, making up 60-80% of cases, with devastating impact on patient lives and projections of being an increasing burden for the future[1]. It has been over a century since AD was first identified, yet the disease remains incurable and challenging to understand. Sex has been shown to be an important factor in AD, with higher prevalence in women afflicted by the disease at a 2:1 ratio[1]. Furthermore, sex differences contribute to differing vulnerabilities in AD, as men progress to death quicker while women show higher cognitive resilience despite tau pathology[2]. With an abundance of electronic medical record (EMR) data available over the past decades, there is an opportunity to deeply investigate the risk factors and pathogenesis of AD to aid in disease prevention and understanding. Here we leverage these data through deep clinical phenotyping and network analysis to provide insight into AD clinical characteristics. While some of the molecular differences between sexes have been examined in AD, one goal of our study is to perform a more comprehensive analysis and investigate the role of sex using large phenomics data.

Methods

Patient cohorts were identified from over five million patients in the UCSF EMR database. Due to deidentification, dates are shifted by at most a year and all patients over 90 years of age are represented as 90 years old. AD patients were identified by inclusion criteria of estimated age of >64 years, and ICD10 codes G30.1, G30.8, or G30.9. Control cohorts were identified from all other patients of >64 years by propensity score (PS) matching (`matchit` R package) on sex, estimated age, race, and death status at a 1:2 AD:control ratio.

To evaluate comorbidities, all diagnoses recorded from patient cohorts were identified with the earliest entry of every diagnosis. Diagnoses were classified based upon Level 2/3 ICD10 categories (L2/3 Name) or Full Diagnosis Name and grouped by ICD10 blocks. Comparisons were made between AD and control cohorts, and sex differences were analyzed with a subset of equal number of male and female patients. For each diagnosis, the proportions of patients in each group were compared using Fischer Exact (if <5 patients in a category) or Chi Squared test. Network metrics were computed for male and female networks and compared.

For medications, the proportions of AD and control patients prescribed each medication were compared using Fisher Exact or Chi Squared tests. For laboratory values, median values among all lab tests were identified, and differences in values were compared between AD and controls using Mann Whitney U-test.

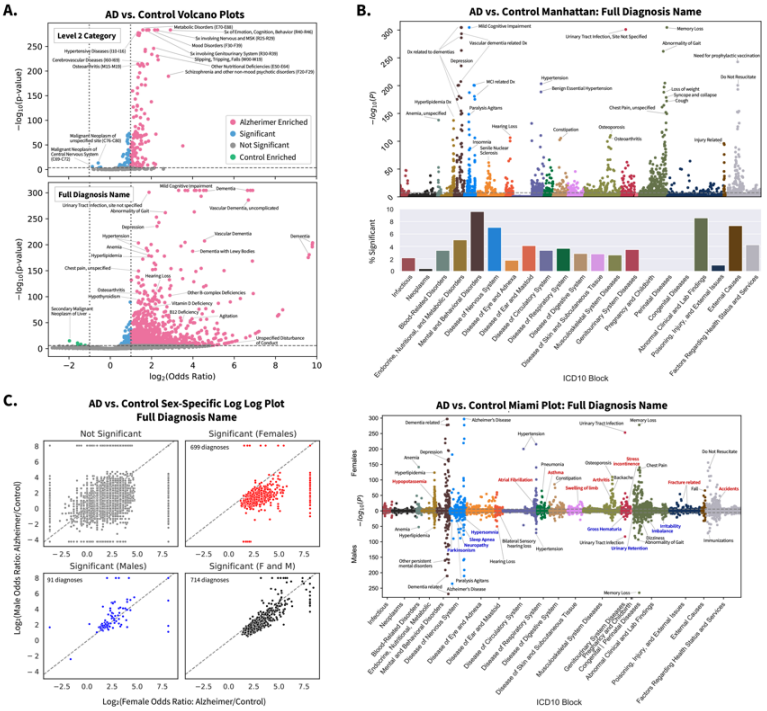


Figure 1. Comparing diagnosis between AD & Controls and with gender stratification. (A) Volcano plot for level 2 categories (top) and full diagnosis names (bottom) compared between AD and control cohorts using Fisher Exact or Chi Squared test. P-value cutoff is Bonferroni corrected (p-val < 2e-8 and 1e-6) with odds ratio cutoff at 2 for AD enriched (pink) or 1/2 for control enriched (green), and remaining significant diagnosis in blue. (B) Above, a Manhattan plot with full diagnosis names colored by ICD10 categories with Bonferroni p-value cutoff. Bottom, percentage of diagnosis in each ICD10 category that is significant. (C) Left: Full diagnosis names compared between AD and controls within each sex. The log of the odds ratio is plotted on the axis, and points are colored by significance (Bonferroni corrected, p-val cutoff > 3e-6). Right: Miami plot of the diagnosis grouped by sex and ICD10 categories. Diagnosis names are colored if significant in only females (red) or only males (blue).

Results

From the UCSF EMR database, we identified 8,804 AD patients and 17,608 PS-matched control patients. Within Level 2 ICD diagnostic categories, 120 significant categories were enriched in the AD group (Figure 1A). Top ICD diagnostic blocks include mental health and behavioral diseases; genitourinary diseases; endocrine, nutritional, and metabolic diseases; and circulatory system diseases (Figure 1B). Within Full Diagnosis Names, 1491 and 7 diagnoses were enriched in AD and controls, respectively (Figure 1A). Top diagnoses in AD include vascular dementia, psychiatric disorders, vitamin deficiency, hypothyroidism, and osteoporosis, while top diagnoses in controls include neoplasms of liver and brain (Figure 1A). When comparing networks, the AD disease network has 243 diagnosis pairs shared by >5% of patients, compared to 1 pair in controls (Figure 2A).

In a sex stratified analysis, female AD patients have diagnoses enriched in injuries, urinary tract infections, osteoporosis, atrial fibrillation, and asthma. Male AD patients have diagnoses enriched in Parkinsonism, hearing loss, sleep disorders, and imbalance (Figure 1C and 1D). When stratifying full diagnosis name networks by sex and shared by >5% of patients within a sex group, female AD patients have 45 shared co-diagnosis pairs (Figure 2B) compared to 14 in male AD patients (Figure 2C), and 0 diagnosis pairs were identified in both control sex groups. Comparison of sex-specific networks show greater closeness centrality and lower eccentricity in female networks (Mann-Whitney U Test, p-value<.01 both metrics).

Within medication differences, the top medications found enriched in AD patients include current treatments like Donepezil, but also vitamin B12 and a variety of dietary supplements and herbs. Medications enriched in controls include aspirin, opioids, furosemide, and dexamethasone. Among significant laboratory value differences, AD patients have higher levels of median blood calcium (p-value 8e-26), red blood cell count (2e-23), serum albumin (2e-14), and cholesterol HDL (1.2e-07), and lower levels of aspartate transaminase (7e-14), white blood cell count (9.1e-12), and ferritin (1e-5).

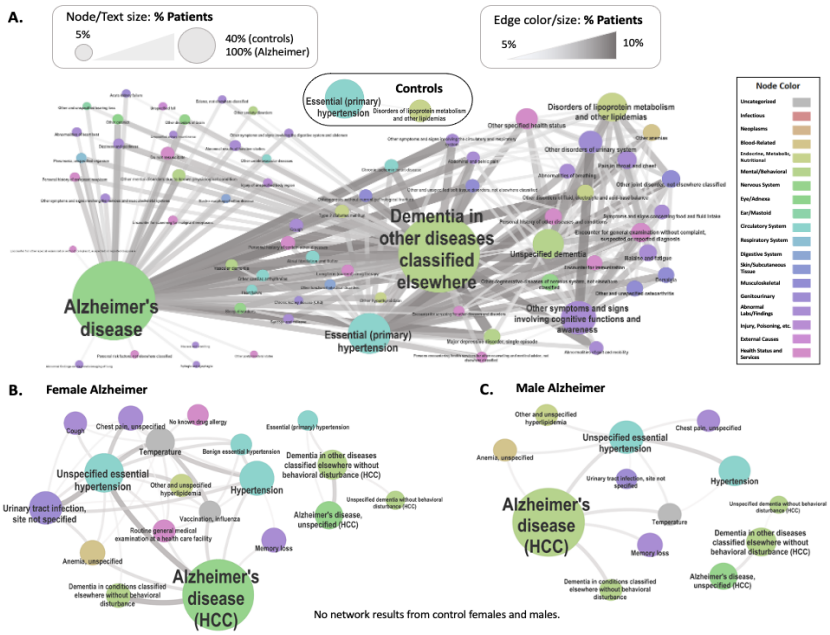


Figure 2. Networks of diagnosis shared between >5% of patients in AD & control cohorts and stratified by sex. (A) Network for level 3 diagnostic categories in AD vs. control patients. Nodes and edges represent >5% of diagnosis or diagnosis pairs shared in each cohort, respectively. (B) Female network of full diagnosis names. Each node and edge represent diagnosis or diagnosis pairs shared by >5% of AD females. No analogous comorbidity network was generated from control females. (C) Analogous network of diagnosis names for males. No network was produced on control males. For each network, the node size, text size, edge size, and edge color represent the number of patients sharing a diagnosis or diagnosis pair. Node colors are based on ICD10 category.

Discussion

Many diagnoses found enriched in our AD cohort have been previously identified as possibly linked with AD, such as midlife hypertension (HTN)[3,4], diabetes[3], vascular dementia[5], osteoporosis[4], and urinary tract infections[6]. Furthermore, from our network analysis, we see higher rates of comorbid conditions among AD patients, particularly with links of HTN-hyperlipidemia, HTN-UTI, and HTN-anemia. Our analysis therefore provides a comprehensive way to identify previously confirmed or new comorbidities and associated factors in AD. Sex stratified analysis shows strong links between HTN-UTI and HTN-chest pain among female AD populations, but not in male AD patients. Female networks also contain more nodes and connectivity than male networks, suggesting greater combined diagnoses in females AD patients. Our analysis identified sex differences that may be linked to AD and that have not previously been explored, such as the enrichment of osteoporosis and UTI in female AD patients, as well as the enrichment of hearing loss and sleep disorders in male AD patients. Furthermore, the enrichment of other neurological pathology found in male AD patients may provide some evidence of lessened resilience in male AD patients' brains[2], as males may have neurological damage either prior or co-occurring with AD disease. Overall, these analyses leverage an extensive clinical dataset to provide a comprehensive overview supporting known or suggested associations with AD, as well eliciting sex-specific differences enriched in AD patients. Nevertheless, currently this analysis only identifies associations with AD. Further work will need to be done to look at diagnoses and other data over time in order to identify temporal relationships. Furthermore, the analyses of the medications and lab test values might lead to hypotheses or preventive or therapeutic strategies, such as identifying medications enriched in controls as potential therapeutic candidates.

References

1. 2020 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2020;16(3):391-460. doi:https://doi.org/10.1002/alz.12068
2. Dubal DB. Chapter 16 - Sex difference in Alzheimer's disease: An updated, balanced and emerging perspective on differing vulnerabilities. In: Lanzenberger R, Kranz GS, Savic I, eds. *Handbook of Clinical Neurology*. Vol 175. Sex Differences in Neurology and Psychiatry. Elsevier; 2020:261-273. doi:10.1016/B978-0-444-64123-6.00018-7
3. Yu J-T, Xu W, Tan C-C, et al. Evidence-based prevention of Alzheimer's disease: systematic review and meta-analysis of 243 observational prospective studies and 153 randomised controlled trials. *J Neurol Neurosurg Psychiatry*. 2020;91(11):1201-1209. doi:10.1136/jnnp-2019-321913
4. Duthie A, Chew D, Soiza RL. Non-psychiatric comorbidity associated with Alzheimer's disease. *QJM Mon J Assoc Physicians*. 2011;104(11):913-920. doi:10.1093/qjmed/hcr118
5. Nucera A, Hachinski V. Cerebrovascular and Alzheimer disease: fellow travelers or partners in crime? *J Neurochem*. 2018;144(5):513-516. doi:https://doi.org/10.1111/jnc.14283
6. Mawanda F, Wallace R. Can Infections Cause Alzheimer's Disease? *Epidemiol Rev*. 2013;35(1):161-180. doi:10.1093/epirev/mxs007

Acknowledgements: We would like to acknowledge Zachary Cutts and Caroline Warly Solsberg for their help in techniques for visualizations and statistical analysis. This work is supported by NIA R01AG060393, R01AG057683.

Outcomes of a Clinical Decision Support (CDS) Tool Informed by Implementation Science: A Cluster-Randomized Controlled Trial to Improve Heart Failure Prescribing

Katy E. Trinkley, PharmD, PhD;[1,3] Miranda E. Kroehl, PhD;[2] Michael G. Kahn, MD, PhD;[1] Larry A. Allen, MD, MHS;[1,3] Tellen D. Bennett, MD, MS;[1,3] Gary Hale, RPh;[1] Heather Haugen, PhD;[1] Simeon Heckman, RN, MS;[1] David P. Kao, MD;[1] Daniel M. Matlock, MD, MPH;[1,3,4] Daniel C. Malone, RPh, PhD;[5] Robert L. Page II, PharmD, MSPH;[1] Krithika Suresh, PhD;[1,3] Chen-Tan Lin, MD[1]

[1] University of Colorado, Aurora, CO, USA; [2]Charter Communications Corporation, Greenwood Village, Colorado; [3]Adult and Child Consortium for Outcomes Research and Delivery Science, Aurora, CO, USA; [4]VA Eastern Colorado Geriatric Research Education and Clinical Center, Aurora, CO, USA [5]University of Utah, Salt Lake City, UT, USA

Introduction

The broad vision of clinical decision support (CDS) tools to improve patient care remains unrealized. To optimize effectiveness, developers are encouraged to apply CDS design best practices (e.g., user-centered design).[1,2] However, comprehensive application of CDS best practices is resource-intensive. As such, institutions often rely on commercially available CDS tools, which may not necessarily follow CDS design best practices. Some have also asserted that commercial CDS tools may be based on content knowledge systems that are uninformative and not clinically relevant, thus less likely to be adopted.[3] However, these assertions are untested.

Retrospective studies suggest that CDS design best practices may improve CDS effectiveness,[2] yet are often minimally applied. Reasons for this may be limited resource availability, skepticism regarding the evidence, or insufficient guidance in how to apply the best practices. Although CDS best practices emphasize the importance of implementation, they provide limited guidance on key implementation science issues including contextual factors that influence the success, sustainability and reproducibility of implementations. Institutions need to understand the return on investment of allocating limited resources to apply CDS best practices with an implementation science framework compared to relying on commercially available CDS tools. Therefore, this study aimed to evaluate the effect of designing an 'enhanced' CDS tool for heart failure prescribing based on CDS best practices and an implementation science framework, as compared to a commercially available CDS tool.

Methods

We conducted an explanatory sequential mixed methods study to evaluate effectiveness and implementation outcomes of the enhanced tool compared to a commercial tool within the electronic health record of 28 primary care clinics across a large regional health system. Both tools aimed to improve evidence-based beta-blocker prescribing for adult patients with heart failure. The design of an enhanced CDS tool was informed by CDS best practices and the Practical, Robust Implementation and Sustainability Model (PRISM) implementation science framework.[4] This included iterative, multilevel stakeholder input (patients, clinicians, managers, leaders) that considered the dynamic interactions of the internal and external environment. The commercial tool followed the vendor-supplied specifications and was not modified for the local context. The first study phase was a pragmatic cluster-randomized trial. The second phase was a series of qualitative interviews with clinicians. Implementation outcomes aligned with PRISM and included patient reach, clinician adoption (did not outright dismiss) and effectiveness (changing prescribing behavior). Differences in effectiveness and adoption rates were tested using a chi-square test. Interviews were thematically analyzed using Atlas.ti software.

Results

Between March 15 and August 23, 2019, the enhanced alert was triggered 106 times for 61 unique patients and 87 unique clinicians. The commercial alert was triggered 59 times for 26 unique patients and 31 unique clinicians (Table 1). Clinician adoption and effectiveness of the enhanced tool was significantly higher than the commercial tool (62.3% vs 28.8% alerts adopted, $p<0.0001$; 14.2% vs 0% alerts changed prescribing, $p=0.006$). Of 21 clinicians interviewed, 15 preferred the enhanced tool and one had no preference. Clinicians preferred the enhanced tool because it 1) was easier to understand the purpose, 2) provided the necessary data to make an informed decision, 3) provided specific

recommendations and made it easy to take action, and 4) used clear language. Five clinicians preferred the commercial tool because of brevity, presence of a dismiss option, or because of the many options within the order set to support holistic management of heart failure (e.g., order multiple meds, labs, imaging studies).

Table 1. Summary of CDS alerts, adoption and effectiveness, n (%)

Characteristic	Enhanced	Commercial
Alerts (of patients who had a visit with primary care during evaluation period)		
Alerts total	106	59
Unique patients with alert*	61	26
Unique clinicians alerted	87	31
Adoption (did not outright dismiss CDS)		
Alerts adopted	66 (62.3)	17 (28.8)
For unique patients	44 (72.1)	13 (0.50)
By unique clinicians	60 (69.0)	13 (41.9)
Effectiveness (changed prescribing)		
Alerts changed prescribing	15 (14.2)	0 (0)
For unique patients	15 (24.6)	0 (0)
For unique patients with first alert	13 (86.7)	0 (0)
By unique clinicians**	14 (16.1)	0 (0)
By attending physicians	9 (60)	0 (0)
By medical resident	2 (14.3)	0 (0)

*Four patients were exposed to both alerts and one clinician was exposed to both alerts.
**One clinician changed prescribing for two different patients.

Discussion

This study suggests applying CDS design best practices with an implementation science framework to CDS tools leads to meaningful improvements in patient reach, clinician adoption and effectiveness of behavior change, as compared to some commercially available CDS tools. Our quantitative findings are substantiated by the results of the qualitative clinician interviews. Although 5 (of 21) clinicians interviewed preferred the commercial alert, their preference was driven by design features that were not prioritized by the majority of interview participants and were associated with lower adoption and effectiveness.

When comparing the enhanced CDS to published evaluations of CDS to improve heart failure prescribing of similar chronic medications, our rates of adoption and effectiveness were higher. Other studies comparing a CDS tool to no CDS demonstrated minimal difference in changing prescribing (23% versus 22% without CDS; 3.6% versus 0.9% without CDS, $p<0.01$).[5,6] In contrast, we found a 24% improvement in prescribing compared to an active control.

These findings suggest that applying CDS design best practices with an implementation science framework may result in more effective CDS tools and ultimately greater improvements in patient outcomes. Future research should assess generalization of these results to other CDS situations and compared to other commercially available CDS tools. Not all commercial CDS tools have the same limitations. Future research is needed to further develop this implementation science-based approach by incorporating rapid, iterative prototyping to expedite the creation of widely adopted, effective and sustainable CDS tools.

References

1. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Informatics Assoc 2003;10:523–30.
2. Horsky J, Schiff GD, Johnston D, Mercincavage L, Bell D, Middleton B. Interface design principles for usable decision support: A targeted review of best practices for clinical prescribing interventions. J Biomed Inform. 2012;45:1202–16.
3. Shah NR, Seger AC, Seger DL, Fiskio JM, Kuperman GJ, Blumenfeld B, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. J Am Med Informatics Assoc. 2006;13:5–11.
4. Feldstein AC, Glasgow RE. A practical, robust implementation and sustainability model (PRISM) for integrating research findings into practice. Jt. Comm. J. Qual. Patient Saf. 2008;34:228–243.
5. Tierney WM, Overhage JM, Murray MD, Harris LE, Zhou X-H, Eckert GJ, et al. Effects of computerized guidelines for managing heart disease in primary care. J Gen Intern Med. 2003;18:967–76.
6. Blecker S, Pandya R, Stork S, Mann D, Kuperman G, Shelley D, et al. Interruptive versus noninterruptive clinical decision support: usability study. JMIR Hum factors. 2019;6(2):e12469.

Maximum-flow formulation improves feature stability of machine learning models trained on high-dimensional whole genome datasets

Maya Varma, BS[1]; Kelley M Paskov, MS[1]; Brianna S Chrisman, MS[1]; Min Woo Sun, BS[1]; Jae-Yoon Jung, PhD[1]; Nate Stockham, MS[1]; Peter Washington, MS[1]; Dennis P Wall, PhD[1]
[1]Stanford University, Stanford, CA, USA

Introduction

The advent of inexpensive whole genome sequencing methods in recent years has led to the creation of supervised machine learning approaches for predicting putative genetic variants from sequence data[1,2]. Since the high dimensionality of variant feature sets paired with a comparatively low number of training samples tends to result in model overfitting, feature selection methods, such as regularization, are often used to narrow the genomic search space and improve model generalizability. Although such methods are widely used, regularized machine learning models tend to face issues related to feature stability and robustness; specifically, slight perturbations to the dataset or model often drastically alter the subset of top-ranked variants determined by the model to be correlated with the phenotype, a phenomenon known as feature instability[3,4]. The absence of stability among predictive features means that variants with high coefficient scores may not necessarily provide insight into the biological mechanisms underlying a condition.

In this work, we present an approach to improve the stability of regularized machine learning methods through incorporation of biological information. We hypothesize that the observed instability of regularized machine learning models trained on large genome datasets results from linkage disequilibrium (LD) between variants. To perform biologically-informed feature selection, we design an algorithm based on maximum flow, which utilizes the presence of linkage disequilibrium to identify a stable set of variants potentially contributing to the phenotype.

We focus our analysis on autism spectrum disorder (ASD), a prevalent neurodevelopmental disorder affecting one in 40 children in the United States. In a previous study, we utilized regularized machine learning to show that variants in a specific subclass of noncoding DNA known as simple repeat sequences (SRS) may be predictive of the ASD phenotype[5]. Here, we extend this work by narrowing the search space and identifying a stable set of putative SRS variants potentially linked with ASD.

Methods

We analyze 232,193 variants in simple repeat sequences (SRS), collected from whole genome sequences of 2182 children with ASD and 379 controls, and we encode genotype calls at each variant location into a binary feature matrix. We perform 5-fold cross-validation with a logistic regression classifier and extract variants assigned non-zero scores from each validation fold. To address class imbalance between the case and control populations, we adjusted classifier weights to be inversely proportional to class sizes.

We assemble the five sets of variants into a graph subject to LD constraints. Specifically, we create a graph G to represent the presence of LD between pairs of variant. Each node n in the graph is defined by a variant v as well as the fold in which it occurs f, which we represent as the tuple $n = (v, f)$. Consider a pair of nodes $n_1 = (v_1, f_1)$ and $n_2 = (v_2, f_2)$; an edge is drawn between the pair if the following criteria are satisfied: (1) n_1 and n_2 are present in neighboring folds such that $f_2=f_1+1$ and (2) n_1 and n_2 are in linkage disequilibrium as indicated by the R^2 value between v_1 and v_2 exceeding 0.8 [6].

We now utilize maximum flow to identify stable variants across folds. We restructure $G=(N, E)$ into a directed, acyclic flow network $L=(N, E)$ such that it is amenable to the maximum flow formulation, a concept well studied in graph theory. To do so, we add a source node s, a sink node t, and directed edges to the graph; we also split each node into two in order to constrain flow through the graph. Our goal is to maximize the total flow passing from the source to the sink node of a graph with respect to the criteria defined above.

The flow through L is computed using the Ford-Fulkerson algorithm. The resulting maximum flow value defines the number of valid paths through the graph, and the nodes along each flow path from the source to sink represent a set of SRS regions that remain stable across folds after accounting for the presence of LD.

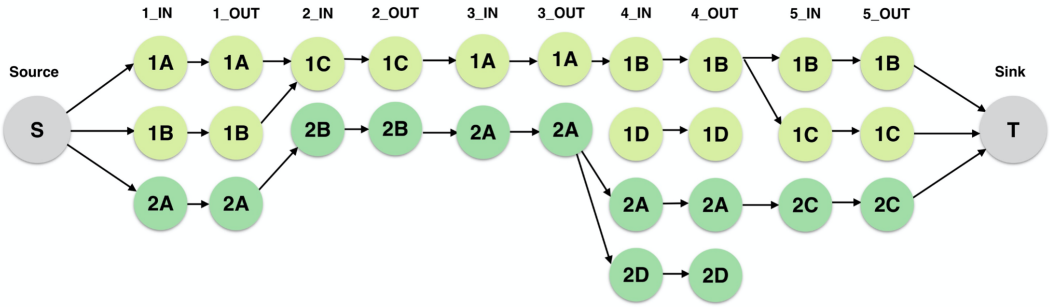


Figure 1: Consider a simplified representation of the dataset, consisting of variants 1A, 1B, 1C, and 1D (located on chromosome 1) as well as variants 2A, 2B, 2C, and 2D (located on chromosome 2). Source and sink nodes are added to the graph, and the variants in each fold are duplicated to constrain flow through the network.

Results

We performed initial measurements of feature stability prior to implementing the maximum-flow formulation. Regularized logistic regression models were trained on five subsets of the 232,193 SRS variant features, and variants with non-zero coefficient scores were extracted. Pairwise comparisons of feature coefficient scores across all ten pairs of feature lists resulted in Pearson correlation coefficients ranging from 0.394 to 0.451.

The maximum flow formulation allowed us to identify 50 stable regions (representing 55 variants). We then performed 5-fold cross-validation across the training set with the new subset of variant features and determined the stability of these features by recomputing the Pearson correlation coefficients. Results show a higher degree of stability, with Pearson correlation coefficients ranging between 0.954 and 0.976.

A literature search showed that several identified variants are located in or near genes associated with neural function.

Discussion

In this work, we developed an algorithm based on maximum flow, which utilizes the presence of linkage disequilibrium in order to perform dimensionality reduction. In contrast to traditional feature elimination methods, the maximum flow approach utilizes biological knowledge to identify a core subset of stable, putative variants. To the best of our knowledge, such a method has never been used before for analysis of high-dimensional datasets. Harnessing information provided by linkage relationships between variants can allow for effective filtration of high dimensional feature spaces, enabling accurate categorization of feature importance for machine learning models.

We then utilized this method to perform a targeted investigation of the noncoding genome, examining the effects of variants in SRS regions on the ASD phenotype. Our analysis extracted a list of 55 candidate variants, which were demonstrated to be highly stable.

Ultimately, the methodology designed in this work allows for the creation of robust, interpretable, and scalable machine learning models that can effectively identify predictive variants from a high-dimensional feature space.

References

[1] Sik D, Ho W, et al. Learning SNP based prediction for precision medicine. *Frontiers in genetics*, 2019

[2] Okser S, et al. Regularized Machine Learning in the Genetic Prediction of Complex Traits. *PLOS Genetics.* 2014.

[3] Tolosi L and Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 05 2011.

[4] Mungloo-Dilmohamud Z, Jaufeerally-Fakim Y, and Pena-Reyes C. Stability of feature selection methods: A study of metrics across different gene expression datasets. *Bioinformatics and Biomedical Engineering*, 2020.

[5] Varma M, Paskov KM, et al. Outgroup Machine Learning Approach Identifies Single Nucleotide Variants in Noncoding DNA Associated with Autism Spectrum Disorder. *Pacific Symposium of Biocomputing*, 24:260–271, 2019.

[6] Wenbo M and Zhang W. Molecular Approaches, Models, and Techniques in Pharmacogenomic Research and Development. In *Pharmacogenomics*, pages 273–294.

A COVID-19 Application Ontology for the ACT Network

Shyam Visweswaran, MD, Ph.D.[1]; Malarkodi J. Samayamuthu, MD[1]; Michele Morris, BA[1]; Griffin M. Weber MD, Ph.D.[2]; Douglas MacFadden, MS[2]; Philip Trevvett[2], Jeffrey G. Klann, PhD[2,3]; Vivian Gainer, MS[3]; Shawn N. Murphy MD, Ph.D.[2,3]

[1]University of Pittsburgh, Pittsburgh, PA; [2]Harvard Medical School, Boston, MA; [3]Mass General Brigham, Boston, MA

Introduction

The Accrual to Clinical Trials (ACT) network is a federated network of Clinical and Translational Science Award (CTSA) hubs that has implemented an efficient and extensible electronic infrastructure to transform clinical and translational research[1]. The network consists of local Informatics for Integrating Biology at the Bedside (i2b2) electronic health record (EHR) data repositories that are integrated by the Shared Health Research Information Network (SHRINE) platform. The SHRINE platform employs a user-friendly query language that enables querying data across the sites in the network using medical terminologies where in each terminology the terms are arranged in a hierarchy for easy navigation. In the context of SHRINE/i2b2, we call a terminology with hierarchical relations an application ontology.

In response to the global pandemic caused by SARS-CoV-2, we mobilized the ACT network, which links EHR data on more than 150 million patients across 50 CTSA hubs to support COVID-19 research at a national scale. To do so, we rapidly developed and deployed a COVID-19 application ontology and augmented EHR data to support the terms in the ontology.

Methods

The COVID-19 ontology has several unique features to enable users to conveniently query with terms that are related to the course of illness and outcomes (see Figure). 1) We identified and categorized **emerging terms** from ICD-10-CM, CPT-4, HCPCS and LOINC terminologies that were introduced in response to SARS-CoV-2. 2) We created **computable phenotypes** to characterize the course of illness and outcomes in COVID-19 that included illness severity, respiratory therapy management, and level of care. We developed computable phenotypes for three levels of illness severity – moderate, severe, and death – and for four levels of respiratory therapy management – supplemental oxygen, intubation, mechanical ventilation, and extracorporeal membrane oxygenation (ECMO) – and for each of these phenotypes we collected a set of relevant codes from ICD-10, CPT-4, and DRG. 3) We created several **derived terms**, such as, "Moderate Illness (Derived)". These derived terms are useful in mapping data from EHRs of patients who are currently hospitalized and for which ICD-10 or CPT-4 codes may not be available. 4) We created **harmonized value sets** for the growing number of SARS-CoV-2 nucleic acid antigen and antibody tests. The harmonized values comprised of positive, negative, equivocal, and pending values and allowed mapping of variously reported results to a set of four values. 5) We identified terms in **existing** ACT ontologies that were likely to be useful for COVID-19 research and included them in the COVID-19 ontology for convenience. For example, we identified and added classes of medications that are relevant to COVID-19 research.

To rapidly obtain input from a diverse group of ACT members and to quickly identify errors we communicated through online meetings, a shared GitHub repository and an i2b2 server dedicated for reviewing the ontology.

Results

In a span of eight weeks, we developed, released and deployed three versions of the ontology on the ACT COVID-19 network that consists of 9 ACT sites. The current ontology consists of 52,476 terms from the domains of diagnosis, procedures, medications and laboratory tests. Counts of patients with SARS-CoV-2 nucleic acid laboratory test results from the ACT COVID-19 network are shown in the Table. The ontology files for the i2b2/SHRINE platform with accompanying documentation are available freely on GitHub at https://github.com/shyamvis/ACT-COVID-Ontology. All ACT ontologies including the COVID-19 ontology can be viewed at http://dbmi-ncats-test01.dbmi.pitt.edu/webclient/.

Preliminary feedback from users of the ACT COVID-19 network has been encouraging. Users found the computable phenotypes and the harmonized SARS-CoV-2 laboratory test values to be critical in identifying relevant cohorts. Furthermore, the ontology is enabling other COVID-19 research efforts. For example, the National COVID Cohort Collaborative (N3C)[2] that is building a centralized national EHR data resource for COVID-19 research and the Consortium for Clinical Characterization of COVID-19 by EHR (4CE)[3] that is collecting and conducting EHR data-driven studies of COVID-19 have leveraged terms from the ontology.

Conclusions

We developed an ontology to enable real-time cohort discovery across the ACT network that has several unique features that are relevant to COVID-19 research. The ontology has been deployed on the ACT COVID-19 network of 9 sites and is now being deployed on the full ACT network of 50 sites. In addition to supporting COVID-19 research on the ACT network, the ontology is enabling other research efforts such as the N3C and the 4CE.

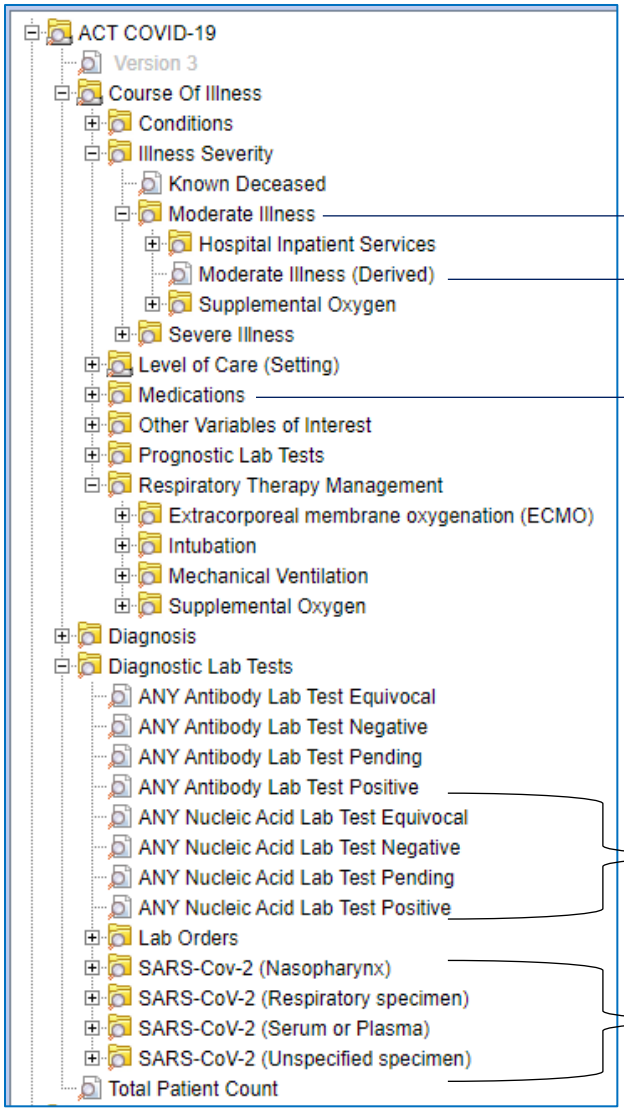


Table. Counts of patients with SARS-CoV-2 nucleic acid laboratory test results from the ACT COVID-19 network.

CTSA hub	Positive	Negative	Equivocal	Pending
Site 1	9,595	95,095	0	175
Site 2	3,855	79,715	1,195	40
Site 3	1,215	46,800	110	25
Site 4	11,010	87,080	405	530
Site 5	8,465	90,490	0	0
Site 6	235	840	0	0
Site 7	3,025	46,440	0	0
Site 8	4,030	48,720	35	0
Site 9	2,710	66,535	225	15

Figure. Screenshot of ACT COVID-19 ontology with illustrative examples of computable phenotype, derived term, existing terms, harmonized value set, and emerging terms.

References

1. Visweswaran S, Becich MJ, D'Itri VS, Sendro ER, MacFadden D, Anderson NR, Allen KA, Ranganathan D, Murphy SN, Morrato EH, Pincus HA. Accrual to clinical trials (ACT): A clinical and translational science award consortium network. JAMIA open. 2018 Oct;1(2):147-52.

2. Melissa H, Christopher C, Kenneth G. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. Journal of the American Medical Informatics Association. 2020 Aug 17.

3. Brat GA, Weber GM, Gehlenborg N et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. npj Digit. Med. 3, 109 (2020). https://doi.org/10.1038/s41746-020-00308-0

Transparency, Reproducibility, and Team Science in the National COVID Cohort Collaborative (N3C)

Anita Walden, M.S.[1], Davera Gabriel, RN[2], Julie McMurry, MPH[3], Andrew Williams, Ph.D[4], Vignesh Subbian, Ph.D[5], Kenneth Gersing, M.D.[6], Nomi L. Harris, M.S.[7], Christopher G. Chute, M.D. Dr. PH [2], Melissa Haendel, Ph.D[1]

[1]Oregon Health & Science University, Portland, OR; [2]Johns Hopkins University, Baltimore, MD; [3]Oregon State University, Corvallis, OR; [4]Tufts University, Boston MA; [5]The University of Arizona, Tucson, AZ; [6]National Center for Advancing Translational Sciences, Bethesda, MD; [7]Lawrence Berkeley National Laboratory, Berkeley, CA

Background

Rapidly identifying treatments, care strategies, longer-term outcomes, and the biological mechanisms underlying COVID-19 as a new disease is a challenging problem that necessitates interdisciplinary teams of experts. Team science facilitates sharing knowledge to "produce exceptionally high-impact research[1]" and solve problems beyond the scope of one field[2]. This includes implementation of shared concept sets, analyses and other tools supporting analyses.[3,4] This pandemic presented an urgent need and opportunity to demonstrate how open team science can more rapidly answer critical questions and develop patient care strategies.

It is critical that data provenance and informatics methods applied to observational healthcare data be made fully transparent and reproducible. Major manuscript retractions[5,6] that have changed the course of ongoing clinical trials have made the need for reproducibility even more pressing. To expedite reproducible, transparent, collaborative analytics for COVID-19 research, the National Center for Data to Health (CD2H) quickly established the National COVID Cohort Collaborative (N3C). The N3C[7] is a broad partnership across US academic medical centers (almost 70 as of this writing) to harmonize electronic health record data.

Methods

The N3C has implemented a uniquely open team science approach to navigate the societal, technical, regulatory, and clinical obstacles. Engagement of the research networks was the first step to building a transparent and diverse team with the necessary knowledge. This included subject matter experts of Accrual to Clinical Trials Network (ACT), National Patient-Centered Clinical Research Network (PCORnet), Observational Health Data Sciences and Informatics (OHDSI), and TriNetX. Governance was established through shared decision making between the NIH and the community around data transfer, data access, data use, code of conduct, guiding principles and publication and attribution. The N3C Data Enclave, which is the analytical platform for the data, maintains full provenance for all data and analysis results so they can be shared with full attribution. Because the Enclave contains data that is the result of an advanced ingestion and harmonization pipeline, statistical code captured as a by-product of the enclave functionality, and made available to researchers, can support reuse of analyses against different datasets in addition to supporting reproducibility of evidence[3,4,8].

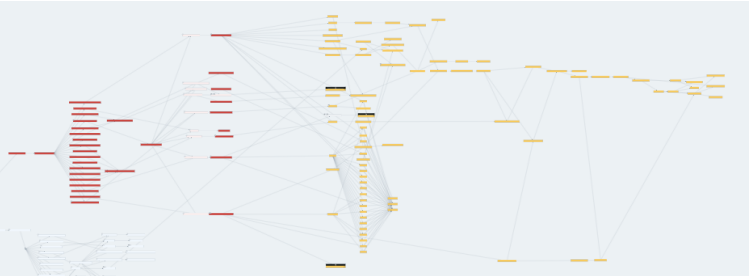
Results

Collaboration. Six workstreams were rapidly formed within the consortium: Data Partnership & Governance, Phenotype and Data Acquisition, Data Ingestion and Harmonization, Collaborative Analytics, Synthetic Data, and Implementation. The workstreams include 4 major subgroups and more than 20 self-organizing domain teams with expertise in clinical domains, statistics, data science, the platform/training, and machine learning. Within eight months of its establishment, the N3C effort has come to involve over 1,200 members representing more than 300 institutions across 47 states nationwide, and 14 foreign countries. Almost 200 of these members collaborated to author the marker manuscript [7] alone. Moreover, these members represent disciplines as diverse as clinical medicine and social science, statistics, librarianship, public health, and computer science all working together to address the complex socio-technical landscape of data access, preparation, and analyses.

Provenance and Attribution. Software, code sets, and other resources deployed within the N3C Enclave are fully attributed using required ORCIDs and the Contributor Attribution Model (https://contributor-attribution-model.readthedocs.io/en/latest/). The N3C Enclave also contains a sophisticated graph model for tracking all actions by all users for security purposes as well as the robust attribution of all contributions to any given resource, workflow, or result. Enclave reports provide a full list of ORCIDs for all participants based on the provenance graph. Reports

and other dataset descriptors and concept set definitions are also included in the Zenodo N3C community for community feedback and public availability.

Reproducibility. The breadth and complexity of data and analytic resources in N3C is larger than can be easily understood and fully reflected at the outset of any given N3C project. Initial exploration by project teams will reveal data development needs and required changes to analytic plans. To promote the clarity of connection between projects and the data and code used to implement them, N3C will support and promote a two-stage project structure. The first phase will encompass all exploration required to fully specify a project that is machine-executable against data known to meet project requirements and using analytics that are appropriate to the project aims. Data exploration in this phase will use either simulated data or a reserved subset of the real clinical data. At the end of the first phase the project plan will be revised. This revised and fully specified project will be one executed in the second phase against full or the real data depending on the code used in the first phase to produce the results of record for the project.



Provenance graph of an N3C workflow, showing resources and their connectivity to the people who created them.

Discussion

N3C mobilized the research community to come together to develop an informatics infrastructure to tackle COVID-19 research. Collaborating in an open, transparent environment has pushed the traditional boundaries to allow innovative solutions that allow team science to evolve. There has been tremendous engagement and participation across the United States, including several countries. The initial phase has been very successful, but there are still challenges to overcome, such as determining use of shared project management tools and the best methods for communication. For example, there are multiple communication channels, but information gaps across the collaborative and within the CTSA organizations exist. This effort is rapidly advancing to produce knowledge about COVID-19.

Acknowledgment

The work reflects the collaboration of many from across the N3C, CTSAs, NCATS, and the many organizations and companies whose members provided ongoing input, support, and participation. This work has been funded through the National Center for Advancing Translational Sciences, National Institutes of Health, under award number U24 TR002306.

References

1. Wuchty S, Jones BF, Uzzi B. The increasing dominance of teams in production of knowledge. Science. 2007;316: 1036–1039. doi:10.1126/science.1136099
2. Porter A, Rafols I. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics. 2009;81: 719–745. Available: https://akjournals.com/view/journals/11192/81/3/article-p719.xml
3. Code share. Nature. 2014;514: 536. doi:10.1038/514536a
4. Barnes N. Publish your computer code: it is good enough. Nature. 2010;467: 753. doi:10.1038/467753a
5. Mehra MR, Ruschitzka F, Patel AN. Retraction-Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. Lancet. 2020;395: 1820. doi:10.1016/S0140-6736(20)31324-6
6. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621. N Engl J Med. 2020;382: 2582. doi:10.1056/NEJMc2021225
7. Haendel M, Chute C, Gersing K, N3C Consortium. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. J Am Med Inform Assoc. 2020. doi:10.1093/jamia/ocaa196
8. Simon GE, Richesson R, Weinfurt K, Hernandez AF, Curtis LH. Statistical Code for Clinical Research Papers. Annals of Internal Medicine. 2019. p. 80. doi:10.7326/l18-0613

Communicating Results of Predictive Models to Patients

Colin G. Walsh, MD, MA[1,2,3], Mollie M. McKillop, PhD, MPH [4], Patricia Lee, MLS[5], Joyce W. Harris, MA[1], Christopher Simpson, MA[1], Laurie Lovett Novak, PhD, MHSA[1]
[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN; [2]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; [3]Department of Psychiatry and Behavioral Sciences, Vanderbilt University Medical Center, Nashville, TN; [4] IBM Watson Health, Cambridge, MA; [5] Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN

Introduction

Prognostication is fundamental to medicine. The output of rules, risk scores, and increasingly complex predictive algorithms have added to our prognostic capabilities. Widespread development of predictive algorithms and interest in them across academia and the healthcare industry has increased their use. We currently rely on humans to deliver poor and complex prognoses to support patients and aid in making sense of this information. Novel ways of collecting, interpreting, and prognosticating using patient data increasingly means that patients receive prognostic information outside of traditional face-to-face interactions between provider and patient. This information flow is increasingly being scaled as algorithms may provide new risk estimates, recommendations around estimates, and automated decision making. Yet it is unclear how best to communicate results of predictive analytics.

We sought to understand the state of the literature around communication of results of predictive algorithms from providers and provider organizations to patients. Because of the diverse nature of the potential literature in this space and the hypothesized lack of rigorous gold-standard studies in this domain, we structured our study as a scoping review informed by the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) Statement[1].

Methods

Primary data from the medical literature investigation was identified and extracted. We tested three distinct concepts in our legacy PubMed search between September 9, 2019, and November 12, 2019: 1) communication; 2) predictive analytics, artificial intelligence, deep learning, big data, machine learning, risk scoring; and 3) communication between the patient or caregiver and doctor/physician/health care provider or patient/caregiver and health system communication. Title and abstract screening were performed by all authors for eligibility for full text review by two blinded independent reviewers. If two reviewers did not agree, blind adjudication was conducted by a pre-assigned third reviewer. Studies published on or after the year 2000 in English and for which the title or abstract indicated evaluation of communication of algorithmic results to patients in the rest of the manuscript were included for full text review. Final selection inclusion criteria required: a primary endpoint on patient communication metric; a predictive analytic tool; and either a decision aid or shared decision-making between patient and provider. Figure 1 shows how we identified articles for review.

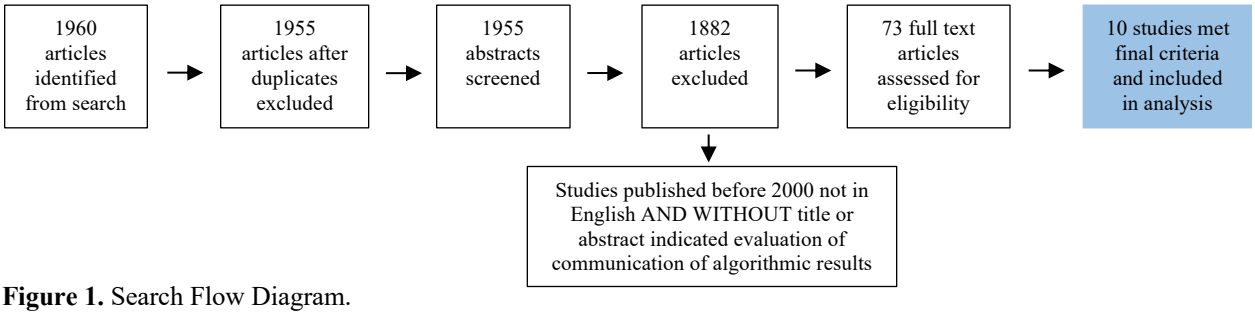


Figure 1. Search Flow Diagram.

Results

Few studies (N=10) examined communication related to the results of predictive analytics to patients. Included studies reviewed by application area (where some papers reviewed fit multiple categories) are summarized in Figure 2.

Figure 2. List of papers reviewed by topic.

Paper	Disease Prevention Aids	Treatment Decision Aids	Medication Harms Reduction	Presentation of Cardiovascular Risk Information
Sheridan et al. 2006	✓			✓
Bonner et al. 2018	✓			✓
Asimakopoulou et al. 2008	✓			✓
Grover et al. 2007	✓			✓
Skinner et al. 2005	✓			
Persell et al. 2015		✓		✓
Flynn et al. 2015		✓		
Mühlbauer et al. 2019		✓		
Hakone et al. 2016		✓		
Fried et al. 2017			✓	

We found few studies that explicitly discussed patient-provider communication and prognosis prediction together. Prognostic decision aids increased communication between patients and providers, yet data on measurable behavior change and health outcomes were mixed. Of the selected studies themes for communicating predictive analytics included contextualizing results to add to perceived credibility, understanding, and satisfaction with model output. Design choices such as pictograms to convey probabilistic information are acceptable in people with lower literacy and may facilitate improvement in acquisition of specific probabilistic information and general impression of risk scores.

Conclusion

Results were communicated to patients across a wide range of application areas. More research in this domain is needed. The gaps identified here might inform areas of further inquiry. In the absence of robust research, some considerations are as follows. We identify a need for determining an optimal way of choosing a risk score given patients' and providers' preferences more often dictate behavior regardless of model output. To bridge this gap, we should further embrace design thinking. Overcoming education gaps in literacy and numeracy among patients and providers will support the successful adoption of predictive systems in health care.

References

1. Liberati A, Altman D, Tetzlaff J, Mulrow C, Gøtzsche P, Ioannidis J, Clarke M, Devereaux P, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of clinical epidemiology. 2009 Oct 1;62(10):e1-34.

An Evaluation of Clinical Natural Language Processing Systems to Extract Symptomatic Adverse Events from Patient-Authored Free-Text Narratives

Yue Wang, PhD[1], David Gotz, PhD[1], Ethan M. Basch, MD, MSc[1], Arlene E. Chung, MD, MHA, MMCi[1]
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Introduction

Symptomatic adverse events (AEs) such as nausea are common among patients enrolled in cancer clinical trials. Historically, this information has been collected and reported into research databases by clinical staff using a set of AE grading criteria maintained by the National Cancer Institute (NCI) called the Common Terminology Criteria for Adverse Events (CTCAE). In NCI's Patient-Reported Outcomes version of CTCAE (PRO-CTCAE) software system, patients can also provide supplemental free-text narratives about their AEs. 58% of patients submit supplemental AE information when given this opportunity[1]. More importantly, there was not considerable overlap between supplemental AEs submitted by patients and those elicited in trial-specific questionnaires, providing evidence for the value of collecting free-text, patient-authored AEs[1]. In our prior work, we also found that the majority (88%) of the symptom concepts within patient narratives could be manually mapped to the Medical Dictionary for Regulatory Activities (MedDRA), which is the standard lexicon for reporting AEs to regulatory agencies such as the FDA[1]. However, the manual process of mapping symptom concepts to lexicons is labor-intensive and limits the widespread collection of free-text AEs. Clinical natural language processing (NLP) has the potential to accelerate recognition and mapping of these symptom concepts and could enable real-time extraction, mapping, and reporting of patient-authored AEs. Off-the-shelf NLP systems, if high-performing, could allow for systematic text processing to be applied, but have not previously been examined for patient-authored AEs. Thus, the objective of this study was to evaluate performance of four widely used clinical NLP systems in extracting symptom concepts from patient-authored free-text AE narratives.

Methods

To determine system performance for extracting AE concepts, four systems that use algorithms ranging from basic pattern matching to deep neural networks were evaluated. Each system was used to map symptom concepts from narratives back to a MedDRA concept, when available, since MedDRA is used for regulatory reporting.

1. **BioPortal**:[2] BioPortal provides web access to a library of biomedical ontologies, and has a RESTful API that annotates documents using terms in user-specified ontologies. The underlying mechanism is multi-word string matching. We specified MedDRA as the target ontology, and used the RESTful API to annotate documents.
2. **MetaMap**:[3] MetaMap employs a set of pattern-matching rules to recognize UMLS concepts within text. The online batch service annotates documents with CUIs. UMLS Metathesaurus was then used to convert each CUI to a MedDRA concept, if available.
3. **cTAKES**:[4] cTAKES assembles a pipeline of pattern-matching and classical machine-learning NLP modules that leverage rich linguistic and semantic information for text analysis. We configured the pipeline to recognize "SignSymptomMention" and "DiseaseDisorderMention." cTAKES generates a CUI for each mention and then each CUI was converted to a MedDRA concept using UMLS Metathesaurus, if available.
4. **Amazon Comprehend Medical (ACM)**:[5] ACM is an NLP service from Amazon Web Services. The entity recognition module employs deep bidirectional long-short term memory (BiLSTM) networks. It can map medical entities to two ontologies (ICD-10-CM or RxNorm). The system was configured to recognize disease and symptom-related entities and to map them to ICD-10-CM codes. The ICD-10-CM codes were then converted to CUIs, which were converted to MedDRA concepts using UMLS Metathesaurus, if available.

Evaluation corpus. A random sample of 100 free-text narratives (documents) were selected from a corpus used in a prior PRO-CTCAE study[1]. Each narrative has symptomatic AEs described in a patient's own words without any character limits. Symptom concepts in each narrative were coded and mapped to MedDRA concepts by two physicians with an adjudicator with 96% inter-rater agreement. On average, a document had 3.4 words and 1.1 symptom mentions; a symptom mention had ~2.3 words. 85% of symptom mentions could be mapped to MedDRA.

Task definition. Given a free-text AE document, the NLP task can be decomposed into two subtasks: 1) concept recognition to identify text spans (each text span consists of one or more words) that mention symptomatic AEs within the document (defined as *symptom mentions*), and 2) concept normalization to map each symptom mention to a corresponding MedDRA concept, as represented by the preferred term (PT) or *none* if no MedDRA concept matched that symptom mention. Given a document as input, we defined the expected output from an NLP system as a set of symptom mentions that were each associated with a MedDRA concept. If a system generated overlapping symptom

mentions (e.g., "rectal bleeding" and "bleeding"), all of them were considered. If a system generated a list of PTs for a symptom mention ranked by prediction confidence (e.g., MedDRA PT 10061525 and 10023643 for "lacrimal disorder"), only the top result was considered. Relaxing the match to "anywhere in the list" did not improve performance substantially (< .02 F_1 increase for all systems, data not reported). Both *strict* and *relaxed* text match conditions were used in the concept recognition subtask. Micro-averaged precision (*P*), recall (*R*), and F_1 score were evaluation metrics for both subtasks.

Results

For the concept recognition subtask, all systems had low precision, recall, and F_1 score under the *strict text match* condition, while the metrics were overall better for *relaxed text match* (Table 1). ACM performed the best with the highest F_1 score in both matching conditions. For the concept normalization subtask, all systems had low precision, recall, and F_1 score for mapping symptom mentions to MedDRA concepts.

Table 1.* Performance across clinical NLP systems by subtask.

Systems	Concept Recognition Subtask						Concept Normalization Subtask					
	Strict Text Match			*Relaxed Text Match*			*Strict Text Match*			*Relaxed Text Match*		
	P	***R***	***F*$_1$**	***P***	***R***	***F*$_1$**	***P***	***R***	***F*$_1$**	**P**	**R**	***F*$_1$**
BioPortal	0.47	0.30	0.37	0.99	0.60	0.74	0.45	0.28	**0.34**	0.98	0.37	0.53
MetaMap	0.37	0.58	0.45	0.76	1.00	0.86	0.27	0.37	0.31	0.60	0.49	0.54
cTAKES	0.38	0.39	0.38	0.98	0.83	0.90	0.33	0.32	0.33	0.96	0.47	**0.63**
ACM	0.60	0.53	**0.56**	0.99	0.92	**0.95**	0.29	0.15	0.20	0.96	0.19	0.32

*Bold indicates the best performing system in terms of F_1 score for each subtask.

Discussion

Focusing on concept normalization, even the best performing system had low performance (strict: 0.34 F_1; relaxed: 0.63 F_1). Similar results were observed in recent shared tasks on extracting and then mapping AEs from patient-authored tweets to MedDRA, where the best performance was obtained by a BioBERT-based deep learning system (strict: 0.34 F_1; relaxed: 0.43 F_1)[6]. This suggests that the task of mapping patient-authored free-text AEs poses significant challenges for NLP systems as they are designed for clinical text. Under the *strict text match* condition, the performance gap between the two subtasks was due to errors in converting UMLS CUIs to MedDRA PTs. For example, "blood in urine" was mapped to C0018965, which was mapped to MedDRA PT 10018867 but not 10018870. Under the *relaxed text match* condition, the performance gap between the two subtasks implies that partially recognized symptom mentions do not sufficiently describe the actual AE. For example, "pain in nails" (onychalgia) is more specific than "pain," and that specificity is important for regulatory reporting. ACM performed the best for concept recognition, which may be due to the potential benefit of its deep sequence tagging algorithm. Amazon has not published their ICD-10-CM mapping algorithm, but error patterns reveal that ACM may use concept embedding matching as opposed to exact string matching, which led to fuzzy and inaccurate mapping results. Based on low performance across these widely used systems, our research reveals that patient-authored symptomatic AE text is sufficiently different from biomedical literature, clinical notes, and patient forum posts, which are the primary targets of these systems. This research highlights the need for new NLP approaches given the goal is to accurately extract and map AEs from patient free-text narratives to standard lexicons for reporting to regulatory agencies.

References

1. Chung AE, Shoenbill K, Mitchell SA, Dueck AC, Schrag D, Bruner DW, et al. Patient free text reporting of symptomatic adverse events in cancer clinical research using the National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). JAMIA. 2019 Apr;26(4):276-85.
2. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue): W541-5. 2011 Jun 14.
3. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. JAMIA. 2010 May 1;17(3):229-36.
4. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. JAMIA. 2010 Sep 1;17(5):507-13.
5. Bhatia P, Celikkaya B, Khalilia M, Senthivel S. Comprehend medical: a named entity recognition and relationship extraction Web service. arXiv preprint arXiv:1910.07419. 2019 Oct 15.
6. Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, et al. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. SMM4H Workshop & Shared Task. 2019:21-30.

Prediction of Blood-Brain Barrier Permeability to Drugs with FDA Adverse Event Report Derived Embeddings

YiFan Wu, MPH[1], Justin Mower, PhD[2], Devika Subramanian, PhD[2], Trevor Cohen, MBChB, PhD[1]
[1]University of Washington, Seattle, WA; [2]Rice University, Houston, TX

Introduction

Adverse event reports (AER) are used for post-market drug safety surveillance, and have also been used for drug repurposing with the assumption that drugs with similar side-effect profiles may have similar therapeutic effects. Methods that directly encode similarity are attractive for this task, with distributed representations, such as neural word embeddings, improving computational performance and accuracy in a variety of domains. In this study, we applied distributed representations for drugs derived from the Food and Drug Administration AER system (FAERS) using `aer2vec`, a method by which drug embeddings emerge from neural networks trained to predict the probability of adverse drug effects given observed drugs[1] (or vice-versa). These representations have been shown to be effective for pharmacovigilance signal detection, but have not yet been leveraged for drug repurposing. We combined these distributed representations with previously used molecular features to predict the permeability of the blood-brain barrier (BBB) to selected drugs, which is an important step in finding treatments for central nervous system (CNS) conditions. We hypothesized that deriving distributed representations of drugs from FAERS data would improve predictive performance of BBB penetration relative to drug similarity statistics derived from the same source.

Methods

Data/Features: FAERS data were obtained from a standardized edition of the database developed by Banda et al[2]. The dataset spans 2004 to 2015 and was standardized by mapping drug names to RxNorm and outcomes to SNOMED-CT. As was the case with prior work using these data for BBB permeability prediction, the AERs were not limited to those concerning CNS side effects. We generated drug embeddings by training an `aer2vec` model to predict P(AE|drug) on FAERS data (following Portanova et al[1]). We obtained two previously developed reference sets for BBB permeability[3,4], and ensured their mutual exclusivity by removing any drugs from the larger set[4] that were present in the smaller one[3]. The molecular properties of lipophilicity (XlogP) and molecular weight (MW) for the drugs were acquired from PubChem. Following previous work, statistically-derived similarity scores were estimated using the Tanimoto coefficients between pairs of drugs on the basis of the AE they are significantly correlated with[5]. Drug-drug similarity features were calculated by comparing each reference set drug to members of an expert-curated six-drug panel of CNS-penetrant agents: haloperidol, olanzapine, risperidone, fluoxetine, mirtazapine, and nortriptyline[6].

Model/Variants: We adapted the approach of McCoy and Perlis[6], substituting distributed representations for Tanimoto similarity features. We estimated the cosine similarity between panel and reference set drugs using `aer2vec` drug embeddings. We also evaluated the utility of using the drug embeddings directly as features. Following McCoy and Perlis[6], AER-derived features were combined with the two molecular features from PubChem. To correct for differences in scale, we applied Z transformation across all features except for the distributed drug vector representations. Using L2-regularized logistic regression classifiers in 5-fold cross-validation (5CV) configuration with the two reference sets separately, we evaluated whether or not the distributed representations improved performance over prior methods using discrete statistical estimates of drug-drug similarity from FAERS data. To further validate the generalizability of our findings we used a cross-training configuration, training on the larger reference set (N=207), and testing on the smaller (N=187). To distinguish the effects of representation type from those of the number of trainable parameters, we reduced embedding dimensionality from 500 to 6 (the number of agents in the expert-selected panel), either by setting this `aer2vec` parameter when training, or by using Principal Component Analysis. Model performance was evaluated using the standard Area Under the Receiver Operating Characteristic Curve (AUC).

Results

Across both datasets and training configurations, the best performing feature combination leveraged information derived from the `aer2vec` embeddings. Adding statistically-derived Tanimoto similarity features to the molecular features improves performance, which is consistent with previous results[6]. Replacing the six statistically-derived panel

Table 1: 5CV and *cross-training* results. Vector-derived similarity score is the dot product between the panel score and vectors. Mean AUC from 5CV and 95% confidence interval using 500-dimensional `aer2vec` embeddings trained on the FAERS data. (Martin: n=187; Gao: n=207)

Models	5CV Martin et. al. dataset	5CV Gao et. al. dataset	Train:Gao; Test:Martin
MW + xlogP	0.711 (0.705, 0.717)	0.738 (0.733, 0.742)	*0.728*
MW + xlogP + discrete statistical similarity score	0.758 (0.751, 0.765)	0.779 (0.773, 0.785)	*0.755*
MW + xlogP + vector-derived similarity score	0.759 (0.756, 0.762)	0.818 (0.813, 0.824)	*0.783*
MW + xlogP + 500-dimensional vectors	**0.807** (0.794, 0.819)	**0.867** (0.862, 0.873)	***0.788***

similarity scores with their embedding-derived counterparts further improves performance. Finally, we observe an additional increase in AUC when replacing similarity scores with the full 500-dimensional `aer2vec` vectors. Results are consistent across the two datasets, as well as in the cross-training configuration (Table 1).

Improvements in performance are retained when constraining embeddings to 6 dimensions, demonstrating superior performance at the same number of trainable parameters (Table 2). As with 500 dimensional embeddings, the best performing models all include embedding-derived features, and models including the embeddings themselves, or their reduced dimensional approximations, showed superior performance to those based on similarity to the six panel drugs.

Table 2: Mean 5CV AUC and *cross-training* results with 6-dimensional vectors within two reference sets

Model	5CV Martin et. al. dataset	5CV Gao et. al. dataset	Training:Gao;Testing:Martin
MW + xlogP + Statistically-derived Similarity Score	0.758 (0.751, 0.766)	0.774 (0.770, 0.778)	*0.755*
MW + xlogP + 6-dimensional vector-derived panel scores	0.775 (0.767, 0.784)	0.849 (0.846, 0.853)	*0.792*
MW + xlogP + 6 principal components	0.796 (0.786, 0.806)	**0.872** (0.868, 0.876)	**0.827**
MW + xlogP + 6-dimensional vectors(L1-regularized)	0.794 (0.783, 0.804)	0.861 (0.853, 0.868)	*0.817*
MW + xlogP + 6-dimensional vectors(L2-regularized)	**0.803** (0.797, 0.808)	0.852 (0.847, 0.858)	*0.820*

Discussion

Our results aligned with previous work in that adding discrete statistically-derived Tanimoto scores to molecular features improved the AUC of logistic regression models.[5,6] However, cosine similarities derived by comparing `aer2vec` derived distributed representations proved more effective than the Tanimoto scores, perhaps on account of `aer2vec` embeddings' additional capacity to generalize. Across all configurations, these advantages in performance were retained after reducing dimensionality from 500 to 6, which shows they are not attributable to having more trainable parameters. Reduced-dimensional models actually outperform full models in some configurations, perhaps on account of noise reduction. Our results indicate that `aer2vec` distributed representations carry information of value for predicting therapeutic effects in addition to their established utility for anticipating harmful side effects.

Acknowledgements

This work was supported by the U.S. National Library of Medicine Grant (R01LM011563) and by the University of Washington Biomedical and Health Informatics Training Grant 5T15LM007442-18.

References

1. J. Portanova, N. Murray, J. Mower, D. Subramanian, and T. Cohen. aer2vec: Distributed Representations of Adverse Event Reporting System Data as a Means to Identify Drug/Side-Effect Associations. *AMIA Annual Symposium Proceedings*, forthcoming, Jul 2019.
2. J. M. Banda, L. Evans, R. S. Vanguri, N. P. Tatonetti, P. B. Ryan, and N. H. Shah. A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data*, 3:160026, 05 2016.
3. Z. Gao, Y. Chen, X. Cai, and R. Xu. Predict drug permeability to blood-brain-barrier from clinical phenotypes: drug side effects and drug indications. *Bioinformatics*, 33(6):901–908, 03 2017.
4. I. F. Martins, A. L. Teixeira, L. Pinheiro, and A. O. Falcao. A Bayesian approach to in silico blood-brain barrier penetration modeling. *J Chem Inf Model*, 52(6):1686–1697, Jun 2012.
5. N. P. Tatonetti, G. H. Fernald, and R. B. Altman. A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. *J Am Med Inform Assoc*, 19(1):79–85, 2012.
6. T. H. McCoy and R. H. Perlis. A tool to utilize adverse effect profiles to identify brain-active medications for repurposing. *Int. J. Neuropsychopharmacol.*, 18(3), Feb 2015.

Sepsis Prediction in the General Ward Setting

Sean C. Yu, MS[1,2]; Aditi Gupta, PhD[1]; Kevin Betthauser, PharmD[3]; Marin H. Kollef, MD[4]; Albert M. Lai, PhD[1]; Philip R.O. Payne, PhD[1]; Andrew P. Michelson, MD[1,4]

[1]Institute for Informatics, Washington University School of Medicine, St. Louis, MO
[2]Department of Biomedical Engineering, Washington University in St. Louis, MO
[3]Department of Pharmacy, Barnes-Jewish Hospital, St. Louis, MO
[4]Division of Pulmonary and Critical Care, Washington University School of Medicine, St. Louis, MO

Introduction

Sepsis, defined as life-threatening organ dysfunction caused by a dysregulated host response to infection, is implicated in one-third to one-half of all inpatient deaths, and was found to be the most expensive condition treated in U.S. hospitals, accounting $38.2 billion or 8.8% of all hospitalization costs.[1-3] Delay in appropriate therapy for sepsis is associated with increased mortality, emphasizing the need for early identification.[4,5] Nearly all prior work on sepsis prediction has focused on the data-rich Intensive Care Unit (ICU) setting, however, it was found that patients who develop sepsis in the general ward setting have worse outcomes compared to those in the emergency department or ICU.[6] Thus, the objective of this research was to predict sepsis six-hours ahead of onset in the data-sparse general ward setting.

Methods

All analysis was conducted using Electronic Health Record (EHR) and administrative claims data from Barnes-Jewish Hospital and the Washington University School of Medicine in St. Louis, a large, academic, tertiary-care referral center. Eligible patients were ≥18 years of age, admitted to the hospital as inpatients or observation status between 1/1/2012 and 6/1/2018, had at least 3 of each vital sign within 24 hours of prediction and had a basic or comprehensive metabolic panel and complete blood count within 24-hours of prediction. Only the first occurrence of sepsis per encounter per patient was included in the analysis.

Sepsis was defined according to the Sepsis-3 consensus definition with modification to include only intravenous antibiotics. Time of onset was determined as the earlier of antibiotic order start time or culture collection. Patients were excluded if they were admitted to the Psychiatry or Obstetrics services, due to highly variable rates of physiologic data collection, or if there were no billing code, vital sign, laboratory, service, room, or medication data. To identify those most likely to benefit from this analysis, patients were also excluded if they were in the ICU 24 hours before prediction time or if they received antibiotics within 48 hours prior to the prediction time. Patients who underwent a surgical procedure within the preceding 72 hours were also not eligible for inclusion to avoid conflation of post-surgical patient status and sepsis.

Features were generated from demographic, service, comorbidity, lab result, vital sign, and medication data. Only information available prior to prediction time was used, e.g., comorbidities were based solely on data from prior admissions. Categorical variables were one-hot encoded. Numerical variables were Box-Cox transformed and standardized. The feature matrix was then split into train/test sets (80%/20%), and the training set was used for random search hyperparameter optimization and training of an XGBoost model.[7] Performance on the test set was evaluated through generation of receiver operating characteristic curve (ROC), precision recall curve (PRC), and the area under each (AUROC & AUPRC). To determine an optimal threshold, a threshold plot was generated in which various traditional performance metrics (precision, recall/sensitivity, specificity, and F1 Score) were plotted against the threshold value. Feature importance was assessed using SHAP values. To estimate the implemented performance of the prediction model, a pseudo-prospective trial was performed in which the model was applied hourly on subjects in the test set. This project was approved with a waiver of informed consent by the Washington University in St. Louis Institutional Review Board (IRB #201804121).

Results

Out of 401,235 total inpatient encounters, 54,086 qualified for inclusion, of which 904 (1.7%) were septic. Sepsis patients were slightly older (65.2 [56.1 – 73.6] vs. 60.7 [49.0 – 71.5], $p < 0.01$), more likely to be white (74.6% vs. 64.8%, $p < 0.01$), and had a higher comorbidity index (11.1 ± 12.6 vs. 9.9 ± 12.0, $p < 0.01$). Sepsis patients had higher rates of sepsis discharge diagnosis code (23.7% vs. 0.7%, $p < 0.01$), longer length of stay (13.0 [8.1 – 19.5] vs. 3.2 [2.0 – 6.2] , $p < 0.01$), and were significantly more likely to die during hospitalization (14.5% vs. 0.7%, $p < 0.01$).

Table 1. Cohort comparison.

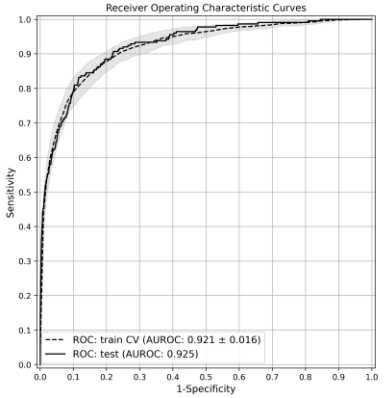
Variable	Total	Sepsis	Non-sepsis	p*
Number of patients, n (%)	54,086 (100.0%)	904 (1.7%)	53,182 (98.3%)	< **0.01**
Age (years), median (IQR)	60.8 (49.2 - 71.6)	65.2 (56.1 - 73.6)	60.7 (49.0 - 71.5)	< **0.01**

Race, n (%)	-	-	-	**< 0.01**
White, n (%)	35,133 (65.0%)	674 (74.6%)	34,459 (64.8%)	**< 0.01**
Black, n (%)	15,639 (28.9%)	163 (18.0%)	15,476 (29.1%)	**< 0.01**
Asian, n (%)	366 (0.7%)	7 (0.8%)	359 (0.7%)	0.876
Other/unknown, n (%)	2,948 (5.5%)	60 (6.6%)	2,888 (5.4%)	0.131
Sex (female), n (%)	25,255 (46.7%)	415 (45.9%)	24,840 (46.7%)	0.657
Body Mass Index, median (IQR)	27.7 (23.7 - 33.0)	27.3 (23.4 - 33.6)	27.7 (23.7 - 33.0)	0.450
Elixhauser index, median (IQR)	8.0 (0.0 - 17.0)	9.0 (0.0 - 18.0)	8.0 (0.0 - 17.0)	**< 0.01**
Length of stay (days), median (IQR)	3.3 (2.0 - 6.5)	13.0 (8.1 - 19.5)	3.2 (2.0 - 6.2)	**< 0.01**
Sepsis discharge diagnosis, n (%)	601 (1.1%)	214 (23.7%)	387 (0.7%)	**< 0.01**
In-hospital mortality, n (%)	486 (0.9%)	131 (14.5%)	355 (0.7%)	**< 0.01**

***p-values generated using chi-squared test for categorical or boolean variables, and Mann–Whitney U test for continuous variables.**

The trained prediction algorithm had a test AUROC of 0.925 and AUPRC of 0.440 (Figure 1). The calibration curve had an r^2 of 0.910 (Figure 2). At the threshold (0.163) which maximized F1 score (0.475), the model had a precision of 0.518, specificity of 0.993, and recall of 0.438. The most important features, according to SHAP values, were maximum respiratory rate, minimum systolic blood pressure, time to prediction time, and median shock index. The pseudoprospective trial revealed that at a tentative threshold of 0.163, the alert fired at least once for 88.9% of sepsis patients and 14.0% of non-sepsis patients. Of those alerted, at the first time of alert, 1167 / 2069 (56.4%) were already being treated (antibiotics, cultures, or ICU transfer within past 48h of alert). Of those who were alerted who were not already being treated, 218 / 902 (24.2%) received sepsis-relevant antibiotics or cultures within 24h of alert.

Figure 1. Receiver Operating Characteristic and Precision Recall Curve



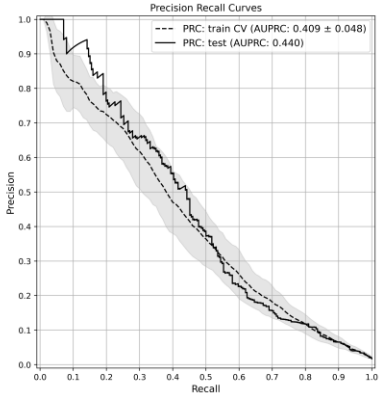


Figure 2. Calibration plot

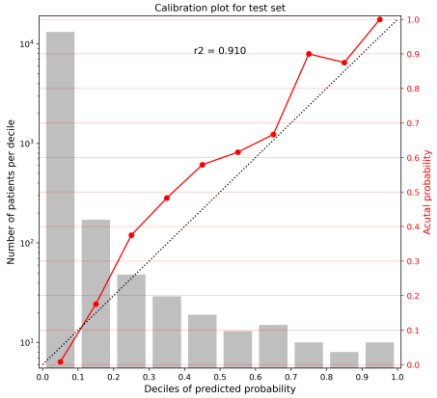
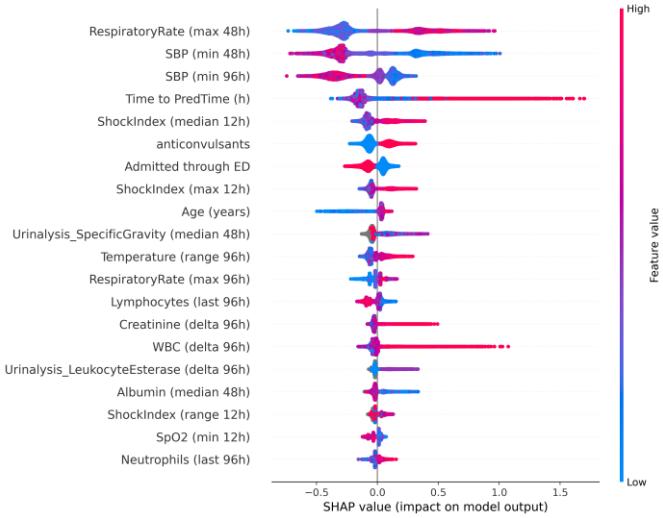


Figure 3. SHAP feature importance plot



Conclusion

Even in the data sparse general ward environment with significant class imbalance, our machine learning prediction model was able to predict sepsis 6-hours ahead of onset with reasonable performance. Next steps include assessment of model in a silent prospective trial and if appropriate, EHR-interface development and ultimately live deployment in clinical practice.

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jama. 2016;315(8):801-10.
2. Rhee C, Jones TM, Hamad Y, Pande A, Varon J, O'Brien C, et al. Prevalence, underlying causes, and preventability of sepsis-associated mortality in US acute care hospitals. JAMA network open. 2019;2(2):e187571-e.
3. Liang L (AHRQ) MBIWH, Soni A (AHRQ). National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2017. HCUP Statistical Brief #261. Agency for Healthcare Research and Quality, Rockville, MD. 2020.
4. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine. 2006;34(6):1589-96.
5. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, et al. Time to treatment and mortality during mandated emergency care for sepsis. New England Journal of Medicine. 2017;376(23):2235-44.
6. Levy MM, Dellinger RP, Townsend SR, Linde-Zwirble WT, Marshall JC, Bion J, et al. The Surviving Sepsis Campaign: results of an international guideline-based performance improvement program targeting severe sepsis. Intensive care medicine. 2010;36(2):222-31.
7. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016: ACM.

Developing an ETL Tool for Converting PCORnet CDM into OMOP CDM

Yue Yu, PhD[1], Nansu Zong, PhD[1], Andrew Wen, MS[1], Sijia Liu, PhD[1], Daniel J. Stone, BS[1], Alanna M. Chamberlain, PhD[1], Davera Gabriel, RN[2], Christopher G. Chute, MD, DrPH[2], Nilay Shah, PhD[1], Guoqian Jiang, MD, PhD[1]

[1]Mayo Clinic, Rochester, MN, USA; [2]Johns Hopkins University, Baltimore, MD, USA

Abstract

The objective of the study is to develop and evaluate an extract, transform, load (ETL) tool for converting the National Patient-Centered Clinical Research Network (PCORnet) common data model (CDM)-based database into the Observational Medical Outcomes Partnership (OMOP) CDM. We also evaluated the ETL process using a real-world medical device dataset. The preliminary evaluation demonstrated promising performance of the data transformation. The ETL tool could facilitate the exchange, pooling, sharing, or storing of clinical data between PCORnet CDM and the OMOP CDM across multiple institutions.

Introduction

The implementation of a common data model (CDM) could "standardize and facilitate the exchange, pooling, sharing, or storing of data from multiple sources"[1] and achieve effective data integration. The most commonly used CDMs regarding medical studies include the National Patient-Centered Clinical Research Network (PCORnet) CDM, the Observational Medical Outcomes Partnership (OMOP) CDM, the Sentinel CDM, and the Informatics for Integrating Biology and the Bedside (i2b2) Star Schema. However, healthcare institutions usually don't support all of these different CDMs. Therefore, it is important to develop tools to facilitate data transformation and interoperability between different CDMs. Some studies have attempted to develop extract, transform, load (ETL) tools to transform different CDMs[2-3]. Nevertheless, to the best of our knowledge, no tooling has been designed to convert the PCORnet CDM into the OMOP CDM directly. In this study, we collaborated with National COVID Cohort Collaborative (N3C)[4-5] to develop an ETL tool that could transform the PCORnet CDM format data into the OMOP CDM. We also evaluated the ETL process to demonstrate the feasibility of the conversion tool.

Methods

Using the PCORnet to OMOP mapping dictionary created by N3C[6], we defined the structure mapping by choosing the appropriate tables/fields from the OMOP CDM (v6.0) for tables/fields in the PCORnet CDM (v5.1). Figure 1 shows the details of table level structure mapping. A total of 12 OMOP CDM tables were mapped with 18 PCORnet tables. The field level structure was also designed between these tables. Then, the ETL tool was built according to the structure mapping using SQL scripts. The ETL tool was designed to achieve three data transformation tasks: 1) value transformation; 2) rule-based transformation; and 3) standard concept code mapping. For the value transformation, we considered the data inconsistency such as the data type and null constraint which may be different for the same data elements between two CDMs. We also established some rules to handle those required fields in OMOP that do not have a counterpart in PCORnet. For example, there is no drug exposure end date information in the PCORnet DISPENSING table. Regarding the concept code mapping, if these codes were from general medical terminologies such as ICD, LOINC, RxNorm, CPT, etc., we mapped those codes to the referred standard codes in OMOP vocabulary. For some concepts which are not from shared medical terminologies such unit concepts, we used a string matching approach to find the mapping between the units and OMOP standard vocabulary. if these codes were specific codes defined in the PCORnet CDM valuesets such as status, type or source information, a manual mapping is required. We also conducted an experiment to evaluate the robustness of the ETL tool by validating the data mapping and conversion process using a real-world

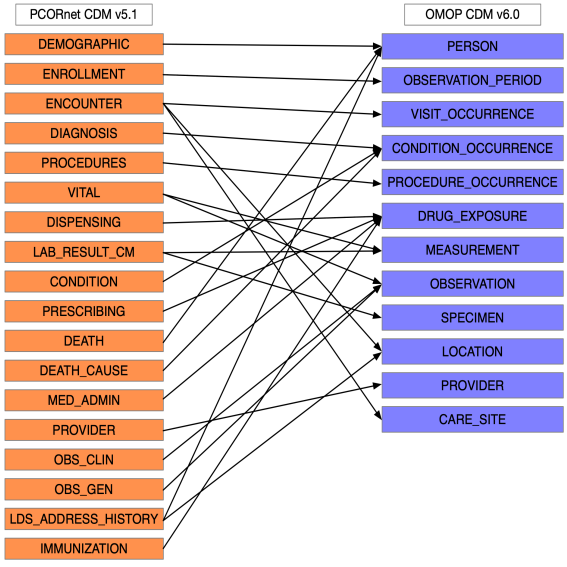


Figure 1. Table level mapping between PCORnet CDM and OMOP CDM.

medical device dataset. For the experiment, we built a cohort that includes randomly selected 100 patients with ventricular tachycardia (VT) and atrial fibrillation (AF) who received cardiac catheter ablation. Then, the related records of the patients were collected from PCORnet CDM and converted into OMOP CDM. The ETL performance evaluation was conducted using the mapping rate representing the proportion of the standard concept codes in PCORnet with the OMOP concept ids successfully identified. For example, the ICD-10 code "I48.0|Paroxysmal atrial fibrillation" has a corresponding OMOP concept_id 45581776 identified. We also designed a query for both CDMs using the same patient identification algorithm to classify the patient into VT and AF subgroups.

Results

Table 1 shows the mapping rate that represents the proportion of the standard concept codes in PCORnet CDM with the OMOP concept ids successfully identified. The mapping rate for all of the concept-related fields was above 97.9%, and 3 of 7 fields achieved a mapping rate of 100%.

Table 1. Concept code mapping rate between PCORnet CDM and OMOP CDM

PCORnet Field	OMOP Field	Mapping Rate	PCORnet Field	OMOP Field	Mapping Rate
PROVIDER_SPECIALTY_PRIMARY	specialty_concept_id	100% (5,222/5,222)	RXNORM_CUI, NDC, MEDADMIN_CODE, VX_CODE	drug_concept_id	97.9% (69,229/70,716)
FACILITY_TYPE	place_of_service_concept_id	100% (62/62)	OBSCLIN_RESULT_SNOMED, OBSGEN_CODE, SMOKING, TOBACCO, TOBACCO_TYPE	observation_concept_id	100% (11,268/11,268)
DX, CONDITION, DEATH_CAUSE	condition_concept_id	99.7% (41,289/41,408)			
PX	procedure_concept_id	99.8% (40,718/40,795)			
LAB_LOINC	measurement_concept_id	99.9% (121,920/121,974)			

Table 2 shows the patient subgroup identification results between two CDMs. We used the diagnosis codes to identify the patient subgroup. The numbers of the patient subgroups identified are the same by different CDMs.

Table 2. Patient subgroup identification results between PCORnet CDM and OMOP CDM

CDM	Number of AF cases		Number of VT cases	
	Paroxysmal AF	Persistent AF	Ischemic VT	Non-Ischemic VT
PCORnet CDM	58	20	21	10
OMOP CDM	58	20	21	10

*One patient may have more than one type of diagnosis, so the total identified case number may be more than 100.

Discussion

In this study, we developed an ETL tool to directly convert PCORnet CDM to OMOP CDM. We also validate the ETL tool by evaluating the concept mapping rate and implementing a patient subgroup identification task. The primary evaluation demonstrated a promising performance of our ETL tool in the data transformation. We also found that the main reason why some concepts were unmatched was due to concepts that have not yet been included in the OMOP vocabulary; examples include the RxNorm codes "amoxicillin/clavulanate" (RxCUI = 19711) and "docusate/sennosides, USP" (RxCUI = 1008340). We plan to work with the OMOP community to add those concepts into the OMOP vocabulary. This study has a number of limitations. First, the patient sample size for the evaluation is small, and as our PCORnet database didn't collect any data for the OBS_CLIN and OBS_GEN table, we could not get a comprehensive evaluation result for the observation-related data. Moreover, we only evaluated the concept mapping rate, and we plan to evaluate the mapping accuracy for the matched concepts as an ongoing effort. In addition, the N3C is working on the validation of the PCORnet CDM valueset mappings with the OMOP concept ids, and we will incorporate the validated mappings in our ETL scripts. As the next step, we will address these limitations by designing more comprehensive evaluation methods and using more data from multiple data sources/institutions. The ETL scripts are available at https://github.com/yuey11/PCORnet2OMOP_ETL_tool.

References

1. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016;64:333-341.
2. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. Plos One. 2019;14(2):13.
3. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assn. 2016;23(5):909-15.
4. National COVID Cohort Collaborative (N3C). https://ncats.nih.gov/n3c. Accessed Aug. 13, 2020.
5. Melissa H, Christopher C, Kenneth G. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment [published online ahead of print, 2020 Aug 17]. J Am Med Inform Assoc. 2020
6. Common Data Element and Permissible ValueSet mapping information from source PCORNet version 5.1 to OMOP version 5.3.1. https://github.com/National-COVID-Cohort-Collaborative/Data-Ingestion-and-Harmonization/tree/master/CDMDataMaps/PCORNet2OMOP. Accessed Aug. 13, 2020.

Examining Patterns of Computerized Physician Order Entry in Emergency and Inpatient Heart Failure Management

Yiye Zhang, PhD MS[1], Joel Park MD MS[1,2], Yifan Liu MS[1], Richard Trepp, MD[2,3], Jessica S Ancker MPH, PhD[1], Jyotishman Pathak PhD[1], Kelly M. Axsom MD[2,3], Peter A D Steel MA, MBBS[1,2]

[1]Weill Cornell Medicine, New York, NY; [2]NewYork-Presbyterian Hospital, New York, NY; [3]Columbia University Vagelos College of Physicians and Surgeons, New York, NY

Introduction

Heart failure (HF) is a leading cause of morbidity and mortality affecting more than 23 million patients worldwide.(1) Variations in management and guideline-adherence have been cited as important and potentially modifiable factors that contribute to adverse outcomes among adults with HF.(2) As computerized physician order entry (CPOE) in the electronic health records (EHR) captures the sequences of management decisions made by care provider teams, we studied the sequential patterns in CPOE in the management of HF patients in the emergency department (ED) and inpatient settings. Patterns in order placement were discovered using process mining and machine learning methods. The associations of the found patterns with patient characteristics, use of order sets, and care outcomes were studied using statistical modeling. Findings are intended to better describe and inform the current HF management.

Methods

Study data were obtained from at an urban academic medical center between 2012 to 2018. We included HF patients who were treated on the Medicine service. HF was defined by the ICD-9/10-CM of principal, billing, and discharge diagnoses noted in the EHR data. Age, gender, race, ethnicity, marital status, and preferred spoken language were obtained directly from EHR data. Census-tract level social determinants of health were extracted using home locations in the EHR.(3) Visit information including admission timestamps, discharge timestamps, discharge dispositions, insurance types, and All Patients Refined Diagnosis Related Groups (APR-DRG) were extracted for each visits. We excluded self-pay visits. The ICD-9/10-CM of coexisting conditions were extracted to compute Elixhauser scores. Data on CPOE included order names, order set names if applicable, the hours of order creation, order statuses, the roles of provider who entered the order, and order types. Canceled orders were removed from the study. Care outcomes are defined as in-hospital mortality, discharge location (home vs. not home), all-cause readmission (preventable and unpreventable), and length of stay (LOS).

Using extracted data, we mined sequential patterns in CPOE using methods described in Zhang et al.(4) This is a process mining method which first computes the similarity of the timing and types of orders in the visits using the longest common subsequence across patient pairs. Then, it applies hierarchical clustering to determine the number of underlying clusters of patients who have had similar patterns of CPOE. Upon obtaining cluster membership for each visit, we used statistical models to study the relationships among cluster membership, care outcomes, Elixhauser scores, demographics, insurance status, and social determinants of health.

Results

A total of 1626 visits were in the study cohort. The mean ages for female and male patients were 74.0 (sd=0.49) and 69.7 (sd=0.44), respectively. The racial distribution of White, Black, Other, Asian, and Unknown was 39.0%, 20.6%, 15.1%, 12.9%, and 12.4%. Hispanics accounted for 15.7% of the cohort. More than 77% of the patients were English-speaking, and 40.3% of the patients were recorded to be married. The distribution of insurance plans: Medicare, Medicaid, and Commercial were 72.8%, 14.4%, and 12.8% respectively. Over 85% of the patients were admitted to the inpatient setting from the ED. There were 965 unique medication orders in the study data. Non-medication orders were placed from 150 unique order sets. The analysis of order pathways identified 3 clusters of HF visits. Table 1 lists the distribution of care outcomes: discharge home, in-hospital mortality, readmission, LOS, and the average Elixhauser scores across clusters. Table 2 shows common medication orders and common comorbid conditions recorded that most represent the 3 clusters. Statistical significance of the difference was tested using ANOVA for numerical variables and Chi-Square test for categorical variables.

Table 1. Distribution of care outcomes and characteristics across clusters

Cluster	Discharge home*	In-hospital Mortality*	Readmission	LOS (days)*	Avg Elixhauser*
1 (N=1052)	43.4%	2.9%	5.6%	5.2(sd=0.21)	8.2 (sd=0.13)
2 (N=419)	26.5%	8.4%	4.3%	15.4(sd=0.71)	9.1 (sd=0.22)
3 (N=155)	18.1%	13.6%	5.2%	35.9(sd=3.42)	9.4 (sd=0.34)

Table 2. Distribution of common medications and comorbidities across clusters

	Cluster 1 (N=1052)	Cluster 2 (N=419)	Cluster 3 (N=155)
Avg Opioid doses/day*	0.40 (sd=0.03)	0.53 (sd=0.05)	0.95 (sd=0.11)
% Order set *	37.1 (sd=3.3)	32.6 (sd=5.1)	31.9 (sd=8.7)
Common medication orders (% patients)	Furosemide (81%), Atorvastatin (44%), Dextrose (44%), Magnesium Sulfate (41%)	Furosemide (81%), Magnesium Sulfate (69%), Magnesium Oxide (59%), Dextrose (55%), Bumetanide (45%), Atorvastatin (44%)	Magnesium Sulfate (88%), Furosemide (83%), Magnesium Oxide (80%), Fentanyl Citrate (67%), Dextrose (67%), Bumetanide (66%)
Common comorbidity (%)	General Cardiology (12.3%), Pulmonary (6.8%), Implantable Cardioverter Defibrillator (4.4%)	General Cardiology (13.9%), Infectious disease (9.1%), Endocrinology (9.0%), Pulmonary (6.4%), Implantable Cardioverter Defibrillator (5.1%)	General Cardiology (7.1%), Infectious disease (6.1%), Implantable Cardioverter Defibrillator (5.1%), Nephrology (4.0%)

We performed multinomial regression analysis with cluster membership as a dependent variable to identify factors that may explain the variation observed. Patient factors that were significantly associated with the cluster membership included race, marital status, and insurance type, while adjusting for factors including age, gender, campus, Elixhauser score and clinical case type. Black race is more likely than White race to be in cluster 3. Decreased percentage of orders placed from order sets was significantly associated with being classified into clusters 2 and 3, while providers who use order sets more frequently than others are also more likely to be in clusters 2 and 3. Clusters 2 and 3 appeared to represent more complex cases and thus may have been more likely to warrant a la carte orders. In addition, we performed a multivariate analysis with discharge disposition (home) as the dependent variable. Cluster membership, cluster 2 (OR=0.41, p-value <0.001) and cluster 3 (OR=0.11, p-value <0.001) were significantly associated with the outcome while controlling for male (OR=1.9 p-value <0.001), age (OR=0.93, p-value <0.001), and Elixhauser scores (OR=0.95, p-value =0.008), and race.

Discussion

The cluster membership was significantly associated with discharge disposition, in-hospital mortality, and LOS. The ordering patterns of medications likely reflect the patients' severity of illnesses within their respective clusters. Cluster 2 demonstrated frequent utilization of bumetanide suggesting a generally more ill cohort of HF patients, which is in contrast to cluster 1 where patients received predominantly furosemide. Cluster 3 was similar to cluster 2 but was distinctly characterized by multiple and varying types of opioids ordered. Potential explanations for increased opioid ordering in cluster 3 are the need for procedural sedation or end of life care but warrant further investigation. Statistical analysis revealed that determinants of cluster membership included not only clinical complexity but also patients' sociodemographic factors. Future work includes comparing order-driven clusters with known HF stage classifications such as New York Heart Association Functional Classification.

References

1. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. Circulation. 2017;135(10):e146-e603.
2. Kumbhani DJ, Fonarow GC, Heidenreich PA, Schulte PJ, Lu D, Hernandez A, et al. Association Between Hospital Volume, Processes of Care, and Outcomes in Patients Admitted With Heart Failure Insights From Get With The Guidelines-Heart Failure. Circulation. 2018;137(16):1661-+.
3. Cantor MN, Chandras R, Pulgarin C. FACETS: using open data to measure community social determinants of health. Journal of the American Medical Informatics Association. 2017;0(0).
4. Zhang Y, Padman R, Patel N. Paving the COWpath: Learning and visualizing clinical pathways from electronic health record data. Journal of biomedical informatics. 2015.

Application of a Novel Sequential Pattern Mining Method to Study Longitudinal Social Determinants of Health through Young Adulthood: The CARDIA Study

Lindsay P. Zimmerman, MPH[1], Donald M. Lloyd-Jones, MD, ScM[1], Kiarri N. Kershaw, PhD, MPH[1], David H. Rehkopf, ScD, MPH [2], Yuan Luo, PhD[1]

[1]Northwestern University, Feinberg School of Medicine, Chicago, IL;
[2]Stanford University, School of Medicine, Palo Alto, CA

Introduction

Recently, there has been greater focus on social determinants of health (SDOH) and their impact on health outcomes. The World Health Organization (WHO) defines SDOH as the "structural determinants and conditions in which people are born, grow, live, work, and age[1]." There are five key domains of SDOH including economic stability, neighborhood and built environment, education, social and community context, and health and health care. The main statistical approaches used in understanding the association of SDOH with various health outcomes are correlation and regression, incorporating one or a limited set of SDOH factors per study on an individual or area-level. These methods have significant statistical and conceptual limitations; they limit our understanding of the complex relationships between the full breadth of SDOH factors, as well as the longitudinal associations between SDOH and health outcomes like cardiovascular health.

Each year, approximately 840,000 Americans die from cardiovascular disease (CVD)[1]. The prevalence of CVD is increasing, and large disparities remain across racial, ethnic, and economic groups[2,3]. To address this increasing burden of CVD, there has been a shift to focus on the public health and preventive strategies to address cardiovascular health (CVH), which is a broader and more positive construct beyond the absence of CVD[4]. CVH, as defined by the American Heart Association, is determined by seven health factors and behaviors and can be measured for all individuals, including those in younger aged groups[4]. However, there remain widespread disparities in CVH, even at young ages, that are poorly understood[5]. SDOH may account for the persistent disparities seen in CVH, but the cross-sectional and longitudinal relationship between SDOH and CVH has not been widely researched. No single variable or domain captures an individual's SDOH. Therefore new, and easily interpretable, methods for studying the exposure patterns and effects of multiple SDOH over time are necessary.

Methods

The primary objective of this work is to identify patterns of SDOH exposure from young adulthood to middle age (across ages 18-45 years) and define exposure subgroups among the total cohort and by race and sex in the Coronary Artery Risk Development in Young Adults (CARDIA) study. CARDIA is an existing, longitudinal biracial cohort study (initial enrollment 1985-1986) with detailed information on cardiovascular risk factors and disease in a geographically diverse sample of young Black and White adults balanced by sex, race, age, and education subgroups[6]. In CARDIA, SDOH have been assessed across nine timepoints for the same participants on both individual and neighborhood levels and across all five domains. The SDOH assessed in CARDIA include measures focused on education, income/wealth, occupation status, family composition, neighborhood, and psychosocial assessments across 30 years of follow-up.

In this context, we will use sequential pattern mining to study how SDOH change over time across young adulthood, and non-negative matrix factorization (NMF) to group the changes in SDOH into meaningfully coherent trends overall and by race. Sequential pattern mining is a type of data mining technique which analyzes sequences of events and identifies recurring subsequences or patterns[7]. NMF is a commonly used unsupervised machine learning method to cluster similar patients and to create meaningful variables from a set of high-dimensional data[8]. We will use Subgraph Augmented Non-negative Matrix Factorization (SANMF) as our primary methodology, developed by Luo (co-author) et al[9]. The primary steps of this analysis are pictured in Figure 1. There will be three main outcomes from this analysis: 1) frequent subgraphs for each SDOH variable independently; 2) overall/data-driven trend groups created by identifying subgraph clusters from all possible subgraphs; 3) domain-specific/theorical trend groups created by identifying subgraph clusters within each SDOH domain.

The application of SANMF to longitudinal SDOH variables is novel. SANMF has been previously applied to other longitudinal scenarios; Luo (co-author) et al. used NMF to group patients admitted to the intensive care unit (ICU) based on their temporal trends in multiple physiologic variables[9]. We will use SANMF to group CARDIA participants

based on their identified SDOH subgraphs. Because we are analyzing count data that is non-negative, SANMF serves as an ideal grouping method by providing easy interpretability with the non-negative constraint.

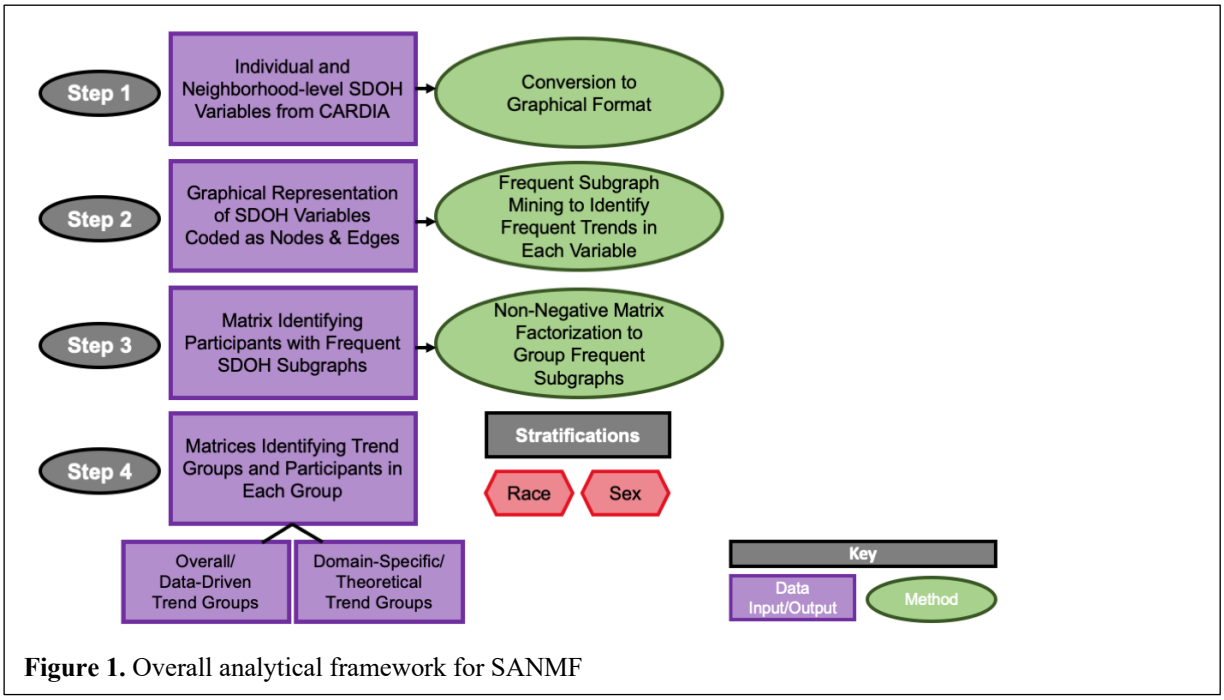


Figure 1. Overall analytical framework for SANMF

Conclusion

By understanding longitudinal SDOH patterns and their clusters at the individual and neighborhood levels, we may be able to generate hypotheses for future work focused on developing timely and multi-component social interventions and policies. The examination of these clusters by race may also help us to better understand the larger structural and social factors underpinning the differences in CVH we see across racial groups. Future work incorporating these SDOH exposures into predictive models for CVH will provide a better understanding of which SDOH exposure patterns may be associated with and predictive of CVH, and at what time during young adulthood. We will present findings from this study at the AMIA 2021 Informatics Summit in March 2021.

References

1. Benjamin EJ, Muntner P, Alonso A, et al. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. *Circulation*. 2019;139(10):e56-e528. doi:10.1161/CIR.0000000000000659
2. Heidenreich PA, Trogdon JG, Khavjou OA, et al. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. *Circulation*. 2011;123(8):933-944. doi:10.1161/CIR.0b013e31820a55f5
3. Graham G. Disparities in cardiovascular disease risk in the United States. *Curr Cardiol Rev*. 2015;11(3):238-245. doi:10.2174/1573403x11666141122220003
4. Lloyd-Jones DM, Hong Y, Labarthe D, et al. Defining and setting national goals for cardiovascular health promotion and disease reduction: the American Heart Association's strategic Impact Goal through 2020 and beyond. *Circulation*. 2010;121(4):586-613. doi:10.1161/CIRCULATIONAHA.109.192703
5. Brown AF, Liang L-J, Vassar SD, et al. Trends in Racial/Ethnic and Nativity Disparities in Cardiovascular Health Among Adults Without Prevalent Cardiovascular Disease in the United States, 1988 to 2014. *Ann Intern Med*. 2018;168(8):541-549. doi:10.7326/M17-0996
6. CARDIA: Coronary Artery Risk Development in Young Adults. https://www.cardia.dopm.uab.edu/
7. Mooney CH, Roddick JF. Sequential pattern mining -- approaches and algorithms. *ACM Comput Surv*. 2013;45(2):1-39. doi:10.1145/2431211.2431218
8. Gillis N. The Why and How of Nonnegative Matrix Factorization. *arXiv:14015226 [cs, math, stat]*. Published online March 7, 2014. Accessed December 19, 2019. http://arxiv.org/abs/1401.5226
9. Luo Y, Xin Y, Joshi R, Celi LA, Szolovits P. Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements. In: *AAAI*. ; 2016. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewFile/11843/11562

Predicting Primary Cancers Based on HL7 Fast Healthcare Interoperability Resources (FHIR) and Resource Description Framework (RDF)

Nansu Zong, PhD[1], Victoria Ngo, PhD[1], Daniel J. Stone, BS[1], Andrew Wen, MS[1], Yiqing Zhao, PhD[1], Yue Yu, PhD[1], Sijia Liu, PhD [1], Ming Huang, PhD [1], Chen Wang, PhD[1], Guoqian Jiang, MD, PhD[1,*]
[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[2]University of California Davis Health, Sacramento, CA, USA
[*]Corresponding author

Introduction

Early detection of primary cancers is a key research area that facilitates optimal treatment in precision oncology. As information technology is largely adopted in healthcare, clinical data utilization offers great promise for disease prediction (1), where diverse types of electronic health record (EHR) data can improve the prediction (2). With the development of the genetic test, gene mutations can be identified, which provides the potential for cancer prediction. On the other hand, conventional predictive models are usually based on bag-of-features (BOF), where the features are treated independently and patterns cannot be fully explored. In contrast, network-based representation of data that embed potential correlations shows great potential (3). Here, we propose a work that leverages Fast Healthcare Interoperability Resources (FHIR) (4) and Resource Description Framework (RDF) (5) to harmonize the phenotypical and genotypical features for cancer prediction. The full version of the study is now under a review process.

Method

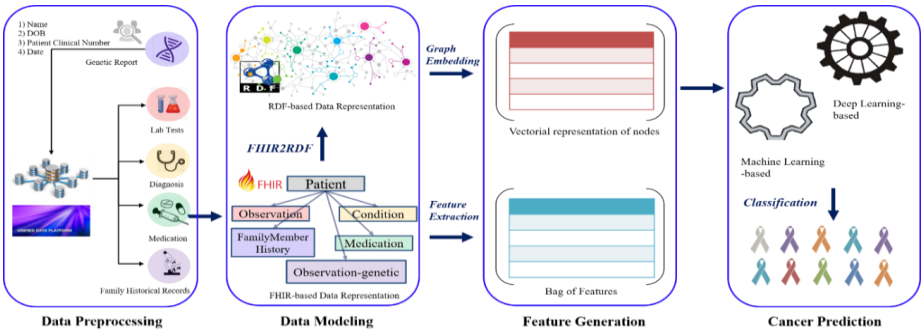


Figure 1. The framework of network-based cancer prediction with FHIR and RDF.

Five types of cancer data, "genetic information", "lab tests", "diagnosis", "medication" and "family historical records", are represented with FHIR resources and further converted to RDF (Figure 1). A graph embedding algorithm, Node2vec, is used to learn the features from the RDF network, and multiple classification models are adopted for the prediction. We have conducted a proof-of-concept study based on a collection of genetic reports of 1,011 cancer patients from Foundation Medicine, and the corresponding EHR data from the Mayo Clinic pan-cancer cohort to predict nine cancers, which are the colon (ICD-9: 153.9), pancreas (157.9), ovary (183), prostate (185), connective and other soft tissue (171.9), thyroid gland (193), breast (174.9), liver (155), and bronchus and lung (162.9).

Results

We designed six tasks to evaluate our work, in which we briefly show four here due to the page limit. We employed seven predictive models, which are Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Deep Neural Network (DNN), Convolutional Neural Network (CNN), Graph Convolutional Networks (GCN). In Task 1, we compared all the combinations of the feature generation and classification methods. Figure 2 shows the best result was achieved by using BOF+Node2vec and Random Forest (AUC = 96.19%).

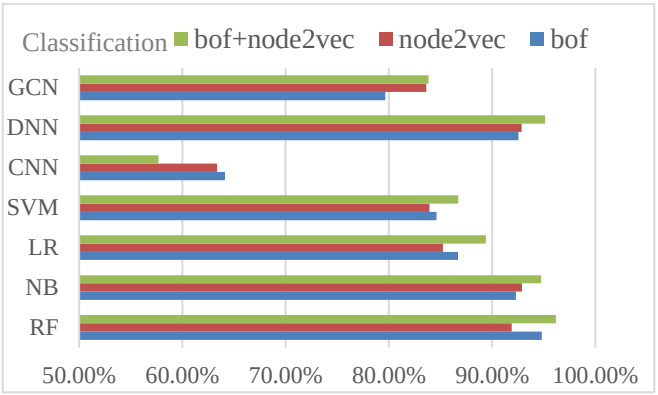


Figure 2. Task 1 the comparison of the combinations of features and classification methods.

In Tasks 2 and 3, we designed evaluation strategies via different validation methods, prediction of primary cancer with internal validation and prediction of the cancer of unknown primary (CUP) cases. For the internal validation, all the cancers except colon, bronchus and lung, show promising results (AUC ROC > 95%). For the CUP cases, only breast, connective and other soft tissue show promising results (AUC ROC > 90%).

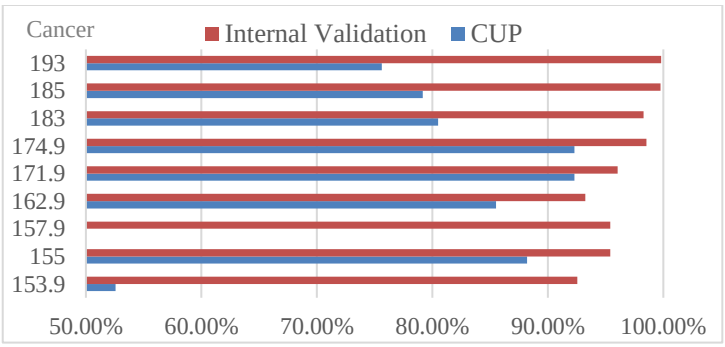


Figure 3. Tasks 2 and 3 prediction for nine cancers based on the internal valuation and CUP cases.

In Tasks 4, we showed how the prediction would be affected by the time-depended data sources. The ideal model "Diagnosis"+ "Medication"+"Lab test"+"Gene" (DML+G) shows impressive prediction results, where it reaches ARUC ROC 91 % at 24 months in advance. Compared to the ideal model (DML+G), the best-performed model "Diagnosis"+ "Medication"+"Lab test" (DML) among all the possible models only achieved promising results (>90 %) within 3 months.

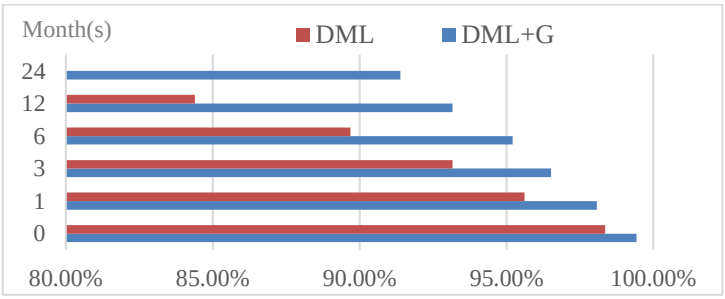


Figure 4. Task 4 time effect of cancer prediction

Discussion and conclusion

In this study, we developed a network-based framework leveraging the FHIR resources and RDF for cancer prediction. Our contributions can be summarized as, 1) exploration of utilizing FHIR RDF technology to provide a network-based solution for the prediction of primary cancers; 2) exploration of the diverse data sources to generate the features in prediction, especially the contribution of leveraging genetic information with phenotypic features observed in the experiments. Despite the profound value of the work as proven in this study, there are several limitations needed to be further addressed. First, we could not differentiate the germline mutation and somatic mutation in our model due to the strategy used in Foundation Medicine for sample collection in the genetic test, which introduces bias to the system. Second, the DML+G may be infeasible as most genetic tests are usually based on the specimen collected from the biopsy or surgery. We consider that the DL model has more potential as it is more practical to learn a large number of phenotypical information for cancer prediction. Third, the notable failures for CUP cases indicate that the patterns learned in the training data can be applied for predicting CUP cases. The CUP cases usually do not have phenotypical symptoms at the origin site when spread at the early stage thus they are not considered as a single type of cancer. Our future work endeavors to address the limitations to improve the developed predictive models.

Acknowledgments

This work was supported by funding from Genentech Research Fund in Individualized Medicine, NIH NIGMS (K99GM135488), BD2K (U01HG009450), and FHIRCat (R56EB028101).

References

1. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Medical care. 2010:S106-S13.
2. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine. 2018;1(1):18.
3. Peng J, Guan J, Shang X. Predicting Parkinson's disease genes based on node2vec and autoencoder. Frontiers in genetics. 2019;10.
4. Bender D, Sartipi K, editors. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. Proceedings of the 26th IEEE international symposium on computer-based medical systems; 2013: IEEE.
5. Lassila O, Swick RR. Resource description framework (RDF) model and syntax specification. 1998.

Analysis of Racial and Socioeconomic Factors of COVID-19

Sahar Abdullah[1*], Caroline Vieira[2*], Siu Hui[2,3], Katie Allen[3], Umberto achinardi[2,3], Eneida A Mendonca[2,3]

IUPUI[1], Indiana University [2], Regenstrief Institute[3]

Introduction

In the past, people have had an increased rate of contracting illnesses outside health and physiological factors, such as race and income. One example of the effects can be exhibited with the recent onset of COVID-19. Studies have shown that Black populations are more likely to contract this strain of the coronavirus compared to their White counterparts[1,2]. This increased risk can be attributed to external factors, such as income and location. With these factors in mind, we conducted a preliminary study in which we aim to investigate the demographic (race, location, income) aspects of COVID-19 positive patient data.

Methods

The Regenstrief Institute registry, CoRDaCO (COVID-19 Research Data Commons), contained data ranging from January 2018 to July 2020 from the Indiana Network for Patient Care. This health information infrastructure includes data from 5 major hospital systems around Indiana. The counts and proportions of race, ethnicity, and gender were extracted with R and the number of patients per unique zip code. From there, the zip codes that had 1-5 COVID-19 positive patients were discarded, and the data was generally visualized with calculated rates. This information was then combined with data from public websites with information on income and population. Racial breakdowns were obtained from the five lowest and five highest income zip codes and analyzed using an odds ratio and rates (income with other vs. white).

Results

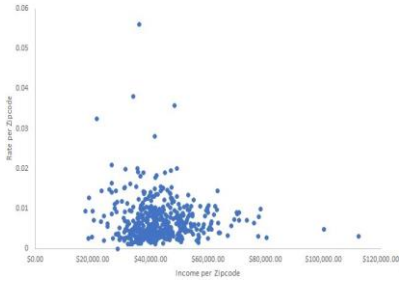


Figure 2. The rate of positive cases compared to income per zip code.

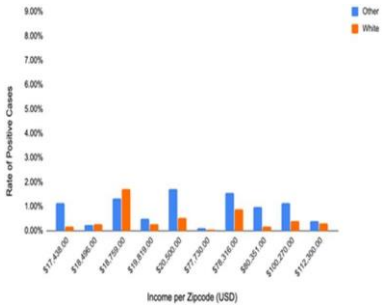


Figure 1. The rate of COVID-19 positive patients in white and non-white groups within the five highest and five lowest incomes per zip code.

Table 1. Odds ratio of COVID-19 positive non-white compared to white patients within the five highest and five lowest incomes per zip code.

Zipcode	Household Income	Odds Ratio	p-values
46601	$17,438.00	7.18	0.0002
47305	$18,496.00	0.9147	0.9103
46402	$18,759.00	0.7752	0.5234
47807	$19,819.00	1.9661	0.0421
47901	$20,500.00	3.3436	0.0082
46032	$77,730.00	1.2332	0.6151
46278	$78,316.00	2.7286	0.0001
46530	$80,351.00	5.7445	0.0001
46033	$100,270.00	1.7682	0.0096
46814	$112,300.00	2.5853	0.0362

Conclusion

The data observed from different websites and hospital/census databases provided much insight into the relationships between income, race, and COVID-19. When analyzing the rates of frequencies with incomes, it was found that zip codes with lower incomes had a higher rate of positive cases of COVID-19 compared to zip codes with higher incomes (Figure 1). Moreover, when conducting this analysis on the five highest and five lowest incomes, we observed that non-whites had a higher rate of COVID-19 than their white counterparts when at the same income level (Figure 2). In terms of the ten specific zip codes that were analyzed, most of the zip codes had an odds ratio higher than one, indicating that people of other racial backgrounds were being infected at a higher rate than those of white populations (Table 1). For the most part, the odds ratio was higher in areas where the majority of the population was white. Such conclusions could be attributed to a variety of factors that can be markers for unequal wealth distribution, along with there being fewer non-white people living in the area. Though such analysis provides insight into external factors related to COVID-19, there are some limitations, including the number of zip codes used for the odds ratio, gaps in the data, difficulty obtaining information about zip codes, and caveats with standardized and collecting healthcare data. Further exploration will focus on utilizing census tracts and more complex analysis of the data. By accomplishing these tasks, greater insight into the relationship between these factors can be delineated.

References

1. Haywood, Eboni et al. “Hospitalization and Mortality among Black Patients and White Patients with Covid-19”. The New England Journal of Medicine, 27 May 2020.
2. Rentsch, Christopher et al. “Covid-19 by Race and Ethnicity: A National Cohort Study of 6 Million United States Veterans”. Insert publisher, 18 May 2020.
3. Hawkins, Devan. “Differential occupational risk for COVID-19 and other infection exposure according to race and ethnicity”. American Journal of Industrial Medicine, 4 June 2020.

Weight-based Integrated Algorithm for Modeling Time-to-Event Data of Multiple Databases without Data Sharing

Ji Ae. Park, MS[1], Yu Rang. Park, Ph.D[1]
[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, South Korea

Background: Modeling for time-to-event for clinical events of interest, called survival analysis, is popular in clinical study in that it can estimate the prognosis of a disease, not whether or not a disease has occurred, and that patients with different follow-up periods including censoring data can be included in the model. A generalizability of time-to-event model can be improved by considering follow-up data of various patients in the model through multi-institution data. Despite the benefits of utilizing multi-institutional data, it is difficult to share the medical data of individuals with confidential characteristics. A methodological solution should be considered to model time-to-event data of multi-institutional data without sharing patient-level data.

Methods: The weight-based integrated algorithm (WIM) for Cox proportional hazard model is carried out in two stages. In the first step, 200 Cox models are generated through resampling at each institution, and the 200 integrated coefficients are calculated by integrating coefficients of the Cox model for each institution based on weights assigned to each institution. In the second step, 200 survival functions were estimated through 200 weight-based integrated coefficients. Based on 200 weight-based coefficients and survival curves, point and interval estimation of WIM for Cox model are performed using the mean and percentile method. We conducted an experiment using real multi-institutional data to verify the developed WIM for Cox model. It selected 10 hospitals (a total of 5,614 ICU stays) from the electronic Intensive Care Unit (eICU) Collaborative Research Database. The outcome was length of day from the date at ICU admission to the date at mortality during hospitalization, and 7 features was used. To evaluate validity of our model compared to centralized model, which was built by combining all the data of 10 hospitals, we used proportional overlap of confidence intervals(CIs)[1] (0.5 or less indicates a significant difference at a significance level of 0.05; 2 indicates two CIs overlapping completely).

Results: In the experiment, the proportion of overlap of the CIs for 7 features in both WIM for Cox model and the centralized Cox model was over 1. The median time from hospitalization to an event was 3.6 days, and a patient with average values of features was selected for estimation of survival rate. The 3.6 survival rates of WIM and centralized Cox model were 92.24% (95% CI, 91.78%-92.76%) and 92.44% (95%CI, 91.62%-93.26%), respectively (Figure 1). The proportion of overlap for the 3.6-day survival rate was 1.21. The 95% CI estimated by WIM for Cox model included survival rate of centralized Cox model at all time points.

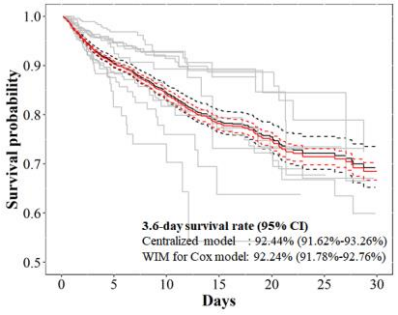


Figure 1. Cumulative survival rate curve. A black solid and dashed line indicate survival rate and 95% CI for centralized Cox model, respectively. A red solid and dashed line indicate survival rate and 95% CI for WIM for Cox model. Grey solid lines indicate survival rates for 10 Cox models of 10 hospitals.

Conclusion: The proposed WIM for Cox model is a privacy-protecting analytic method for modeling time-to-event data. Through the experiment using real multi-institutional data, the WIM for Cox model was shown that not only the coefficient of the model but also the survival curve can be estimated with high accuracy.

References

1. Geoff C, Fiona F. Interval estimates for statistical communication: problems and possible solutions. 2005.

Medication information resource tools: Who uses them and for what purpose are they used?

Shilo Anders PhD[1], Laurie L. Novak PhD MHSA[1], Nawshin Kutub PhD[2], Daniel France PhD MPH[1], Christopher Simpson MA[1], Courtney VanHouten MA[2], Karlis Draulis[2], Rubina Rizvi MD PhD[2], Tiffani J. Bright PhD[2], and Anita M. Preininger PhD[2]
[1]Vanderbilt University Medical Center, Nashville, TN, USA; [2]IBM, Cambridge, MA, USA

Introduction
Pharmacological knowledge bases (PKBs) provide current medication-related information[1] that healthcare personnel use on a regular basis. Understanding how PKBs are selected, accessed, and used in clinical practice can reveal barriers to widespread adoption and use. To better understand medication-related information needs and use of PKBs across clinical environments, we surveyed current PKB users at a large, tertiary academic medical center, Vanderbilt University Medical Center (VUMC), located in Nashville, Tennessee.

Methods
Participants were invited via email by clinical or informatics leaders. Consenting clinical staff within VUMC across a variety of clinical environments and settings (inpatient/outpatient, adult/pediatrics) including pharmacists, providers, nurses, and others who have used PKBs in the past year were eligible to participate in a survey exploring use PKBs. In order to better understand a specific user experience and workflow, we asked additional survey questions to participants that focused on a single PKB, Micromedex®. Surveys were administered and study data collected using REDCap according to best practices for anonymity, confidentiality, privacy and security.[2] We compared the types and frequencies of information sought, modes of use, and overall experience and perceptions about the PKBs. Results were stratified by role, practice area, and years of experience; significance of differences were calculated by (Chi square analyses). Factors that influenced the usage of PKBs were evaluated in this survey using a series of open-ended, free-response questions and thematic analyses conducted by three researchers. The study was determined exempt by the VUMC Institutional Review Board (IRB).

Results
We surveyed a total of 155 medical personnel. Respondents included nurses (31%), residents (25%), fellows (19%), pharmacists (8%), nurse practitioners (7%), attending physicians (5%), and others (5%). PKBs were accessed through web browser (82%), mobile application (55%) and/or the electronic health record (EHR) (48%). In response to open-ended questions regarding factors which influence usage of PKBs, users cited accessibility as the most important feature driving use of a PKB, followed by reliability and general ease of use. Of the 63 Micromedex users further surveyed, the most common uses involved side effects/indications (26%), drug-drug interaction (24%), dosing/dosing adjustment (24%), IV compatibility (23%), and administration (19%). Information in Micromedex was most often accessed using a keyword search (62%), followed by navigation of headings (59%). We found statistically significant differences (P = 0.008) between user roles and frequency of use of Micromedex; nurses and pharmacists used Micromedex daily or several times per week; nurse practitioners and residents used it several times per month; fellows used it less than once per month. There were no significant differences in mode of use of Micromedex by role, years in role, or practice area.

Discussion and Conclusion
PKBs help hospital staff care for patients in numerous ways related to medication management. Ease of accessibility and minimal interruption of the individual's workflow were the factors most associated with PKB use. Thus, how the EHR affords access to the PKB is critical. Micromedex, frequently used by nurses and pharmacists, is most often consulted for drug information needs including side effects, interactions, dosing, IV compatibility and administration. The relationship between user role, training, modes of access (web, EHR, or mobile application), previous exposure to similar tools, and specific modes of information retrieval are currently under further investigation.

References

1. Preininger AM, South B, Heiland J, Buchold A, Baca M, Wang S, Nipper R, Kutub N, Bohanan B, Jackson GP. Artificial intelligence-based conversational agent to support medication prescribing. JAMIA Open. 2020.
2. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG, Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed Inform. 2009 Apr;42(2):377.

Age Aware Model of Phenotype Risk Score

Layla Aref, BA[1], Lisa Bastarache[1], MS, Jacob J. Hughey, PhD[1]
[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Introduction

Electronic health records (EHR) have provided a powerful resource in studying different diseases and their genetic etiology. In the case of mendelian disorders, however, this can be a challenging task since they are generally rare in the population. Common strategies, such as genotype and phenotype association studies, are often inadequate in capturing the full phenotypic pattern of rare diseases as they examine phenotypes independently.

The Phenotype Risk Score (PheRS) is a method developed to address this problem by scoring patients based on their similarity to the clinical features of a mendelian disease as described in the Online Mendelian Inheritance in Man (OMIM)[1,2]. This method allows us to identify individuals with different combination of mendelian phenotypes who share the same genetic cause. The PheRS was able to distinguish between cases and controls of a number of mendelian diseases and uncovered potentially pathogenic variants that were previously unknown.

The Current PheRS model combines sub-phenotypes characterizing each mendelian disease weighed by their log inverse prevalence in the population. This weighting scheme assigns the same weight to all individuals with a particular phenotype and does not take into account time dependent factors, such as the timeline and frequency of phenotype emergence. Previous research has shown that modeling the risk of having a disease phenotype as a function of age increases the power for detecting significant associations in genome-wide association studies (GWAS)[3]. We hypothesize that, similarly, an age aware PheRS model will more accurately summarize the phenotypic characteristics of mendelian disorders and quantify the risk of harboring them.

Materials and Methods

We are calculating the PheRS for each mendelian disease through a weighted aggregation of its clinical sub-phenotypes. For this, we use the annotations provided in the Human Phenotype Ontology (HPO) using the OMIM mapping. These annotations are then mapped to diagnostic codes extracted from Vanderbilt University Medical Center (VUMC) Synthetic Derivative (SD), a de-identified version of VUMC's EHR.

Using Cox proportional hazard regression, we can quantify the risk of developing each disease sub-phenotype given the age of first visit and either the age at which the phenotype is first recorded in the EHR or the age of last visit. The risk values can then be used to construct new weights that are inversely related to the patient's predicted probability of having different sub-phenotypes.

To evaluate the final PheRS model, we are using a set of 16 mendelian diseases with known cases. We can then compare the performance of the two models based on how well each can distinguish between cases and controls of the gold standard set. Additionally, we can test the sensitivity of the new PheRS model in capturing significant pathogenic variants using genetic association studies.

Conclusion

The age aware model of PheRS can be used as a high throughput approach to generate evidence for pathogenicity of rare variants or identify unknown protective variants. PheRS can also provide the means to study mendelian disease inheritance patterns.

References

1. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science. 2018 Mar 16;359(6381):1233–9.
2. OMIM - Online Mendelian Inheritance in Man [Internet]. [cited 2020 Aug 10]. Available from: https://omim.org/
3. Hughey JJ, Rhoades SD, Fu DY, Bastarache L, Denny JC, Chen Q. Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. BMC Genomics. 2019 Nov 4;20(1):805.

A Review of Discordance in Disparate Clinical Data through the Construction of a General Pediatric Cancer Population

Ashley Batugo[1], Hanieh Razzaghi, MPH[1], Charles Bailey, MD, PhD[1]
[1]Children's Hospital of Philadelphia, Philadelphia, PA, US

Introduction

As pediatric cancer remains the leading cause of death in children past infancy, the National Cancer Institute recently introduced the Childhood Cancer Data Initiative with a goal to effectively use and share pediatric cancer data. With the emergence of clinical data warehouses, which provide a comprehensive view of a single patient by consolidating medical, operational, and clinical data[1], extracting cohorts of patients through various streams became much more feasible. Because cancer care is highly complex, and summary labels such as validated diagnoses or treatment plans are often not available, pediatric cancer patients must also be identified through various data sources. In this study, we discuss a three-fold approach to discovering this cohort of patients and address the discordance in the three analytical approaches taken.

Methods

Clinical source data stored in the EHR-based data warehouse at the Children's Hospital of Philadelphia were used to extract inpatients and outpatients with cancer from January 1, 2000 to August 24, 2020, using one of three criteria:

- **Method 1**: Visits to an oncology service
- **Method 2**: International Classification of Diseases (ICD) 9/10 CM codes indicating a malignant neoplasm diagnosis: ICD9: 140-239 EXCEPT 210-229 OR ICD10: C00-C96
- **Method 3**: Hospital Tumor Registry (Patients with any of the following primary or metastatic malignant conditions: Leukemia, central nervous system tumors, neuroblastoma, bone tumors, rhabdomyosarcoma, kidney tumors, liver tumors, retinoblastoma, germ cell and gonadal tumors, Hodgkin's disease, non-Hodgkin's lymphoma, and other malignant tumors)

Results

A total of 37,092 unique patients were identified. The overlap in patients is shown in Figure 1. In the registry cohort of patients, about 7.9% did not have visits to any oncology clinic. Instead, visits were largely in a radiology or endocrinology department. About 77.3% of those patients were identified as having thyroid tumors, leukemia, central nervous system tumors, neuroblastoma, or other malignant tumors. For those patients only with visits to an oncology service and identified as having malignant ICD codes, 45.1% were present in the registry with 75.9% having conditions that were instead 'neoplasm ruled out', 'benign central nervous system tumors', 'histiocytic disorders', 'neurofibromatosis', and 'non-malignant tumors.' For patients with only visits to an oncology service, 24% were recorded in the registry with 77.6% of these having a diagnosis of 'neoplasm ruled out', 'hematopoietic stem cell donors,' 'non-malignant hematologic disorders', 'conditions predisposing to cancer', and 'non-malignant tumors.'

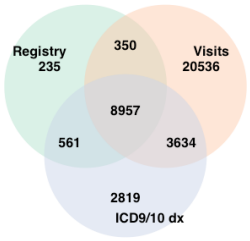


Figure 1. *Overlap in Patients*

Conclusion/Discussion

Consolidating disparate sources is crucial when seeking to discover a hospital-wide cohort of patients. This is especially true for pediatric cancer as patients with different conditions have largely heterogenous care pathways. Visits to an oncology service alone often indicated false positive patients, however this conclusion requires further exploration. Similarly, patients were at times falsely diagnosed as having a malignant condition through ICD visit diagnoses. Often, it was in the tumor registry where a diagnosis was confirmed as malignant. Nevertheless, registries should not be used as the only source of truth for diagnoses given the latency of this manual process in identifying patients. In the case of this cohort, remaining inclusive while establishing sub-cohorts by degree of confidence as having cancer, may allow us to more effectively capture a general pediatric cancer population by condition.

References

1. Evans RS, Floyd J, Pierce, L. Clinical use of an enterprise data warehouse. AMIA Annu Symp Proc. 2012;2012: 189-198

From Traditional Anonymization to State-of-the-Art Privacy Protection – Overview of Features of the ARX Data Anonymization Tool

Lena Baum, M.Sc.[1,2], Marco Johns, M.Sc.[1,2], Thierry Meurers, M.Sc.[1,2], Fabian Prasser, Ph.D.[1,2]
[1]Berlin Institute of Health, Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany;
[2]Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

Introduction

Using anonymization or de-identification, data is transformed to reduce the risks to the privacy of individuals when data is processed for secondary purposes or disclosed to third parties. Developing tools which support this process is challenging due to the technical complexity of many anonymization methods and associated scalability and usability issues. The ARX Data Anonymization Tool is a comprehensive open source software for anonymizing structured individual-level health data. It was first presented at the 2014 AMIA Annual Symposium [1], has been continuously developed and extended with further functionalities [2] and is freely available on the project website (https://arx.deidentifier.org). The objective of this poster is to provide an overview of the current features of the software, a comparison with related projects, and an outlook on future developments.

Methods

Using ARX data can be transformed semi-automatically while privacy risks are traded off against output data utility. The tool can be used to de-identify data according to the Safe Harbor and Expert Determination methods defined by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule and for applying quantitative anonymization approaches required by the EU General Data Protection Regulation (GDPR). The three most distinctive properties of the software are: a comprehensive set of supported features, high scalability, an intuitive graphical frontend. In the poster, we will provide an overview of the software along three important axes:

1. **Transformation models** – The software supports more than 10 different ways of transforming data, including methods, such as generalization, top- and bottom-coding, suppression, sampling and micro-aggregation.
2. **Privacy models** – ARX supports more than 10 different methods for quantifying and protecting privacy, including traditional syntactical models, such as k-anonymity, statistical models for population uniqueness, and state-of-the-art semantic models including a game-theoretic approach and (ε,δ)-differential privacy.
3. **Utility models** – The tool supports more than 10 different models for quantifying the utility of data during the anonymization process, including general-purpose methods reflecting data fidelity or changes to value distributions as well as application-specific models, e.g. for creating privacy-preserving machine learning models.

In addition, selected highlights will be presented, including plugins for Extract-Transform-Load (ETL) tools and an integrated reliable computing framework that ensures that ARX delivers on the promised privacy guarantees. Furthermore, we will present future extensions, including data masking features, risk assessment based on Minimal Sample Uniques and an integration of the R statistical programming framework for utility analyses. Based on the presented features, ARX will be compared to related open source tools, including the UTD Anonymization Toolbox, the Cornell Anonymization Tool, sdcMicro, Amnesia, Open Anonymizer and μ-Argus (see [3] for an overview of some related tools).

Results

The results of our comparison will show that ARX supports a much broader spectrum of methods than any comparable open source tool. Moreover, almost all methods supported can be combined arbitrarily and all methods are supported through the graphical user interface. Also in terms of scalability, ARX outperforms related tools.

Conclusion

This poster provides an overview of the features supported by ARX and shows that it is much more comprehensive than related tools. That these are important features of the software is demonstrated by its broad adoption and international recognition, which we will focus on in another poster at AMIA 2021.

References

1. Prasser F, Kohlmayer F, Lautenschlaeger R, Kuhn KA. ARX – A comprehensive tool for anonymizing biomedical data. In: AMIA Annual Symposium Proceedings 2014 (Vol. 2014, p. 984). American Medical Informatics Association.
2. Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX – Current status and challenges ahead. Software: Practice and Experience. 2020 Jul;50(7):1277-304.
3. Meindl B, Templ M. Feedback-based integration of the whole process of data anonymization in a graphical interface. Algorithms. 2019 Sep;12(9):191.

A Natural Language Processing Pipeline for the De-identification of Clinical Notes at an Academic Medical Center

Arnav Bhandari, Michael Horvath, Joseph Rigdon Ph.D., Umit Topaloglu Ph.D., FAMIA, Wake Forest School of Medicine, Winston Salem, NC

Introduction

Healthcare is expected to generate 2.3 zettabytes of unstructured data in 2020, which presents regulatory and computational challenges for research purposes. To comply with the Health Insurance Portability and Accountability Act (HIPAA), Protected Health Information (PHI) must be redacted to enable researchers' easier access. This project highlights the implementation process of a de-identification pipeline at the Wake Forest Baptist Medical Center using natural language processing (NLP). This process has been approved by the Institutional Review Board (IRB) and the Chief Privacy Officer of this institute.

Methods

MITRE identification Scrubber Toolkit[1] (MIST) software is a widely used NLP based tool for identifying PHI in unstructured data. In our data set, PHI includes NAME, LOCATION (related to patient), DATE, PHONE, HOSPITAL, IDNUM, and AGE. These tags are chosen per the safe harbor method. Similar identifiers are grouped, e.g., IDNUM encapsulates any alphanumeric identifiers. NAME encapsulates both the patient and the doctor's name. This project initially considers pathology reports (N=368,355). A random representative sample of n=860 reports are extracted, and any PHI is manually tagged. These records are randomly chosen to get a relatively uniform distribution of records (~100) from each year and percentage equivalents of each type, allowing for changing formats. The extracted reports are randomly split (80/20) into a training set (n_1=688) and a test set (n_2=172).

The tagging is completed via MIST's web-based interface. The final pipeline includes calling MIST functions and internally developed masking python code. The pipeline is ported into a Docker container for easy deployment and is linked to an external oracle database which stores the unprocessed and NLP based de-identified notes.

Results and Discussion

		True labels (hand-tagged)		
		PHI	Non-PHI	Total
MIST label	PHI	1497	64	1,561
	Non-PHI	47	62,809	62,856
	Total	1544	62,873	64,417

Table 1: Confusion Matrix containing the overall PHI vs. non-PHI at token level

The test set contains 64,417 labels in n_2 = 173 reports. Performance on the test set is measured by the precision, recall, and F-measure scores. Our model has a 95.9% precision (positive predictive value), 97.0% recall (sensitivity), and F-measure 96.43% (95% CI = 95.82%, 97.09%). A 95% bootstrap confidence interval for the F-measure was calculated by taking 1,000 samples of size 64417 (from Table 1). The percentage of PHI redacted is considerably high, at 96.95% (1497/1544). Notably, the tags for NAME, IDNUM and DATE have a recall above 99%.

Notably, this performant pipeline averages processing 1 report per 0.077 seconds. This model is applied to all N=368,355 reports to obtain de-identified reports indexed by Elasticsearch. We have created visualizations and made these notes searchable through Kibana dashboards. This is accessible to CITI-trained researchers at the Wake Forest School of Medicine for exploration of de-identified reports in the preparatory research phase. The novelty of this approach lies in the performant implementation and streamlined availability of indexed de-identified notes which has been approved by our IRB. This approval has laid the foundation for models to be created for other types of notes to further benefit our research community.

References

1. Aberdeen J, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. International Journal of Medical Informatics. 2010;79(12):849-859. doi:10.1016/j.ijmedinf.2010.09.007

Evaluation of SOFA score for Outcome Prediction in COVID-19 ICU Patients

Kriti Bhattarai, BS[1], Mackenzie Hofford, MD[2], Sean C. Yu, MS[1], Seunghwan Kim, MS[1], Aditi Gupta, PhD[1], Albert M. Lai, PhD[1], Philip R.O. Payne, PhD[1], Andrew P. Michelson, MD[1,3]

[1]Institute for Informatics, Washington University in St. Louis, St. Louis, MO

[2]Division of General Medicine, Washington University in St. Louis, St. Louis, MO

[3]Division of Pulmonary and Critical Care, Washington University in St. Louis, St. Louis, MO

Introduction

The Sequential Organ Failure Assessment (SOFA) scoring system was developed in an attempt to quantify the severity of illness to estimate risk of mortality [1]. While hospital triage systems have used SOFA score as an element of resource triage decisions, it has not been well validated for estimating mortality in COVID-19 patients. In this analysis, we evaluate predictive performance of the SOFA score in COVID-19 ICU patients for outcome (mortality) prediction.

Methods

All patients with a positive COVID-19 PCR test who were admitted to an ICU for at least 24 hours at Barnes-Jewish Hospital, a large tertiary-care academic medical center in St. Louis. MO, between 03/12/2019 and 10/24/2020 were eligible for inclusion. All inpatient information, including demographic, vital sign, laboratory, flowsheet, mortality, and admission/discharge/transfer data were extracted from the electronic health record. The SOFA score was calculated according to established criteria [1]. Based on the results, each patient was assigned a score out of 24, with a higher score indicating a greater severity of illness. Replacement criteria was applied for all missing values. Missing mean arterial pressure (MAP) and Fraction of inspired oxygen (FiO2) was calculated from blood pressure and O2 flow rate, respectively following established methods [2,3]. If O2 flow rate was missing, we assumed an FiO2 of 21%, equivalent to room air. Missing PaO2 values were estimated from SpO2 values [4]. For all other missing data, mean, median, last observation carried forward and K-Nearest Neighbor (KNN) imputation techniques were tested. Data with no prior measurements was assumed to be missing not at random due to clinical judgement and imputed as normal. Mortality rate, initial SOFA score, maximum SOFA score, mean SOFA score, and ΔSOFA score (difference between maximum SOFA score and initial SOFA score) was calculated for all patients within the first 24 hours of admission.

Results

Out of 4527 COVID-19 positive patients, 864 patients met the inclusion criteria. Out of the 864 patients, in-hospital mortality occurred in 306 (35.42%). The area under receiver operating characteristic (AUROC) curve and area under precision recall curve (AUPRC) for mortality prediction within the first 24 hours of admission are shown in Figure 1. SOFA score with no imputation presented larger area under curve (AUC) in comparison to the SOFA score with other imputation techniques (AUC with no imputation = 0.7404; AUC with median imputation = 0.7349; AUC with mean imputation = 0.7344; AUC with KNN Imputation =0.7343).

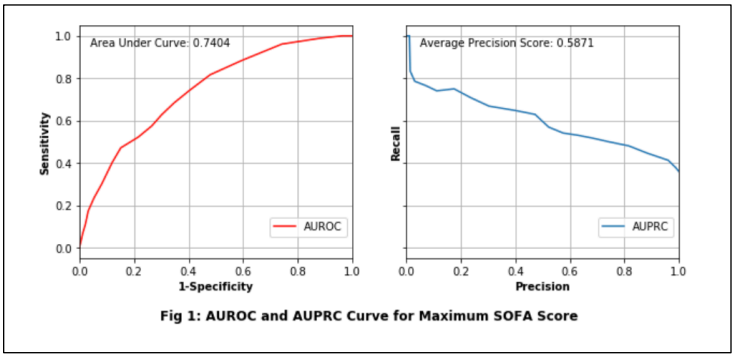


Fig 1: AUROC and AUPRC Curve for Maximum SOFA Score

Conclusion

This initial retrospective analysis attempts to validate the SOFA score for mortality prediction in COVID-19 ICU patients. Our results indicate that maximum SOFA predicts patient mortality relatively well in our COVID-19 ICU cohort and may be a useful predictor of outcome. Future direction includes integrating Acute Physiology and Chronic Health Evaluation (APACHE) II score and building a machine learning pipeline with electronic health record variables to determine the comparative validity.

References

[1] Ferreira FL, Bota DP, Bross A, Mélot C, Vincent J. Serial Evaluation of the SOFA Score to Predict Outcome in Critically Ill Patients. *JAMA*. 2001;286(14):1754–175

[2] Brzezinski WA. Blood Pressure. In: Walker HK, Hall WD, Hurst JW. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. Boston: Butterworths; 1990. Chapter 16.

[3] Wettstein RB, Shelledy DC, Peters JI. Delivered Oxygen Concentrations Using Low-Flow and High-Flow Nasal Cannulas. Respiratory Care. 2005; 50(5) 604-09

[4] Pandharipande, Pratik P et al. "Derivation and validation of Spo2/Fio2 ratio to impute for Pao2/Fio2 ratio in the respiratory component of the Sequential Organ Failure Assessment score." *Critical care medicine* vol. 37,4 (2009): 1317-21.

Schematic Design of a Health Information Exchange Performance Reporting Platform

Nathan E. Botts, PhD[1,2,3], Eric C. Pan, MD, MSc, FAMIA[1,2,3], Nelson S. Hsing, ScD[1], Omar Bouhaddou, PhD[1,4], MS, Bharathi Vedula, MS[1,4], Jim Malpass[1,4], Jeffrey E. Anderson, MD, MS[1]
[1]Veterans Health Information Exchange, Veterans Health Administration, Washington, DC; [2]Westat, Rockville, MD; [3]JP Systems, Clifton, VA; [4]Innovet Health, Los Angeles, CA

Background and Need

The Program Performance and Data Analytics team are tasked with providing analytical and statistical support for the Department of Veterans Affairs (VA) Veterans Health Information Exchange (VHIE). Reports generated by this team support VA leadership in strategic decision-making, inform VHIE Community Coordinators as to the status of exchange within their area of management and assist VHIE operational managers with analyses and trends in system quality and interoperability over time [1]. A primary objective is to inform VA business needs: improve Veteran care through effective clinician utilization of health information exchange. Progress toward this goal is monitored through Key Performance Indicators (KPIs) (e.g., rate of exchange) and Critical Success Factors (CSFs) (e.g., eHealth Exchange adaptor system uptime). Consequently, measurement of progress and achievement of VHIE goals, objectives, KPIs, and CSFs are all derived from distinct data points that dictate the architecture of an analytic reporting system (Figure 1). For example, a key indicator of VA clinician adoption would include transactional information on inbound HIE traffic trends to VA requiring upwards of three different audit tables to formulate. The more detailed poster that will describe this research will expose the understanding derived from mapping these direct connections and the broader impact on performance evaluation throughout the evolution of the VHIE system.

VHIE Analytic Performance Reporting Platform Schema

As VHIE systems continue to evolve and migrate to commercial platforms, it will be essential to understand the impact and level of effort required to develop VHIE Analytics reporting platforms so that they can continue to measure, and monitor progress and performance [2].

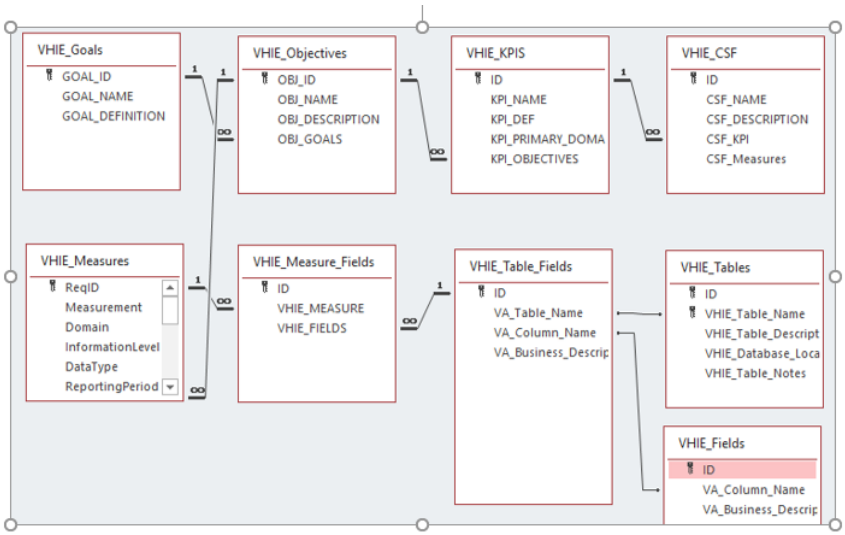


Figure 1. VHIE Analytic Performance Reporting Platform

This poster will describe the data used and how those data are mapped against VHIE goals, objectives, KPIs, and CSFs. These data are retrieved from VHIE repositories and logs that include systems such as the Veterans Authorizations and Preferences, the Master Veteran Index, and the VHIE gateway and adaptor. These systems are used to monitor reporting domains for Veteran consent collection, patient matching success, and volume of retrievals and disclosures of patient health summaries. Schemas such as this will provide a valuable blueprint for similar HIE efforts.

References

[1] Botts, N., Bouhaddou, O., Bennett, J., et al. (2014, November). Data quality and Interoperability challenges for eHealth exchange participants: Observations from the Department of Veterans Affairs' Virtual Lifetime Electronic Record Health Pilot Phase. Annual Meeting of the American Medical Informatics Association, Washington, DC.

[2] Donahue, M., Bouhaddou, O., Hsing, N., Turner, T., Crandall, G., Nelson, J., & Nebeker, J. (2018). Veterans Health Information Exchange: Successes and Challenges of Nationwide Interoperability. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2018, 385–394.

Extracting COVID-19 Related Symptoms from EHR Data: A Comparison of Three Methods

Hannah A. Burkhardt[1], Nicholas Dobbins[1], Brenda Mollis[1], Margaret Au[1], Kris Pui Kwan Ma[1], Meliha Yetisgen[1], Angad Singh[1], Matthew Thompson[1], Kari A. Stephens[1]

[1]University of Washington, Seattle, WA, USA

Introduction

The COVID-19 pandemic has claimed over 310,000 lives in the United States[1]. A promising resource for discovery in COVID-19's symptom progress is data documented in electronic health record (EHR) systems as part of clinical care. Such data are stored in disparate locations within the EHR, requiring multiple extraction methods. We compared the symptom detection rates of three extraction methods to assess the comparative utility of each source of COVID-19 related symptoms within the EHR.

Methods

Symptoms were extracted from EHR data for all patients who were tested for SARS CoV-2 through May 31, 2020 from a single large healthcare system in the state of Washington. Three methods were used: 1) extraction of ICD-10 codes, which reflected symptoms and diagnoses documented for medical billing, 2) regular expression matching of clinical notes using a COVID-19 note template developed for standard use across the health system, and 3) a previously reported and evaluated Natural Language Processing (NLP) pipeline[2,3] applied to clinical notes. Patients were considered to either have or not have each of 11 different symptoms (fever, cough, shortness of breath, sore throat, rhinorrhea, headache, GI symptoms, general aches and pains (myalgia), anosmia, ageusia, and chills) by each of the 3 methods if they were documented in the EHR in the 10 days prior to SARS CoV-2 PCR lab test. ICD codes and NLP and pattern parsing outputs were matched to one of the 11 symptoms. We obtained descriptive statistics on the unique and overlapping symptoms detected by each of these extraction methods. A small sample of notes was manually annotated for symptom presence by the authors and compared to automatically extracted symptoms to validate NLP performance.

Results

SARS CoV-2 PCR tests were conducted across 25,115 unique patients, who were given 32,924 total tests between February 29 and May 31, 2020. COVID-19 related symptoms were extracted at differential rates across sources within the EHR (see Figure 1). On average, tested patients had 1.1 (SD 1.9) symptoms documented within 10 days before a SARS CoV-2 PCR test, with cough (24%), myalgia (23%), and fever (20%) being the most common. However, 65% of tests had no associated symptoms identified. NLP detected the most symptoms of all the extraction methods, namely 88.2% of all symptoms, and 66.5% were detected only by NLP. The ICD data source added 3,554 (10.0%) symptoms that were not already captured by NLP, and the parsing of notes using regular expression extraction from a known structure added 636 (1.8 %) more symptoms. In a small sample of 10 manually annotated notes, NLP demonstrated an average sensitivity of 79% and an average specificity of 77%.

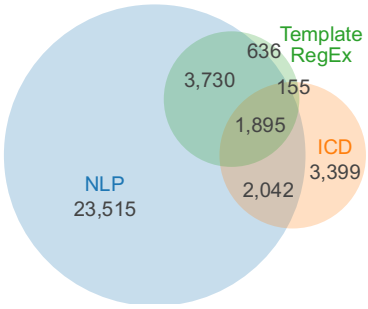


Figure 1. COVID-19 related symptom totals and overlap between extraction methods.

Discussion & Conclusion

All three extraction methods contributed to COVID-19 symptom detection, with NLP detecting the large majority of symptoms and template parsing detecting the least number of symptoms. A standardized note template containing a discrete checklist of COVID-19 related symptoms led to simple and highly accurate text parsing; however, the template was used infrequently, and NLP extraction was able to parse most of the template-derived symptoms. ICD codes directly provide discrete symptom data; however, NLP captured more symptoms than ICD codes, possibly because clinical narrative tends to be more detailed and captures information peripheral to the chief complaint. Given NLP methods resulted in the highest extraction rate of COVID-19 related symptoms, using only methods such as note template parsing and structured data extraction of ICD codes may miss a significant amount of symptom data.

References

1 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;**20**:533–4. doi:10.1016/S1473-3099(20)30120-1

2 uw-bionlp/uwbionlp-parser. https://github.com/uw-bionlp/uwbionlp-parser (accessed 24 Aug 2020).

3 Yetisgen M, Vanderwende L, Black T, *et al.* A New Way of Representing Clinical Reports for Rapid Phenotyping. In: *Proceedings of AMIA 2016 Joint Summits on Translational Science*. San Francisco: 2016.

Characterizing Respiratory Symptoms and COVID-19 Trends from OMOP-CDM database for Public Health Reporting

Marcello Chang, MS[1*], Jonathan Lu, MS[1*], Birju Patel, MD, MPH[1], Nigam H. Shah, MBBS, PhD[1], Jonathan H. Chen, MD, PhD[1]

[1]Center for Biomedical Informatics Research, Stanford School of Medicine, Stanford, CA

* indicates equal contribution

Introduction

While healthcare systems have focused on overall case reporting to inform the COVID-19 response, finer information on patient demographics, respiratory illness, tests and results is also important. For example, population shifts in respiratory symptoms for patients not yet tested for COVID-19 may be difficult to identify with currently mandated data. The CDC does not typically have access to joint trends in patient characteristics, as these require a robust and interoperable health information infrastructure. Our objective was to demonstrate an example of public health analytics for COVID-19 by using our clinical data warehouse to provide a synthesis of demographic and clinical data from our healthcare system.

Methods

We queried the Stanford Medicine STARR-OMOP[1] de-identified data warehouse, which conformed to the OMOP-CDM schema[2] and can be accessed via Google BigQuery. We composed SQL queries for three phenotypes: 1) visits related to respiratory illnesses identified by related concept families of ICD10CM codes, 2) SARS-CoV2 Nucleic Acid Amplification tests identified by LOINC codes, and 3) results associated with those SARS-CoV2 tests. To characterize the relationship between testing and symptoms, we additionally filtered SARS-CoV2 tests that resulted up to 14 days after respiratory diagnoses. The data were then stratified by age, gender, race, and ethnicity. Metadata conveying processing parameters were also included in the final export.

Results and Lessons Learned

From 6/10/2020 to 7/21/2020, we supplied 5 data updates, cumulatively reporting on 44,240 patients with respiratory symptoms and 79,200 patients receiving SARS-CoV2 tests. We were able to extract clinical data and create a public health update in less than three minutes, which could be re-run on-demand. Example reports are available publicly at tinyurl.com/stanfordcovidcdcmmwr

There were several factors contributing to our success, including the availability of de-identified STARR-OMOP data for prototyping prior to establishing a data transfer agreement with public health agencies and the use of a common data model enabled rapid design iteration as the health questions evolved. The use of a standardized schema allows our analysis pipeline to be transferred to other institutions to enhance public health reporting. Doing so would be a major step ahead in addressing the data woes surrounding COVID19 response[3]. Limitations of the data include missing values in certain fields, which may reflect workflow issues or limitations of the data transformation for our analytic warehouse.

Conclusion

Data from the OMOP-CDM can facilitate serving an urgent public health need. Compared to traditional analytics, which often involve tailored queries, a common data model can facilitate rapid and iterative reporting for clinical trends during a pandemic.

References

1. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj, D, Mesterhazy J, Pallas J, Desai P, Shah NH. A new paradigm for accelerating clinical data science at Stanford Medicine. Arxiv preprint.
2. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, Lei J van der, Pratt N, Norén GK, Li Y-C, Stang PE, Madigan D, and Ryan PB. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform. 2015; 216: 574–578.
3. Maxmen, Amy. Why the United States is having a coronavirus data crisis. Nature 2020, doi:10.1038/d41586-020-02478-z.

Ranking-based Convolutional Neural Network Models for Peptide-MHC Binding Prediction

Ziqi Chen, Ph.D. student[1], Martin Renqiang Min, Ph.D.[2], Xia Ning, Ph.D.[1,3,4]
[1]Computer Science and Engineering, The Ohio State University, Columbus, OH, USA;
[2]Machine Learning Department, NEC Labs America, Princeton, NJ, USA;
[3]Biomedical Informatics, The Ohio State University, Columbus, OH, USA;
[4]Translational Data Analytics Institute, The Ohio State University, Columbus, OH

Introduction Computational predictive models have been widely used to predict the binding between peptides and major histocompatibility complex (MHC) genes [1]. In this study, we developed two new deep convolutional neural networks (CNNs) with attention mechanism for allele-specific peptide-MHC binding prediction. We conducted a comprehensive study on the model architectures and ranking-based learning objectives so as to accurately prioritize the most promising peptides that need be experimentally assessed for the design of peptide vaccine.

Methodology We developed the following CM model and SCM model for peptide-MHC binding prediction. CM, as presented in Figure 1, is composed of 1D convolutional layer, a self-attention layer and a fully connected layer. The 1D convolution layer applies multiple kernels of length k to learn the k-mer embeddings. The self-attention layer is employed to assign an attention weight to each k-mer embedding. The weighted sum of all the k-mer embeddings is then used as the input to the fully connected layer for the binding prediction. The SCM model, as presented in Figure 1, extends the CM by having global kernels to extract global features for peptides. The global features are concatenated with the peptide embeddings for the binding prediction. We also proposed three learning objectives based on pair-wise hinge loss functions to prioritize promissing peptides. Given a pair of peptides of different binding levels, these objectives require the score of peptide of higher binding level greater than that of lower binding level by a margin m. The first objective $\mathcal{L}_{\text{v}}$ defines the margin m between two peptides using the difference of their binding affinities. The second objective $\mathcal{L}_{\text{l}}$ defines the margin using the difference of their binding levels. The third objective $\mathcal{L}_{\text{i}}$ extends $\mathcal{L}_{\text{l}}$ by requiring similar peptides to have similar predicted scores. These objectives are compared with the baseline objective $\mathcal{L}_{\text{ms}}$ which uses mean square loss as in many regression models. Note that the ranking based learning objectives formulate the peptide-MHC prediction as to rank peptides for each MHC allele instead of accurately predicting binding affinities, and thus a novel ranking problem with great robustness against measurement inaccuracy in training data.

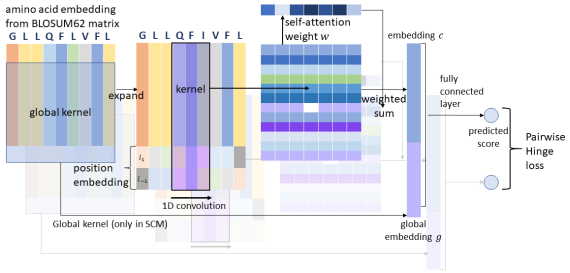


Figure 1: Architectures of CM and SCM

Datasets and Experimental Results We assembled dataset from the Immune Epitope Database (IEDB) with 202,510 binding affinity measurements across 128 alleles. We built our models by combining CM and SCM with different learning objectives, and compared our models with the state-of-the-art baseline MHCflurry [1] with $\mathcal{L}_{\text{ms}}$. We evaluated these methods with 5-fold cross validation and report their best performance in terms of the average of all 7 metrics across the 128 alleles. Table 1 presents the average improvement over MHCflurry +$\mathcal{L}_{\text{ms}}$ on IEDB dataset. The results show that the combination SCM with $\mathcal{L}_{\text{v}}$ has the highest improvement over the baseline among all the models (e.g., 12.45% improvement over the baseline on $\texttt{ROC}_{10}$). The loss function $\mathcal{L}_{\text{v}}$ enables significant improvement on most metrics (e.g., for MHCflurry, $\mathcal{L}_{\text{v}}$ enables improvement over all metrics). This demonstrates the strong potential of our new deep learning models in prioritizing the most promising peptides.

Table 1: Performance Comparison (%)

method	loss	$\texttt{AR}_{100}$	$\texttt{HR}_{100}$	$\texttt{AR}_{500}$	$\texttt{HR}_{500}$	AUC	$\texttt{ROC}_5$	$\texttt{ROC}_{10}$
CM	$\mathcal{L}_{\text{v}}$	8.12	7.82	1.30	3.46	2.88	7.71	5.49
	$\mathcal{L}_{\text{l}}$	3.40	4.56	0.47	3.38	2.21	4.87	2.39
	$\mathcal{L}_{\text{i}}$	6.89	7.72	2.60	3.64	3.23	7.42	4.98
	$\mathcal{L}_{\text{ms}}$	-8.37	-0.54	-8.04	-3.45	-1.73	-7.07	-7.59
SCM	$\mathcal{L}_{\text{v}}$	**12.11**	**10.46**	**6.84**	**8.63**	**4.87**	**18.26**	**12.45**
	$\mathcal{L}_{\text{l}}$	7.41	7.46	4.90	5.94	3.92	13.79	8.41
	$\mathcal{L}_{\text{i}}$	8.63	6.40	5.87	6.85	4.27	13.48	9.28
	$\mathcal{L}_{\text{ms}}$	5.55	6.93	2.56	3.82	2.35	13.71	9.10
MHCflurry	$\mathcal{L}_{\text{v}}$	10.62	10.04	5.81	5.27	3.86	12.09	9.12
	$\mathcal{L}_{\text{l}}$	7.99	7.69	5.06	4.33	3.97	10.30	7.27
	$\mathcal{L}_{\text{i}}$	7.28	6.58	5.88	5.83	4.10	9.97	8.20
	$\mathcal{L}_{\text{ms}}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00

"ARx" denote the average rank of peptides with binding affinity less than x μM. "HRx" denote the hit rate of peptides with binding affinity less than x μM. "ROCx" denote the area under the ROC curve up the x-th false positives. The numbers in the table are percentage improvement compared to the baseline MHCflurry with $\mathcal{L}_{\text{ms}}$ loss. The best performance under each metric is **bold**.

References

1. Timothy J. O'Donnell, Alex Rubinsteyn, Maria Bonsack, Angelika B. Riemer, Uri Laserson, and Jeff Hammerbacher. MHCflurry: Open-source Class I MHC binding affinity prediction. *Cell Systems*, 7(1):129–132.e4, 2018.

Knowledge Management of Clinical Pharmacogenetic Information – Complexities and Challenges

Christine M. Cheng, PharmD, Brad Green, BS, George A. Robinson, BSPharm, and Jeff Bubp, PharmD
First Databank Inc, South San Francisco, CA

Introduction

The incorporation of pharmacogenetic (PGx) testing in clinical practice has become an increasingly recognized strategy for optimizing medication use.[1] PGx data can support different electronic health record applications including clinical decision support, dashboard surveillance, or as a reference tool. A clinical decision support application should provide succinct alerting with recommended prescribing actions to take and why. A dashboard application might include a patient's medication list along with actionable genetic results and drugs for which testing may be considered. A referential workflow might involve a comprehensive report of a patient's genetic test results, whether actionable or not.

Data model

A pharmacogenetic information model for clinically relevant drug-gene-phenotype triads ("triads") for different PGx applications must account for gene-based recommendations that depend on a variety of factors. These factors can be broadly grouped into the following areas: (1) variant specificity, (2) treatment intent, (3) drug formulation characteristics, and (4) patient characteristics. Variant specificity includes single and multiple gene results, genotype, phenotype, activity score, and presence or absence of certain low function alleles that can affect drug response. Treatment intent refers to the drug's reason for use, as recommendations may be specific to a drug's indication (e.g., clopidogrel/CYP2C19 phenotype in ACS-PCI patients) or to uses that require higher initial doses (e.g., tricyclic antidepressants/CYP2D6).[2,3] Drug formulation characteristics are important since recommendations may apply to certain routes of administration but not others (e.g., systemic versus topical tacrolimus/CYP3A5).[4] Patient characteristics such as age (e.g., pediatric vs adult recommendations for atomoxetine/CYP2D6 phenotype) and ancestry or race may also influence gene-based guidance.[5]

PGx recommendations should be stratified in a manner that can enable interruptive alerting when clinician action is appropriate versus those that simply provide informational content expected within a "look-up" or referential application. Separate recommendations may be associated with genetic test status (e.g., result present, pending, never ordered). Mechanism of drug-gene interaction (e.g., polymorphic drug disposition, adverse drug reaction susceptibility) as well as potential clinical consequence (e.g., increased toxicity, poor efficacy) with reference citations and evidence ratings should also be included.

Our PGx model, which is based on iterative review of PGx information in drug labeling and practice guidelines, includes attributes that allow for variant comprehensiveness, drug and patient characteristics, interoperability and alert construction and classification. Due to the lack of current PGx lab interoperability and standardization, our model also accounts for variations in result terminologies used across laboratories and health systems and can accommodate future lab interoperability standards should they be implemented. Examples of pharmacogenetic information modeling and sample alert constructs for display in clinical decision support systems will be presented.

References

1. Hicks JK, Aquilante CL, Dunnenberger HM, et al. Precision pharmacotherapy: integrating pharmacogenomics into clinical pharmacy practice. J Am Coll Clin Pharm. 2019;2:303-313.
2. Scott SA, Sangkuhl K, Stein CM et al. Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. Clin Pharmacol Ther. 2013; 94: 317–323.
3. Hicks JK, Sangkuhl K, Swen JJ, et al. Clinical Pharmacogenetics Implementation Consortium Guidelines (CPIC) for CYP2D6 and CYPC19 genotypes and dosing of tricyclic antidepressants: 2016 update. Clin Pharmacol Ther. 2017;102:37-44.
4. Birdwell KA, Decker B, Barbarino JM et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for CYP3A5 genotype and tacrolimus dosing. Clin Pharmacol Ther. 2015;98:19-24.
5. Brown JT, Bishop JR, Sangkuhl K, et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6 genotype and atomoxetine therapy. Clin Pharmacol Ther. 2019;106:94-102.

What can online Ovarian Cancer Forum present us about patients' information needs? - A Text-Mining Approach

Vivian Hui[a], Zhendong Wang[c], Young Ji Lee[a,b]
a. Department of Health and Community Systems, School of Nursing, University of Pittsburgh, USA
b. Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, USA
c. Department of Informatics and Networked Systems, School of Computing and Information, University of Pittsburgh, USA

Introduction: Ovarian cancer (OvCa) is the most lethal gynaecological cancer with the highest rate of recurrence. OvCa patients request abundant information on the internet across the continuum of cancer care.[1] Social media has emerged as a resourceful hub for patients to share their experiences and seek help from each other. It has also provided a salient database for clinicians and researchers to extract unique self-reported patients' experiences, which are not commonly reported in clinical data. Previous studies have shown that patients who suffered from chronic illness use online health communities (OHCs) to discuss similar health issues and symptoms with peers.[2,3] Leveraging OHCs to understand the unmet needs of OvCa patients would add value to cancer research; however, there is a dearth of research examining the OHCs among OvCa patients. Thus, the objective of this study is to use topic modeling to understand the most influential topics among OvCa patients through the analysis of online forum posts.

Method: Data were extracted from an online OvCa forum (i.e., National Ovarian Cancer Coalition), which contains more than 900 posts that include initial thread and comment. We extracted the id, user no., URL, title of the post, post content, and post comments from the forum. We used lemmatization and synonym grouping functions to filter words with the same meaning with Python 3.8.0 programme NLTK package. After tokenization, stemming, stop words removal, symbols removal, and synonym matching, Latent Dirichlet Allocation (LDA) was used to identify topics. We explored the dataset with a manual investigation to determine the appropriate number of topics. Topic and keywords distributions are displayed by pyLDAvis (Figure 1). The access of data has been approved by NOCC and reviewed by the institutional review board in our institution.

Result: 909 posts were extracted from 460 users in the OHC. Stage, treatment, new, chemo, cancer, surgery, trial, question, tumor, recurrence, year, cell, group, expert, clinical were the most frequent fifteen words that appeared from the entire corpus. We identified fifteen topics from our LDA model, which covered across the cancer trajectory (Table 1). Among 15 topics, treatment-related (n=6) played a preponderant role, while keywords of chemotherapy (n=3) and recurrence (n=3) prevail in the forum discussion. Daily life sharing (n=4) about chemotherapy story, gratitude, support, and emotions are commonly found and associated with broad discussion scope.

1.Clinical trial	6.Making a decision for surgery	11.Metastasis and recurrence
2.Chemotherapy	7.Gratitude towards daily life	12.Story about cancer journey
3.Possible treatment options	8.Support for cancer journey	13.Daily life with chemotherapy
4.Recurrence treatment	9.Pathology	14.Chemo-related side effects
5.Questions for treatment plan	10.Negative emotions about treatment	15.Miscellaneous

Table 1. Fifteen topics by LDA topic modeling.

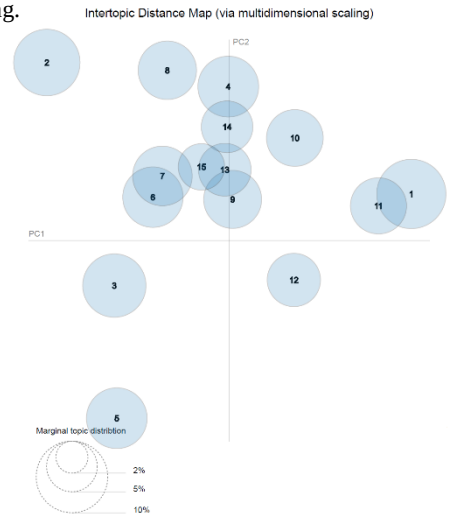


Figure 1. Topic modeling results from pyLDAvis

Conclusion: The topics classified in this corpus reflected the diverse information needs of OvCa patients. OHCs can extract more daily life cancer challenges and difficulties from cancer patients. Clinicians can prioritize their time with patients and caregivers to discuss their most concerned topics from our study. This study shed light on generating meaning from unstructured text. Future research can explore the topic differences and interconnectedness between cancer stages and treatment phases with text network analysis.

References

1. Haase K. R. (2019). The role of the internet in the cancer experience: Synthesizing patient and provider views to forge new directions for care. Canadian oncology nursing journal = Revue canadienne de nursing oncologique, 29(3), 204–209.
2. Bender, J. L., Jimenez-Marroquin, M. C., & Jadad, A. R. (2011). Seeking support on Facebook: A content analysis of breast cancer groups. *Journal of Medical Internet Research*, 13(1).
3. Lee, Y. J., Park, A., Roberge, M., & Donovan, H. (2020). What Can Social Media Tell Us About Patient Symptoms: A Text-Mining Approach to Online Ovarian Cancer Forum. *Cancer Nursing*.

Enabling Data Liquidity for Health Data Science: Initial Experiences with a Suite of APIs for EHR Data

Ryan Craig, MMCi[1], Shelley Rusincovitch, MMCi[2], Ricardo Henao, PhD[2], Ursula Rogers[1,2]
[1]Analytics Center of Excellence, Duke Health Technology Solutions Durham, NC;
[2]AI Health Institute, Duke School of Medicine, Durham, NC

Background

Gaining secure, reliable, and efficient access to the right data sources is critical to the development of health data science. Through a partnership with the Duke AI Health Institute, the DHTS Analytics Center of Excellence (ACE) designed and deployed an initial suite of APIs (application programming interfaces) that has allowed machine learning (ML) processes to programmatically access Electronic Health Record (EHR) data within a basis of appropriate compliance and adherence to patient privacy. The success of these APIs has laid the foundation for further expansion across both structured and unstructured EHR data, and has achieved an important goal in promoting data liquidity.

Design

The rationale for creating these APIs originated from a multitude of EHR data requests for various ML research projects. It was clear that while the cohorts differed, the data elements being requested were often very similar. These one-off data requests take time and analyst effort, therefore the capability to extract EHR data elements directly into the ML pipeline became desirable. API design involved curating common data elements, using PCORNet CDM as a base, while building in flexibility with query parameters for targeted data retrieval. Swagger design documents were created based on the REST standard Open API spec 3.0. Since our ML projects often require a large inventory of detailed data points, the US Core Data for Interoperability (USCDI) was considered and expanded upon during design, ensuring compatibility with the Fast Healthcare Interoperability Resources, Release 4 **(**FHIR R4) standard and other standards endorsed by the Office of the National Coordinator (ONC).

The initial use case driving the development of the API suite (Table 1) was a set of ML projects using natural language processing (NLP) with unstructured clinical narrative. Projects such as Prostate mpMRI, which analyzes biopsy results in pathology reports, successfully piloted the use of these first APIs. Subsequently, the need for other EHR data to accompany the text was realized and a full-suite of the most commonly used structured EHR domains were implemented as APIs; enabling sufficient structured and unstructured data for ML processes and providing the ability to pull a patient's full clinical story.

Unstructured data	*Structured data*
Clinical Notes	Demographics
OR Notes	Encounters
Radiology Reports	Procedures
Pathology Reports	Diagnosis
MyChart Messages	Ordered Meds

Figure 1. APIs currently available

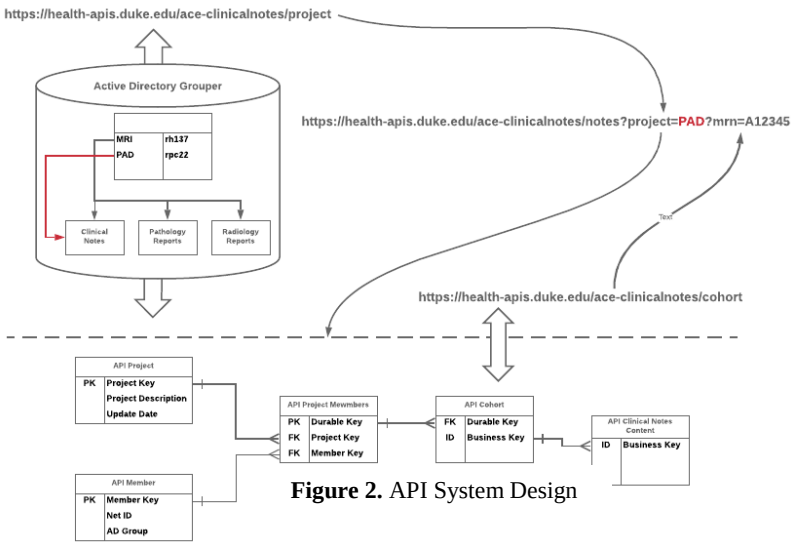


Figure 2. API System Design

Implementation

Figure 1 lists the initial use case APIs and additional APIs developed for complementary structured EHR data. Figure 2 illustrates the system design, which incorporates programmatic filtering to bind patient cohorts to projects and projects to end users; and Active Directory (AD) Groupers, leveraging Duke's existing security to control access to projects. User access to these APIs takes place in the Duke Protected Analytics Computing Environment (PACE).

Future Development

As the need for real-time data becomes a high priority, future enhancements will allow for the same resource to support both retrospective analysis and near real-time events without placing technical burden on the production EHR. This will be accomplished by pointing the APIs to a real-time data lake. This shift will allow for downstream developers to focus on creating action-oriented applications, support a breadth of data science needs and help us reach our ultimate goal of data liquidity.

Application of an informatics needs assessment to guide discussions on the development of a roadmap for an informatics-savvy injury center

Melvin Crum[1], MS, Kamran Ahmed[1], MD MS, Mamadou Misbaou Diallo[1], MD MPH, Jeffrey Gordon[1], PhD
[1]Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Introduction

An informatics-savvy health agency vision statement is a strategic resource guiding its development. A clear informatics vision is essential to prioritizing strategic objectives as well as organizing informatics resources to improve the efficacy and efficiency of evidence-based decision-making. Driving these improvements is the need to assess a health departments informatics-savviness across all phases of its operations. This presentation highlights the preliminary findings of the National Center for Injury Prevention and Control's (NCIPC) Office of Informatics (OI) baseline informatics needs assessment and describes how its outcomes guided discussion in developing an informatics-savvy injury center roadmap.

Approach

The NCIPC informatics needs assessment was adapted from the Public Health Informatics Institutes (PHII) Informatics-savvy Health Department Self-Assessment Tool[1]. The assessment was completed by NCIPC informatics staff and included 26 questions evaluating the centers informatics needs around three focus areas, including: the informatics vision and strategy, information system modernization efforts, and workforce development. The assessment outcomes were used to set strategic priorities and guide discussions for developing the NCIPC informatics-savvy injury center roadmap.

Discussion

The informatics needs assessment found that efforts around the office of informatics vision and strategy as well as workforce development plan had been documented, but the process was not fully systematized; while ongoing and systematized efforts were underway with regards to the offices system modernization efforts. These findings informed the development of an actionable roadmap to increase the center's informatics capacity. The development of a strategic roadmap around the center's data and information systems modernization efforts and informatics vision and strategy were prioritized as outcome data from these focus areas were highly-variable and the components most closely aligned with the CDC's Data Modernization Initiative, an effort aimed to transform public health data systems and save lives[2]. Additionally, the development of the NCIPC strategic roadmap guided the development of NCIPC informatics assets toolbox and engagement plan. Future efforts will build on the Office of Informatics current work by enhancing NCIPC informatics systems usability, functionality, and interoperability, increasing NCIPC's access and use of its informatics assets, and providing input for targeted trainings to support workforce development efforts.

Conclusion

This presentation shows how an informatics needs assessment outcomes can be used to guide discussions on the development of an informatics-savvy injury center roadmap. Additionally, the informatics needs assessment added value to the center's work by supporting the development of a center-wide informatics engagement plan as well as helped to prioritize implementation of strategic objectives identified within the roadmap.

References

1. Public Health Informatics Institute. Informatics-Savvy Health Department: A Self-Assessment Tool [Internet]. 2019 [cited 2020 Aug 13]. Available from: https://www.phii.org/info-savvy/self-assessment-tools
2. Centers for Disease Control and Prevention. Data Modernization Initiative Catalog of Data Resources [Internet]. 2020 Jul 31 [cited 2020 Aug 13]. Available from: https://www.cdc.gov/surveillance/dmi/index.html

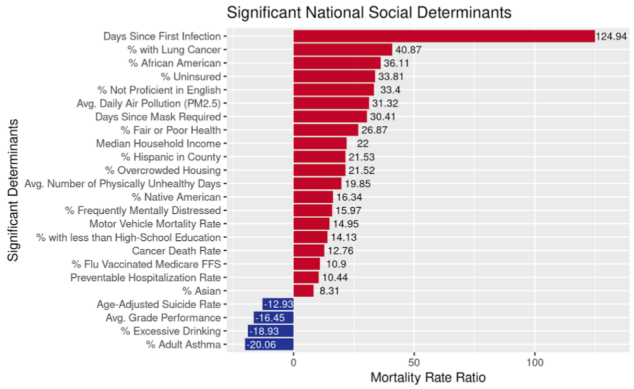
Social Determinants Associated with COVID-19 Mortality in the United States

Shayom Debopadhaya[1], Ariella D. Sprague[1], Hongxi Mou[1], Tiburon L. Benavides[1], Sarah M. Ahn[1], Cole A. Reschke[1], John S. Erickson, PhD[1], Kristin P. Bennett, PhD[1]
[1]Rensselaer Polytechnic Institute, Troy, NY

INTRODUCTION: The US has experienced hundreds of thousands of COVID-19 deaths. These deaths are not uniformly distributed, and several counties are experiencing higher-than-average death rates. It is hypothesized that social determinants have contributed to these disparities in COVID-19 mortality. Social determinants, such as a county's access to healthcare, rates of education, indicators of health, environment quality, and economic status can affect the impact of a disease, so a growing amount of literature highlights the need to identify the social determinants of COVID-19 [1]. Yet, there is a lack of diverse screening of the social determinants of COVID-19.

METHODS: To address gaps in the literature and leverage readily available data, this study identifies which social determinants are associated with significant changes in COVID-19 mortality rate at the county level as of July 6, 2020. Uniquely, we account for a comprehensive list of comorbidities and the impact of differing state policies such as closings and reopenings. The study consists of an initial study to find significant high-risk factors from 24 reported in the literature. Then, using the significant high-risk factors as controls, 41 social determinants are evaluated to find those associated with COVID-19 mortality. Both the initial study to identify controls and the association study utilize negative binomial mixed models to analyze county level data (n=3093 counties), statistically corrected for possible false discoveries using the Benjamini-Hochberg Procedure. Model performance and fit are validated using a procedure from Wu et al. [2]. Further details on the data sources and analysis are described in the full manuscript [3].

RESULTS: The statistically significant results (Benjamini-Hochberg corrected p-value < 0.05) from the association analysis are shown. The social determinants are listed on the y-axis. The x-axis shows the Mortality Rate Ratio, which is the change in COVID-19 death rate per unit increase of social determinant. The social determinants that increase risk of mortality are red, while determinants that decrease the risk of mortality are blue. Our models showed robustness in this process, as the false discovery rate was acceptably below 0.05 at 0.0217.



DISCUSSION: After adjusting for high-risk factors and differing state policies, we identify that ethnic minorities, poor access to healthcare, immigrants, socioeconomic inequalities, and early exposure to COVID-19 are associated with increased COVID-19 mortality, while the prevalence of asthma, suicide, and excessive drinking are associated with decreased mortality. Overall, our results indicate that social inequality puts disadvantaged populations at risk, which must be addressed through future policies and programs. We also reveal possible relationships between lung disease, mental health, and COVID-19. Because of the limitations of an ecological study, these associative relationships found in county-level data need to be further explored on a clinical level.

References

[1] F. Ahmed, N. Ahmed, C. Pissarides, and J. Stiglitz, "Why inequality could spread COVID-19," *The Lancet. Public Health*, vol. 5, no. 5, p. e240, 2020.

[2] X. Wu, R. C. Nethery, B. M. Sabath, D. Braun, and F. Dominici, "Exposure to air pollution and COVID-19 mortality in the United States," *medRxiv*, 2020.

[3] S. Debopadhaya, A. D. Sprague, H. Mou, T. L. Benavides, S. M. Ahn, C. A. Reschke, J. S. Erickson, and K. P. Bennett, "Social determinants associated with COVID-19 mortality in the United States," *medRxiv*, 2020.

Compound Prioritization via Ranking and Graph Representation Learning

Vishal Dey, BS[1], Xia Ning, PhD[1,2,3]
[1]Computer Science and Engineering, The Ohio State University, Columbus, OH;
[2]Biomedical Informatics, The Ohio State University, Columbus, OH;
[3]Translational Data Analytics Institute, The Ohio State University, Columbus, OH

Introduction

In silico methods have been extensively developed to prioritize promising drug candidates in order to speed up drug discovery. To tackle the compound prioritization problem, learning-to-rank methods[1] have recently gained attention. However, these methods typically use fixed molecular fingerprints to represent compounds. Taking advantages of Graph Neural Network (GNN) representation learning, we develop a comprehensive learning-to-rank method for effective compound prioritization that jointly learns molecular graph representations via GNN and a scoring function using the representations. We denote our method as gnnCP.

Methods

We consider the compound prioritization problem as to correctly rank compounds in terms of their binding affinities with respect to a protein target. To achieve so, gnnCP represents compounds using latent features that are learned from molecular graph structures via a new, self-attention directed message passing neural network (A-DMPNN). A-DMPNN generates the compound representation, using pooling with a self-attention mechanism over the learned atom features out of directed message passing (DMPNN)[2]. A linear scoring function is then applied on the learned representations to score and rank the compounds. The gnnCP methods minimizes a ranking-based objective function that approximates the fraction of miss-ordered pairs in the ranking list. We evaluate all the methods using a set of 105 single-target bioassays from PubChem Bioassay via 5-fold cross validation using concordance index (CI), recall@k (R@k), Normalized Discounted Cumulative Gain@k (ndcg@k)[1], R@k% and ndcg@k%. We compare gnnCP with the following feature vectors with the same scoring and loss functions: (i) binary Morgan fingerprints (Morgan), (ii) Morgan count fingerprints, (Morgan-c), (iii) bioassay-specific compound features computed using Tanimoto coefficient on binary Morgan fingerprints (Morgan-ba)[1], and (iv) 200-dimensional RDKit descriptors[2] (RDKit200).

Results

Table 1 presents the performance comparison between A-DMPNN, DMPNN and the baselines. We observe that A-DMPNN and DMPNN perform significantly better than all the baselines across all performance metrics. This demonstrates that the learned representation out of gnnCP can effectively encode useful molecular substructure information, and thus are more effective for compound prioritization. We also observe that A-DMPNN achieves substantial improvement over DMPNN especially in terms of recall metrics. Unlike mean pooling in DMPNN, attention mechanism in A-DMPNN can differentially focus on atoms based on the relevance of each atom to the prioritization problem.

Table 1: Overall Performance Comparison

method	CI	R@3	R@5	ndcg@3	ndcg@5	R@5%	ndcg@5%
Morgan	0.7059	0.5429	0.6442	0.8138	0.8163	0.4200	0.8384
Morgan-c	0.7109	0.5454	0.6545	0.8145	0.8193	0.4371	0.8461
Morgan-ba	0.6874	0.5003	0.6255	0.7892	0.7971	0.3752	0.8160
RDKit200	0.6869	0.5194	0.6316	0.7897	0.7965	0.3962	0.8128
DMPNN	<u>0.7310</u>	<u>0.6432</u>	<u>0.7086</u>	<u>0.8542</u>	<u>0.8469</u>	<u>0.5790</u>	<u>0.8960</u>
A-DMPNN	**0.7482**	**0.6857**	**0.7402**	**0.8808**	**0.8674**	**0.6857**	**0.9361**

The best/second best performance under each metric is **bold**/<u>underlined</u>.

Discussion and Conclusion

We developed a comprehensive learning-to-rank method gnnCP to better rank the compounds based on their binding affinities. Our experimental results demonstrate that gnnCP significantly outperforms the molecular fingerprint-based methods in compound prioritization. Future work could include 1) interpretation of self-attention weights and understanding compound substructures that are important for prioritization, 2) compound prioritization via GNN with respect to multiple compound properties, etc.

References

[1] Liu J, Ning X. Differential Compound prioritization via bidirectional selectivity push with power. Journal of Chemical Information and Modeling. 2017;57(12):2958–2975.

[2] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling. 2019;59(8):3370–3388.

Analysis of EHR-driven Standardized Nursing Data to Explore Psychospiritual Care for Hospitalized Patients

Fabiana C. Dos Santos, MSN[1], Tamara G.R. Macieira, PhD[1], Yingwei Yao, PhD[1], Hwayoung Cho, PhD, RN[1], Olatunde O. Madandola, MPH, RN[1], Ragnhildur Bjarnadottir, PhD, MPH[1], Diana J. Wilkie, PhD[1], Karen Dunn Lopez, PhD[2], Gail M. Keenan, PhD, RN[1]

[1]University of Florida, Gainesville, Florida, [2]University of Iowa, Iowa City, Iowa

Introduction

Spirituality has been identified as a crucial component of nursing care and includes experiences of connectedness with the transcendent or the sacred. Psychospiritual interventions go beyond and encompass psychotherapy techniques with spiritual practices to help patients achieve spiritual and psychological growth in life stressors such as terminal illness.[1] Existing evidence suggests that patients become more involved with religious practices in clinical settings, and a sense of spirituality magnifies with age.[2] To date, little is known about the nature and frequency of psychospiritual interventions provided by nurses and their impact on patient outcomes. The purpose of this study was to examine psychospiritual care and its relationship with patient age by analyzing nursing care of hospitalized patients documented in electronic health records (EHRs) coded with standardized terminologies.

Methods

A secondary data analysis was conducted using de-identified nursing care plan data from 9 units in 4 hospitals retrieved from the Hands-on Automated Nursing Data System (HANDS), an electronic application and database that nurses use to enter and track standardized nursing diagnoses, interventions, and outcomes during a patient's continuous stay on a unit (i.e., episode of care)[3] producing interoperable data. Based on the Nursing Interventions Classification (NIC), we determined psychospiritual intervention as the plan of care having one or more of the following 9 interventions during a hospital stay: Spiritual Growth Facilitation; Spiritual Support; Presence; Active Listening; Therapeutic Touch; Meditation Facilitation; Religious Ritual Enhancement; Emotional Support; and Calming Technique. Descriptive analysis, group comparison, and binary logistic regression were used to examine psychospiritual care provided and its association with patient age. Fixed effects estimation approach was adopted at the institution level to account for hospital and unit unobserved characteristics.

Results

Among 34,466 episodes of care, psychospiritual interventions delivered by nurses were provided in 3,858 (11%). The number of psychospiritual interventions in each episode ranged from 1 to 4, with 1 intervention delivered in 93% of episodes. The nursing interventions commonly provided were Active Listening (59%), Emotional Support (21%), and Calming Technique (16%). 15% of patients between 18 to 64 years old received psychospiritual care compared with 8% of those 65 and older. The prevalence of psychospiritual interventions by hospital/unit can be found in Table 1. The small community hospital (3/Med) provided psychospiritual care to 54% of patients, in contrast to 2%-14% in other hospitals. In the small community hospital, the primary outcome for patients who received psychospiritual intervention was pain control (94%). The delivery of psychospiritual interventions was associated with the patient being younger ($p = .002$).

Table 1. Psychospiritual nursing interventions by episodes and settings

Hospital/Unit	1/Geron	1/ICU	1/Med	2/Geron	2/Med	3/Med	4/Cardiac	4/ICU	4/Neuro
Episodes (n)	7,536	691	4,231	1,490	3,151	3,953	5,079	1,323	7,012
Psychospiritual (%)	2%	11%	14%	6%	8%	54%	8%	6%	2%

1. Large Community Hospital; 2. Large Community Hospital; 3. Small Community Hospital; 4. University Hospital.

Conclusion

Using interoperable nursing data, we generated evidence that psychospiritual care is being provided differentially in hospital medical-surgical units and more frequently to younger patients suffering from pain. This study provides insight into the needs and characteristics of psychospiritual care in hospital units and may guide future nursing care plans. While we found being younger to be significantly associated with receiving psychospiritual interventions, further research is needed to understand this relationship. Future studies are planned that will examine the relationships among subsets of psychospiritual interventions, age, diagnoses, and outcomes.

References

1. Corwin D, Wall K, Koopman, C. Psycho-Spiritual Integrative Therapy: Psychological intervention for women with breast cancer. Journal for Specialists in Group Work 2012;37(3): 252-273.
2. Moberg D.O. Research in spirituality, religion, and aging. Journal of Gerontological Social Work 2005;45:11-40.
3. Keenan G, Yakel E, Yao Y, et al. Maintaining a consistent big picture: meaningful use of a web-based POC EHR system. Int J Nurs Knowl 2012;23(3):119-33.

Prediction of Gestational Diabetes Mellitus in Overweight and Obese Caucasian Women using Machine Learning

Yuhan Du, BE[1], Fionnuala M McAuliffe, MD[1], Catherine Mooney, PhD[1]
[1]University College Dublin, Dublin, Ireland

Introduction

Gestational Diabetes Mellitus (GDM) is an adverse pregnancy complication linked to many short- and long-term consequences for both mothers and babies. We explored machine learning techniques to develop models to predict GDM in overweight and obese Caucasian women in early second trimester. Our preliminary results based on baseline maternal characteristics and blood biomarkers are presented in this manuscript.

Method

This research is a secondary analysis of the PEARS study (ISRCTN 29316280), a randomized controlled trial on the prevention of GDM using a behavioural antenatal lifestyle intervention in overweight and obese women[1]. As the majority of the participants are Caucasian, we focused on this ethnic group only in this research.

The candidate features for the early prediction of GDM are maternal anthropometry, demographic characteristics and blood biomarkers at 14.91 ± 1.65 weeks gestation. Features with greater than 20% missing values were excluded and the remaining were imputed using multiple imputation by chained equations (MICE). 70% of the women were randomly selected as the training set and the 30% as an independent test set. An ensemble of p-values from Mann-Whitney-U test, the Pearson product-moment and the Spearman rank correlation coefficients using a fast correlation based filter, beta-coefficients of logistic regression, error-rate-based and Gini-index-based variable importance measure in random forest, was used to select the features. Synthetic Minority Oversampling Technique was applied to balance the training set. Five machine learning algorithms (C5.0 decision tree, random forest (RF) and support vector machine (SVM) with linear, polynomial and radial kernel) were trained in five-fold cross validation repeated five times optimizing the area under precision-recall curve (AUC-PR). The models were evaluated on the independent test set.

Results

Of the 439 Caucasian women included in this research, 13.90% (61) were diagnosed of GDM. The features selected are: maternal age, weight, fasting glucose, insulin and C-peptide. As shown below, SVM with polynomial kernel performed best with highest AUC-PR of 0.60 and second highest area under Receiver Operating Characteristics curve (AUC-ROC) of 0.81 on the test set. At 5% and 10% false positive rate, this model achieved sensitivity of 0.44 and 0.67 respectively.

	C5.0	Random Forest	SVM Linear	SVM Polynomial	SVM Radial Basis
AUC-PR	0.53	0.48	0.58	0.60	0.55
AUC-ROC	0.79	0.79	0.81	0.81	0.83

Discussion

This research explored the development of prediction models for GDM with a novel focus on the overweight and obese Caucasian group while carefully addressing the class imbalance problem. The models achieved good performance, showing potential in assisting the early prediction of GDM in a clinical setting. Further research will be conducted on data modeling using remotely accessible maternal characteristics only to reduce hospital visits during the COVID-19 pandemic.

References

1. Kennelly MA, Ainscough K, Lindsay KL, O'Sullivan E, Gibney ER, McCarthy M, et al. Pregnancy exercise and nutrition with smartphone application support: a randomized controlled trial. Obstetrics & Gynecology. 2018;131(5):818–826.

Open Integrated Analysis of Multi-institutional Data using ICEES

Karamarie Fecho, PhD[1*], Stavros Garantziotis, MD[2], Ashok Krishnamurthy, PhD,[1] Emily Pfaff,[3] Charles Schmitt, PhD[2], Shepherd Schurman, MD[2], Samantha Shuptrine, MPP[4], Hao Xu, PhD[1] Stanley Ahalt, PhD[1]

[1]Renaissance Computing Institute, University of North Carolina, Chapel Hill, NC, USA; [2]National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA; [3]North Carolina Clinical and Translational Sciences Institute, University of North Carolina, Chapel Hill, NC, USA; [4]Social & Scientific Systems, Silver Spring, MD, USA

Introduction

The Integrated Clinical and Environmental Exposures Service (ICEES) is a novel, regulatory-compliant, disease-agnostic service that provides open access to UNC Health clinical data that have been integrated at the patient level with a variety of public environmental exposures data. ICEES is accessible via an OpenAPI, either by command-line calls or a Swagger user interface. ICEES allows users to openly and rapidly conduct exploratory statistical analyses designed to uncover important relationships between exposures and health outcomes. Importantly, we have validated ICEES in the context of an asthma use case.[1] The Environmental Polymorphisms Registry (EPR) is a ~20-year study of nearly 20,000 participants. The EPR is based at the National Institute of Environmental Health Sciences (NIEHS), which is located ~13 miles from UNC Hospitals, creating overlap in UNC Health patients and EPR participants. Given the overlap, and recognizing the value of integrating UNC Health clinical data, EPR survey and SNP data, and public exposures data, we aimed to: (1) quantify the overlapping population; (2) securely transfer a subset of EPR data to UNC; (3) integrate the data; (4) openly expose the data via ICEES; and (5) use ICEES to replicate independent UNC[1] and EPR[2] findings on asthma.

Methods

All study procedures were approved by the Institutional Review Boards at UNC and NIEHS. For integration of UNC and EPR data, we used a custom software tool termed DAREL and created a crosswalk using the following identifiers: date of birth; first three letters of first name; first eight letters of last name; and sex. These identifiers were used as they were determined to be optimal in a pilot study. We estimated overlap with UNC Health for both the overall EPR cohort and an asthma-specific cohort. We then integrated the UNC Health and EPR data, with rows representing patients and/or participants and columns representing UNC Health or EPR variables. The de-identified integrated feature table was then exposed via the ICEES OpenAPI. The data were accessed using the Swagger interface and command-line CURLs. ICEES results were returned as JSON output and in tabular form. ICEES queries focused on two dependent variables that are indicative of asthma exacerbations: emergency department (ED) or inpatient visits for respiratory issues (UNC Health metric) and self-reported ED visits for asthma (EPR metric). Race and exposure to airborne particulate matter were chosen as independent variables as prior independent work at UNC[1] and EPR[2] demonstrated their significance. The significance threshold was set at $\alpha = 0.05$.

Results

In December 2019, 2,770,607 patients were part of UNC Health, and 19,388 participants were enrolled in the EPR. 7,233 EPR participants were also UNC Health patients (37.3% of all EPR participants). 4,130 EPR participants were included in the EPR asthma cohort. Of those, 947 (22.9%) had a self-reported diagnosis of asthma and complete survey and SNP data. Of these 947 EPR participants, 218 (23.0%) were also UNC Health patients and included in the UNC asthma cohort. We then queried ICEES to determine the impact of race and exposure to particulate matter on asthma exacerbations, using both UNC and EPR metrics. Our results indicated an increase in the proportion of asthma exacerbations among African Americans compared to Caucasians ($P < 0.001$ EPR) and persons exposed to relatively high levels of particulate matter ≤ 2.5-μm diameter compared to those exposed to lower levels ($P < 0.001$ UNC, $P < 0.001$ EPR). While the effect of race was not significant for the UNC metric, a trend was apparent.

Conclusion

Our results demonstrate that ICEES can be used to openly access integrated data from UNC Health and EPR and conduct rich integrative statistical analyses designed to generate insights into asthma and other diseases.

References

1. Fecho K, Pfaff E, Xu H, Champion J, Cox, Stillwell L, et al. J Am Med Inform Assoc. 2019;26(10):1064-73.
2. Schurman SH, Bravo MA, Innes CL, Jackson WB, McGrath JA, Miranda ML, et al. Sci Rep. 2018;8(1):12713.

**Presenting author. Apart from first/lead and last/senior authors, all others are listed alphabetically.*

Funding support from NIH (OT3TR002020) and Intramural Research Program of NIEHS.

Challenges in capturing comorbidities using Medicare data

Alexander Fiksdal, PhD[1]; Dana De Alasei, MA, MS[1]; Hojjat Salmasian, MD, MPH, PhD[1,2]
[1]Brigham and Women's Hospital, Boston, MA; [2]Harvard Medical School, Boston, MA

Introduction

Using comorbidities for risk-adjusting clinical outcomes (e.g., mortality) is a standard practice. Claims data, coded using International Classification of Diseases (ICD) codes, are often used to capture comorbidities. This approach is not only widely used in health systems research, but also is used by Agency for Healthcare Quality and Research (AHRQ) to calculate risk-adjusted outcome measures for each hospital based on claims data submitted for Medicare patients. Similar measures and risk-adjustment approaches are used by various ranking and reimbursement programs focused on healthcare quality. Medicare only accepts up to 25 diagnosis codes for each admission, while many admissions are assigned more than 25 diagnosis codes. Our objective was to quantify how many comorbidities may be missed due to this arbitrary threshold and compare those with comorbidities that are captured in the Medicare data.

Methods

We used in-house coded diagnosis data for one year of inpatient discharges from one large academic medical center in the northeast United States. We used the Elixhauser comorbidity index for this study;[1] specifically, we used the AHRQ version.[2] We restricted the data to patients with Medicare as their primary or secondary insurer. No other exclusion criteria were applied. The data was run through the Elixhauser algorithm twice; once including all diagnosis codes for each admission, and once using only the first 25 diagnosis codes. We compared the output descriptively and using standard parametric statistical tests. The study was exempted from IRB review.

Results

Of 15,394 encounters (10,828 patients) included in the study, 2,415 (15.7%) were assigned more than 25 diagnosis codes. At least one Elixhauser comorbidity was exclusively associated with the diagnosis codes 26 or after in 735 cases (4.8% of all admissions). Table 1 shows the most common comorbidities captured in the first 25 diagnosis codes (i.e. in Medicare data), as well as the most common comorbidities found exclusively in position 26 and after (i.e. not in Medicare data). The mean number of comorbidities per encounter was 3.39 when restricted to the top 25 diagnoses, and 3.46 when including all diagnoses ($p < .001$). Due to the logic of the AHRQ Elixhauser algorithm, we found 2 edge cases (0.01%) where the inclusion of diagnosis 26 and after resulted in the removal of a comorbidity that would be captured using only the first 25 diagnosis codes.

	First 25 diagnoses	Diagnosis 26 and after
1	Hypertension	Deficiency anemias
2	Fluid and electrolyte disorders	Obesity
3	Renal failure	Valvular disease
4	Chronic pulmonary disease	Hypertension
5	Depression	Neurological disorders
6	Deficiency anemias	Renal failure
7	Hypothyroidism	Depression
8	Diabetes with complication	Diabetes with complication
9	Weight loss	Chronic pulmonary disease
10	Congestive Heart Failure	Fluid and electrolyte disorders

Table 1 – Most common comorbidities capture using data included in or excluded from Medicare submission.

Conclusion

For a substantial number of encounters, Medicare's limit on the number of diagnosis per encounter resulted in incomplete capture of comorbidities at a large academic medical center. The missed comorbidities appeared categorically different than those captured through the first 25 diagnosis codes. This limits the accuracy of studies that use Medicare data and utilize comorbidity indices for risk adjustment.

References

1. van Walraven C, Austin PC, Jenings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Medical Care. 2009 (47):626-633.
2. Healthcare Cost and Utilization Project. *Elixhauser Comorbidity Software Version 3.7*. Available at https://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp [accessed July 24, 2020]

Preliminary Experience with Teleophthalmology in Residency Education

Michael J. Fliotsos, B.S.[1], Abdulaziz Alaqueel, B.S.[1], Michael V. Boland, M.D., Ph.D.[1], Yesha S Shah B.S.,[1] Fasika A. Woreta, M.D., M.P.H.[1]
[1]Johns Hopkins Medicine Wilmer Eye Institute, Baltimore, MD, USA

Abstract (50-75 words)

In this preliminary study, we present our institution's initial experience with a teleophthalmology tool for use by ophthalmology residents to share findings with supervising ophthalmologists during emergency department consultations. Over a six-month period, the device captured over 400 images during 56 unique clinical encounters. The device was used most often in assessment of optic nerve pathology such as papilledema and optic neuropathy (30%), visual acuity/visual field deficits (18%), and visual changes related to diabetic retinopathy (13%).

Introduction

Emergency department (ED) attendance has been steadily increasing in the United States over the past several decades, from 108 million visits in 2000 to 130 million visits in 2010.[1] Integration of telemedicine services is one approach to reduce the increasing burden on EDs. As telemedicine continues to expand, it may also be useful in augmenting resident education. The use of fundus photography and ocular coherence tomography (OCT) imaging in the ED setting among ophthalmology resident physicians to communicate their examination findings remotely to supervising ophthalmologists has not been studied. The current study presents preliminary data on the implementation of teleophthalmology tools in resident education at our institution.

Methods

The study was conducted at the Johns Hopkins Hospital Wilmer Eye Institute (Baltimore, MD, USA) and was approved by the Institutional Review Board. Our ophthalmology residency program acquired a TopCon 3D OCT-1 Maestro System (TopCon Medical Systems Inc., Oakland, NJ, USA). Five first-year ophthalmology residents were trained on use of the device in capturing OCT images and photos of the retina and were asked to utilize the device in their evaluation of patients presenting to the ED with urgent eye complaints. Use of the device was at the discretion of the resident based on the clinical scenarios for which such images would be useful for diagnosis and clinical management, such as measurement of optic nerve thickness or documentation of retinal lesions. Residents communicated their findings remotely with supervising ophthalmologists via the electronic health record, where images were exported. Images were extracted from the device and retrospective chart review was conducted to obtain information on demographics and the final diagnosis recorded by the ophthalmology resident. Images obtained on the device were reviewed by two independent graders for subjective image quality on a scale of 1 (lowest) to 3 (highest).

Results

From December 1st, 2019 to May 25th, 2020, the device was used to assess patients in 56 unique encounters, capturing 453 fundus photos and OCT images of the retina and optic nerve (average 8 images per encounter). The average age of patients was 50.5 years old and 45% were male. In terms of race/ethnicity, 48% of patients were white/Caucasian, 36% were African-American, 9% were Hispanic/Latino, and 7% were Native American or Asian. Average subjective image quality was 1.8 out of 3 between the two independent raters. The imaging device was utilized most commonly for pathology of the optic nerve, such as confirmed or suspected papilledema (n=12, 22%) and optic neuropathy (n=5, 9.8%); this was followed in prevalence by assessment of new-onset visual acuity/visual field defects (n=10, 18%) and visual changes relating to diabetic retinopathy with or without hemorrhage (n=7, 13%). Less common diagnoses included retinal tear/detachment (n=5, 9%), workup of embolic events such as stroke and central retinal artery occlusion (n=4, 7%), and workup of traumatic eye injuries (n=4, 7%).

Conclusion

Teleophthalmology tools in our residency program were utilized by resident physicians most often in assessment of potential pathology of the optic nerve, followed by non-specific visual changes and pathology of the retina relating to diabetes. Future research will focus on potential improvements to the resident user experience, such as troubleshooting for images of suboptimal quality, as well as an imaging protocol for specific diagnoses that can compare outcomes of patients assessed with and without teleophthalmology.

Reference

1. Channa R, Zafar SN, Canner JK, Haring RS, Schneider EB, Friedman DS. Epidemiology of eye-related emergency department visits. JAMA Ophthalmology. 2016;134(3):312-319.

Disparities in patient portal enrollment and telehealth use among oncology patients

Meera Garriga, BA[1], Sumi Sinha, MD[1], Nishali Naik[1], Brian W. McSteen, BS[1], Sasha Yousefi[1], Anobel Odisho, MD, MPH[1], Amy Lin, MD[1], Lauren Boreta, MD[1], Julian Hong, MD[1]
[1]University of California, San Francisco, San Francisco, CA, USA

Abstract
Patient portals allow patients to access their medical information and communicate with providers, which may be particularly important to coordinate complex oncologic care. Disparities in patient portal usage may become more exaggerated as patients become increasingly reliant on remote communication during the COVID-19 pandemic. Here, we identify disparities in portal enrollment amongst oncologic patients and show that portal activation is associated with increased telehealth visits.

Introduction
Care for oncologic patients requires multidisciplinary, longitudinal coordination. Patient portals allow patients to access their medical information from electronic health records (EHR) and easily communicate with providers, which can improve treatment coordination and increase patient participation in their care. Unfortunately, disparities in portal usage may affect care. The COVID-19 pandemic may exaggerate the downstream effects of these disparities as patients become increasingly reliant on remote communication. Building on prior smaller studies, we present a large longitudinal study utilizing a right-censored approach to investigate patient portal enrollment among oncology patients. The objective of this study was to evaluate disparities in time to portal enrollment across age, race, ethnicity, language, and marital status, and assess the relationship between portal enrollment and telehealth use.

Methods
We conducted a retrospective review of all adult oncology patients seen at the University of California, San Francisco cancer center from January 2011 to December 2019. Data regarding patient demographics, telehealth encounters, and portal enrollment was extracted from the EHR. Patient time to enrollment in the portal over the study period and associations with demographic characteristics were assessed using the Kaplan-Meier method, log-rank test, and Cox proportional hazards method. The relationship between portal enrollment and telehealth visits (conducted on a platform independent from the portal) in patients with a single portal enrollment status and at least 10 total visits was assessed using multivariate logistic regression (the 22% of patients who switched enrollment status partway through the study were excluded). Analysis was conducted in Python 3 and R (Version 4.0.1).

Results
Among 261,027 patients, 128,516 (49%) activated their portal account over the study period. Median time to activation was 278 days. Patients who did not activate in this period had median follow-up of 14 days. Primary non-English speakers (HR 0.40, 95% CI 0.39-0.41; $p<0.001$), Black patients (0.54, 0.52-0.55; $p<0.001$), Hispanic or Latino patients (0.74, 0.73-0.76; $p<0.001$), single patients (0.74, 0.73-0.75; $p<0.001$), and male patients (0.92, 0.91- 0.93; $p<0.001$) were less likely to activate. Activation decreased with age (0.99 per year, 0.99-0.99; $p<0.001$). These disparities persisted in a multivariate model.

Of the 4,095,865 encounters that occurred over the study period, 736,438 (18%) were conducted via telephone or video. Of these, 508,289 (69%) were with patients with activated portals. Controlling for demographic factors, activated patients had a greater likelihood of having had at least one telehealth visit (OR 1.32, 95% CI 1.23-1.41); $p<0.001$). Activated patients also had more telehealth visits over the study period compared to non-activated patients (median 5, IQR 1-11 and 1, 0-5, respectively; $p<0.001$). Black and Hispanic/Latino patients were more likely to have at least one telehealth visit compared to white patients (1.18, 1.03-1.36; $p=0.02$ and 1.38, 1.21-1.57; $p<0.001$).

Conclusion
Despite high levels of overall enrollment over time, there are significant disparities in telehealth engagement among oncology patients. Patients who do not speak English as a primary language and those who identify as Black or Hispanic/Latino have the lowest rates of portal enrollment. Portal activation is associated with an increased likelihood of telehealth visits, although it is interesting to note that race and ethnicity has less of an effect on telehealth than portal enrollment. Addressing disparities in digital access, particularly among non-English speakers, should be prioritized to prevent exacerbations to already existing disparities in cancer care during the COVID-19 response.

Explore the Usage of a Multi-modal Social Risk Decision Support Tool in Primary Care and Patient Characteristics

Weiwei Ge, MS[1], Suranga N. Kasthurirathne, PhD[1,2], Joshua R. Vest, PhD, MPH[1,2]
[1]Indiana University, Indianapolis, IN; [2]Regenstrief Institute Inc., Indianapolis, IN

Background

A pilot study[1,2] was implemented to better address safety-net patients' need for wraparound services referral using a clinical decision support system (CDSS). This project is to explore usage of the CDSS that facilitated referrals to social service providers in response to real-world disruptions and changes. The CDSS provided individualized patient risk scores to primary care providers in two fashions: a scheduled line listing report (batch) that summarized groups of patients simultaneously and a near-real time one patient view via a user interface (UI) to support one-on-one patient care. Three key events occurred after CDSS go-live: 1) a security fix that took UI offline; 2) UI returned after upgrades and feature enhancements; and 3) the COVID-19 pandemic (stay-at-home order began on 3/24/2020).

Method

Using system access log files, we tracked the weekly access of the 441 batch listings and the 444 UI accesses between April 2019 and May 2020 across 15 provider locations. We described the temporal usage graphically and with frequencies. For patients whose scores were accessed through the UI, we were able to link patient characteristics using electronic health records (EHR) to describe the demographics, comorbidities, insurance types, emergency department (ED) revisit and inpatient stay readmission of the patients.

Preliminary Results

Scheduled patients (batch) appointments increased after stay-at-home order began and one patient perview (UI) decreased overtime, especially after stay-at-home order began.

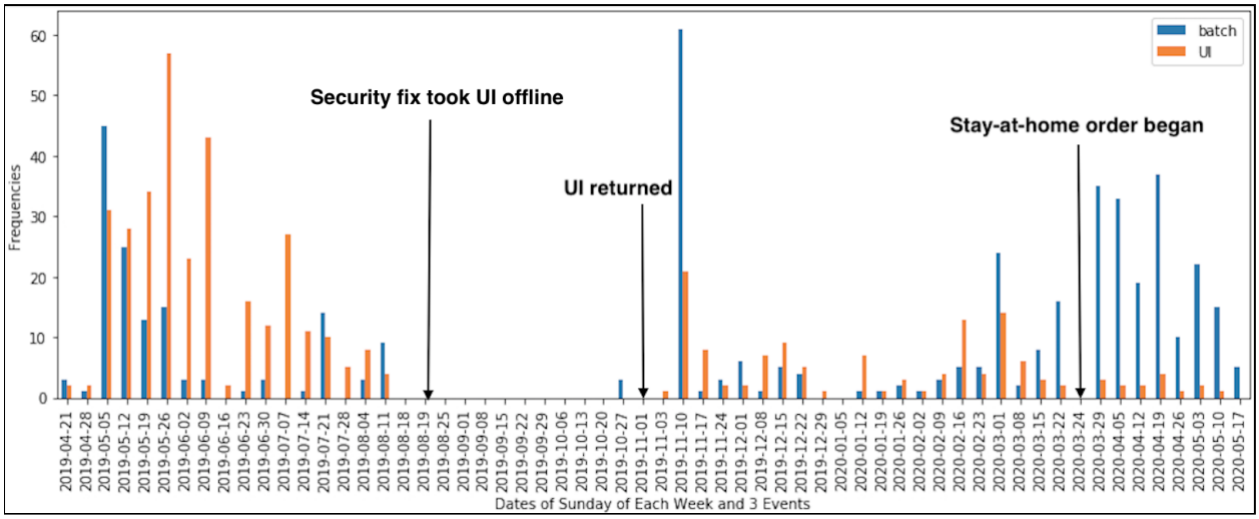


Figure 1. CDSS usage trend by week.

References

1. Vest JR, Menachemi N, Grannis SJ, Ferrell JL, Kasthurirathne SN, Zhang Y, et al. Impact of risk stratification on referrals and uptake of wraparound services that address social determinants: a stepped wedged trial. American journal of preventive medicine. 2019;56(4):e125-e33.
2. Kasthurirathne SN, Vest JR, Menachemi N, Halverson PK, Grannis SJ. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. Journal of the American Medical Informatics Association. 2018;25(1):47-53.

Content-based Recommendation Systems to Improve Reusability of Gene Expression Omnibus Datasets

Braja Gopal Patra, PhD[1,2], Kirk Roberts, PhD[2], Hulin Wu, PhD[2]
[1]Weill Cornell Medicine, Cornell University, New York, NY, USA.
[2] The University of Texas Health Science Center at Houston, Houston, TX, USA

Introduction

Recent scientific discoveries have generated an extensive amount of data and these datasets are being stored in different repositories. Most of these datasets are used only once. Several biomedical dataset repositories are available integrated with search engines, which can help researchers looking to find specific types of datasets. DataMed (https://datamed.org) or Google Dataset Search may be helpful to researchers in finding relevant datasets. Researchers who generally want to find datasets related to their interests, but do not have a particular interest in mind, could benefit from a dataset recommendation system. Further, it is difficult for researchers to keep track and search for new datasets related to their research field in the complex search environment. A dataset recommendation system will help solve the above searching problem and reduce the effort for searching suitable datasets. Once a dataset is found, a researcher may need to find relevant literature in the dataset's domain. We developed content-based recommendation systems for finding datasets suitable for researchers and literature in the domain of datasets.

Materials and Methods

We collected the researcher's publications from PubMed using a web crawler and 122,222 datasets from Gene Expression Omnibus (GEO) repository for developing the dataset recommender. To identify the researchers' area of research, we implemented the Dirichlet process mixture model (non-parametric clustering algorithm) to cluster publications into several research areas. The top datasets were then recommended for each cluster using cosine similarity of dataset vectors and publication cluster vectors. Datasets (title and summary) and publications (title and abstract) were converted to vectors using TF-IDF.[1] For developing a literature recommender, we collected PubMed publications from MEDLINE archive. The most similar publications were recommended for each dataset using BM25.[2] Figure 1 provides an overview of dataset and literature recommenders.

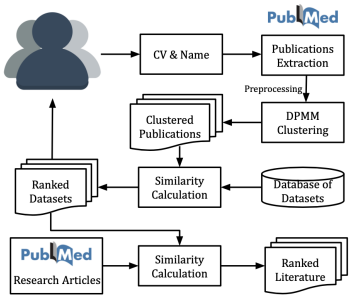


Figure 1: Overview of recommendation systems.

Results

The dataset recommendation system was evaluated by five annotators (with 32 publications on average). Each researcher judged 40 recommended datasets on average by providing 1 to 3 stars. The cluster-specific dataset recommender achieved the maximum precision at 10 (P@10) strict (S) and partial (P) of 0.31 and 0.45, respectively. Here, we divided the cluster-specific P@10s with the number of clusters and then averaged over all evaluators. We also merged the recommended datasets together for all publication clusters based on Round-robin algorithm. Now each researcher can have only 10 recommended datasets, which resulted better than previous and it achieved the P@10 (S) and P@10 (P) of 0.61 and 0.78 based on the five annotators. Three annotators reviewed the literature recommended for 36 datasets. The literature recommendation system achieved the maximum P@10 (S) and P@10 (P) of 0.83 and 0.90, respectively.

Conclusion

To the best of the authors' knowledge, there were no such recommendation systems available. A user-friendly and efficient web-based platform that implements these recommenders is freely accessible at http://genestudy.org/. We hope these recommenders can improve the re-usability of GEO datasets. These recommendation systems can easily be extended to other datasets.

References

1. Patra BG, Roberts K, Wu H. A Content-Based Dataset Recommendation System for Researchers - A Case Study on Gene Expression Omnibus (GEO) Repository. Database. 2020;2020.

2. Patra BG, et al. A Content-Based Literature Recommendation System for Datasets to Improve Data Reusability-A Case Study on Gene Expression Omnibus (GEO) Datasets. Journal of Biomedical Informatics. 2020;p. 103399.

PrecisionFDA Brain Cancer Predictive Modeling and Biomarker Discovery Challenge

Yuriy Gusev, PhD[1,2], Krithika Bhuvaneshwar, MS[1,2], Anas Belouali, MS[1], Samir Gupta, PhD[1], Adil Alaoui, MS, MBA,[1], Holly Stephens[3], Sean Watford[3], Zeke Maier[3], Elaine Johanson[4], and Subha Madhavan, PhD[1,2]
[1]Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington DC; [2]Lombardi Comprehensive Cancer Center, Georgetown University; [3]Booz Allen Washington DC; [4]Precision FDA, Washington DC

Introduction

Brain cancer is the tenth leading cause of death among both men and women[1] and also the second most common cancer among children of age 0-14. It affects more than 700,000 people in the US every year[2]. The 5-year survival rate for the most common malignant brain tumor - Glioblastoma (GBM) was only 5.5% in 2019. While scientific research in understanding the biology of these tumors has helped improve survival of patients with these deadly tumors, further investigation is necessary to develop targeted therapies that can be personalized to a patient's genetics, clinical diagnosis and treatment history. Biomarkers must be identified and validated to allow for modern clinical trials to be designed based on matching molecular features of tumors to targeted therapeutics.

The Food and Drug Administration (FDA) is leveraging crowdsourcing for regulatory science advancement through their precisionFDA platform. It allows for analyses of large biological datasets, while encouraging collaboration, interaction, and data sharing. precisionFDA engages the public to advance regulatory science through crowdsourcing challenges, soliciting voluntary contributions from a group of individuals. Recently, the Lombardi Comprehensive Cancer Center, Innovation Center for Biomedical Informatics at Georgetown University Medical Center and precisionFDA launched the Brain Cancer Predictive Modeling and Biomarker Discovery Challenge[1].

Methods

The dataset used for this challenge was the REMBRANDT data collection[2] which is publicly available via the NCBI GEO repository[3]. It is a large brain cancer dataset that included a total of 671 brain cancer patients with clinical data that included tumor stage, grade and outcome (overall survival status). 541 patients in this dataset had gene expression data and 263 patients had DNA copy number data.

In the challenge, participating teams were asked to develop supervised machine learning and/or artificial intelligence models to identify biomarkers and predict patient outcome (overall survival status) using gene expression, DNA copy number, and clinical data from the REMBRANDT dataset. The challenge was set up in three sub-challenges. In *sub-challenge 1* (SC1), participants were provided with gene expression data, clinical phenotype and outcome data. In *sub-challenge 2* (SC2), DNA copy number data, clinical phenotype and outcome data were provided. In *sub-challenge 3* (SC3), participants were provided a combination of both gene expression and DNA copy number data, clinical phenotype and outcome data. The data for the challenge was released in two phases – Phase 1 and Phase 2. During Phase 1 of the challenge, participants used the Phase 1 dataset to develop machine learning models and identify the most important model features. During Phase 2, participants applied their models to predict Alive/Dead outcome status for patient samples in the Phase 2 dataset; and this was used to score model performance.

We aimed to rank participant teams whose models provided a short list of most informative features for brain cancer, and wrote an evaluation algorithm that would automatically rank the phase 2 submissions based on three metrics – accuracy, sensitivity and specificity. This scoring was done for each sub-challenge. SC3 was given twice the importance as SC1 and SC2 since it contained multiple data types that made the model building and prediction more complex. An overall score was calculated from the individual scores in each sub-challenge. In addition to the evaluation algorithm, badges were awarded to the top 5 performing teams based on several criteria including model robustness; extra credit based on short listed features for potential use in biomarker research; extra credit for utilizing domain knowledge; and overall documentation, usability and overall presentation of results.

Results

The Biomarker Discovery Challenge received 30 submission during Phase 1 and 22 submissions during Phase 2. A wide variety of machine learning models were used by participants, with gradient boosting frameworks and ensemble based methods being the most popular. The top-performing model used 46 features, selected from 40 genes, 4 cytobands, and 2 clinical attributes.

Conclusion

The PrecisionFDA Brain Cancer Predictive Modeling and Biomarker Discovery Challenge demonstrated that crowdsourcing challenges can help identify the most effective machine learning algorithms and also detect a small number of novel molecular features suitable for development of predictive brain cancer biomarkers

References

1. precisionFDA. https://precision.fda.gov/challenges/8. Last Accessed Aug 27, 2020
2. Gusev, Y. et al. The REMBRANDT study, a large collection of genomic data from brain cancer patients. Scientific Data. 2018.
3. NCBI GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108474 . Last Accessed Aug 27, 2020

Uncovering the potential biological mechanisms between a chemical and a phenotype

Christopher Hawthorne, MSc[1], Guillermo H. Lopez-Campos, PhD[1]
[1] Wellcome-Wolfson Institute for Experimental Medicine, Belfast, Northern Ireland, United Kingdom

Introduction

An individual's exposome is an integral part of human health and disease development and can assist the goals of precision medicine and health studies through tailored medical treatment[1]. Biomedical informatics is highlighted as a field prepared to analyse and further exposomic data and research. We previously described a methodology and R package called phexpo[2] which is able to establish potential chemical - phenotype relationships through shared genes. There is a need to provide a better understanding of the potential underlying biological mechanisms driving these relationships. This would provide a holistic biological insight into how environmental influences utilize biological mechanisms to cause or affect a phenotype and allow for greater direction and hypothesis generation in biomedical informatics research. To address this aspect, we have explored the possibility of building on top of our phexpo methodology, aiming to identify the biological mechanisms as pathways and/or Gene Ontology (GO) terms that underlie a chemical and phenotype relationship. We have prototyped this approach analysing warfarin, an anti-coagulant applied in clinical settings against phenotypes such as deep vein thrombosis, to demonstrate the application of this method and how it can provide greater biological comprehension.

Methods

The methodology utilises phexpo to identify bidirectional chemical-phenotype relationships and uses the associated genes between a chemical and a phenotype to generate the potential biological mechanisms (represented by MSigDB v7.0's[3,4] gene set collections) by using a Fisher's exact test. The MSigDB v7.0 dataset was further preprocessed before utilization, with the removal of Kyoto Encyclopedia of Genes and Genomes and BioCarta datasets. Additionally, outdated or non-human gene information was removed using *Homo sapiens* gene info from NCBI (2020-01-14).

Results and Discussion

For warfarin, chemical-phenotype relationships were filtered using a Bonferroni corrected p-value 5E-4 and results were restricted to the top three mechanisms (four if p-value was tied). We identified 5 phenotypes, 176 unique mechanisms, an average of 82 mechanisms per phenotype, a max of 150 (Venous thrombosis) and a minimum of 34 (Abnormality of prothrombin). Warfarin's mechanisms display its known relationship as an anticoagulant through GO and Reactome terms (e.g. GO_REGULATION_OF_COAGULATION, REACTOME_COMMON_PATHWAY_OF_FIBRIN_CLOT_FORMATION, …). This methodology expands the phexpo analyses but also inherits its limitations. Our results show how this methodology can provide a deeper biological relationship between chemicals and phenotypes than previously described.

Conclusion

We believe the application of this methodology represents a novel and deeper insight into the potential mechanistic relationships that link a chemical and phenotype and will provide an opportunity for greater research hypothesis creation and exploration for exposome research.

References

1. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW. The exposome: molecules to populations. *Annu Rev Pharmacol Toxicol*. 2019;59:107-127.
2. Hawthorne C, Simpson DA, Devereux B, López-Campos G. Phexpo: a package for bidirectional enrichment analysis of phenotypes and chemicals. *JAMIA Open*. 2020;3(2):173-177.
3. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550.
4. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425.

Resynthesizing MIMIC-III Personally Identifiable Information Tags to Increase Corpus Utility: Process and Impact Assessment

Paul M. Heider, PhD[1], Gary Underwood, MS[2], Stéphane Meystre, MD, PhD[1,2]
[1] Medical University of South Carolina, [2] Clinacuity, Inc., Charleston, SC

Introduction: The MIMIC-III (Medical Information Mart for Intensive Care) database[1] is a popular resource used across biomedical informatics with about 2 million unstructured clinical text notes from real patient care for natural language processing (NLP) research and development. Due to the necessity of de-identifying the notes to protect patient privacy, personally identifiable information (PII) was replaced with artificial tags. NLP models using MIMIC-III notes must be designed to either treat these artificial tags as natural text or skip spans of text which look like the tags at the risk of ignoring real data. We propose that resynthesizing PII (i.e., replacing with realistic surrogates) will result in a more useful corpus. We present our process with baseline metrics and assess its impact on NLP tasks.

Methods: PII tags in the MIMIC-III text notes come in a range of patterns consisting of a named pattern and an optional indexing number surrounded by a string of brackets and stars (e.g., '[** Known lastname **]' or '[** Hospital 2 **]'). No official list of tags exists. We used pattern-based string extraction (specifically, *grep*) iterated with manual inspection to curate such a list. We mapped each named pattern to categories commonly used in de-identification tasks. We then used the PII resynthesis engine adapted from CliniDeID® to resynthesize the tags. The original pattern type and mapped category guided the content and form of the surrogate (e.g., 'GS' vs. 'Spelvin' vs. 'George Spelvin'). The first author reviewed 266 random tags and their contexts to evaluate the mappings of patterns to categories.

We investigated downstream effects of resynthesizing notes to understand how common uses for the MIMIC-III corpus would be impacted. We hypothesized that a production NLP system would do a better job of extracting information from the resynthesized notes because the surrogates provide realistic contexts. To test this hypothesis, we ran 1000 random notes through Clinacuity's CliniWhiz system (which extracts medical problems, medications and attributes, lab tests and results, and allergens) on both corpora and manually inspected differences. Output differences were categorized such that false positives (FP) denoted an incorrect extraction in only one output corpus and false negatives (FN) denoted a correct extraction in only one output corpus. We also hypothesized that a word embedding model would perform better on NLP tasks when generated from the resynthesized corpus. To test this hypothesis, we generated a word model from each corpus with fastText. Each model was used to train five separate Bi-LSTMs (bidirectional long short-term memory) for named entity recognition (NER) and de-identification.

Results: In our manual inspection of tag context, we found that 72% of the categories were appropriate for the context (that is, name surrogates appeared in contexts that indicated a redacted name). Three percent were correctly identified as PII but not of the correct form (e.g., tagged the wrong type of name) or not the correct type of PII (e.g., phone number vs. ID). The remaining 25% are not likely PII. Lab values were often tagged as numeric identifiers or dates. Lab names or procedures were tagged as names or cities. Full tables of counts and frequencies of named patterns and categories will be included in the poster. The CliniWhiz outputs differed on 1108 annotations (~5%). The original corpus output had more FNs (363 vs. 135). The resynthesized corpus output had more FPs (169 vs. 70). The original text yielded fewer annotations. Finally, the average resynthesized corpus Bi-LSTM model's precision, recall, and F_1-score were significantly higher ($p=0.018$, $p<0.0001$, and $p<0.0001$) for the 2014 i2b2 de-identification shared task[2] while precision ($p<0.0001$) but not recall ($p=0.065$) was significantly lower for NER on the 2010 i2b2 shared task[3].

Conclusion: Resynthesizing PII tags in MIMIC-III notes appears to have a net positive effect on common uses for the corpus. More work needs to be done to refine the mapping from pattern to category to surrogate form, including the possible need to correct individual named pattern instances in the source corpus. Patching those spans of text incorrectly flagged as PII with realistic surrogates may also further increase the utility of this corpus.

References

1. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):160035.
2. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. 2015 Dec;58(Suppl):S11–9.
3. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011 Jun 16;18(5):552–6.

Semi-Automated Corpus Augmentation Methods for Enriching Laboratory Test Names with Value Annotations

Paul M. Heider, PhD[1], Youngjun Kim, PhD[1], Stéphane M. Meystre, MD, PhD[1,2]
[1]Medical University of South Carolina, Charleston, SC; [2]Clinacuity, Inc., Charleston, SC

Introduction: Reference annotated datasets are key to any machine learning-based approach to natural language processing (NLP). The 2010 Integrating Biology and the Bedside (i2b2) NLP challenge corpus[1] is a broad-purpose resource with de-identified clinical notes annotated for problems, tests, and treatments. The 2019 n2c2 NLP challenge corpus[2,3] normalized a subset of these annotations to Unified Medical Language System (UMLS) concept unique identifiers (CUIs). To mitigate the time and money investment of further manual augmentation, we present a general workflow for extending a gold standard corpus semi-automatically to create a 'silver' reference standard corpus. Our particular case adds laboratory test values to names but the general workflow can be implemented for other data gaps.

Methods: The first stage of augmentation is a two-stage filter to reduce the types of annotations to just laboratory test annotations. Filter 1 flags all concepts for enrichment with the UMLS semantic type of "T059" or "T034" as they indicate laboratory tests. Filter 2 flags all concepts annotated with the concept category of PROBLEM or TEST and one of eight UMLS semantic types: T195, T007, T123, T004, T129, T121, T005, or T127. Given the relatively small number, these concepts were manually reviewed for inclusion by the last author. The next stage of augmentation was a rule-based search for laboratory test values around known laboratory test names. In order, we searched for numeric expressions (e.g., "1.3", "30%") after the concept, then before the concept, and then categorical values (e.g., "positive") in either direction. The first match, if any, was annotated. The next (optional) stage was to manually verify the discovered values and their relation to a laboratory test name to create a reference corpus for the other stages. The last author used WebAnno, a web-based annotation tool, to add missed values and correct erroneous relations. Finally, we trained a deep neural network (Bi-LSTM)-based sequence labeling model on the output of the rule-based approach (above) to explore boot-strapping a larger corpus from one annotated only with laboratory test names.

Results: The final set of extracted laboratory test values were analyzed with respect to their location relative to the laboratory test name and the text between the two. Table 1 presents the frequencies of the most common intervening text templates for the 1,106 cases of a laboratory test name followed by a value. For the NER task of annotating laboratory test values, the rule-based annotator had a precision of 91.16, recall of 80.54, and F_1-score of 85.52. Averaging over five runs, the Bi-LSTM annotator had a precision of 93.31, recall of 90.17, and F_1-score of 91.71.

Table 1. Most Frequent Categories of Intervening Textual Material Between a Test Name Followed by a Value

Value Type	(Blank)	Copula	Preposition	Change Verb	Other Lab Value	Temporal	Other Lab Name
Categorical	26.0%	36.1%	3.0%	3.6%	7.1%	7.1%	0.6%
Numerical	58.9%	12.0%	15.3%	3.9%	3.0%	1.1%	1.5%

Conclusion: We have presented a general framework for adapting available corpora to related but more specific needs. These 'silver' standard corpora can be used to boot-strap annotation in an even larger corpus at a level competitive with state-of-the-art techniques (cf. Xu et al.[4] with an F_1-score of 95.54). Our analysis of the context surrounding and intervening between laboratory test names and values should help other developers improve their own automated systems. Our augmentations will be made available via GitHub (https://github.com/musc-tbic) in a privacy preserving manner. Future work includes boot-strapping other corpora and back-porting CUIs from the 2019 to the 2010 corpus.

Acknowledgements: This work was supported in part by the SmartState endowment and a SCRA SACT grant.

References

1. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011 Jun 16;18(5):5526.
2. n2c2: National NLP Clinical Challenges, (n.d.). https://n2c2.dbmi.hms.harvard.edu/
3. Luo Y-F, Sun W, Rumshisky A. MCN: A comprehensive corpus for medical concept normalization. Journal of Biomedical Informatics. 2019 Apr 1;92:103132.
4. Xu J, Li Z, Wei Q, et al. Applying a deep learning-based sequence labeling approach to detect attributes of medical concepts in clinical text. BMC Medical Informatics and Decision Making. 2019 Dec 5;19(5):236.

Evaluation of Data Management of COVID-19 Clinical Trials Using a Cloud-Based Clinical Data-Management Platform

Rezzan Hekmat, MHS[1], Dilhan Weeraratne, PhD[1], Courtney Van Houten, MA[1], Brett R. South, MS, PhD[1], Nawshin Kutub, PhD[1], Van C. Willis, PhD[1], Walker Bradham, BS[1], Robert DiCicco, PharmD[1], Gretchen P. Jackson, MD PhD[1,2], Jane L. Snowdon, PhD[1]
[1]IBM Watson Health, Cambridge, MA, USA, [2]Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

The COVID-19 pandemic presents a unique opportunity for understanding the impact of data-management technologies on streamlining and fast-tracking execution of clinical trials. Drug discovery, pre-clinical testing, and human trials are lengthy and expensive processes. IBM Clinical Development (ICD) is a cloud-based clinical data management system (CDMS) platform that is differentiated by an integrated suite of modules (e.g., automated data integration, patient engagement, advanced reporting and analytics, medical coding supported by machine learning, randomization and trial supply management, endpoint adjudication) in a single platform. This study's goal was to evaluate how ICD can support and facilitate data-management for two COVID-19 vaccine and therapeutic drug candidate clinical trials sponsored by different mid-size pharmaceutical companies and conducted by Veristat.

Methods

ICD usage data and aggregated Veristat operational data (e.g., de-identified study attributes, ICD utilization, and key study milestones, cycle times, and performance indicators) were collected, summarized with descriptive statistics, and compared with industry benchmarks. Surveys and semi-structured interviews of Veristat clinical data managers were conducted to evaluate satisfaction, usability, workflow, performance, impact, and remote use during the pandemic. Interviews were recorded, transcribed, and analyzed using an open coding approach. Thematic analysis was defined deductively by reviewing the direct responses to interview questions and inductively by identifying emerging patterns.

Results

The average time to design and release the COVID-19 study databases in ICD was 10 business days compared to Veristat's usual 30-40 days, a 66.7-75.0% reduction, and to an industry baseline of 68.3 days, an 85.4% reduction (Wilkinson M et al, PMID: 29714600). User interviews (n=3) and surveys (n=4) highlighted ICD's flexibility, customizability, and ease-of-use as the main factors contributing to rapid trial launch. First, the system easily allowed code reuse in database creation, and users routinely consulted a template library to generate new builds efficiently. Bypassing *de novo* coding in database creation allowed users to focus efforts on study-specific customizations. Second, ICD's mid-study update feature allowed users and sponsors to work in a parallel rather than serial fashion, speeding up trial start dates and protocol amendment. Users could 'split' releases of trial databases allowing sponsors to initiate trial start up while completing the databases. In comparison, an industry baseline assessment found one-third of companies are "often" or "always" releasing study-specific databases after their first patient first visit (Wilkinson M et al, PMID: 29714600). In second releases of the databases, users added or made updates to study specific electronic case report forms, edit checks, and rules. Each trial database had 3 mid-study updates after the first go live, and all were associated with protocol amendments. Trial 2 had two additional protocol amendments prior to the release of any part of the database, resulting in five protocol amendments. Compared to industry benchmarks of 1.8 amendments per protocol for a phase 1 clinical trial (Getz K et al, PMID: 30227022), protocol amendments in Veristat's COVID-19 trials were more frequent, likely due to rapidly evolving science about the virus. 75% of survey respondents thought ICD helped to reduce the impact of mid-study updates. Usage data revealed databases were locked for only one hour on average to execute mid-study updates. 100% of respondents thought ICD aided data management and facilitated working from home. Average usability, as measured by the System Usability Scale (SUS), was 80.6, which is considered excellent.

Conclusion

Rapid execution of clinical trials is critical to addressing the global public health and economic crises created by the COVID-19 pandemic. ICD facilitated effective and efficient clinical trial start up and data management for two COVID-19 drug and vaccine trials by making database design and changes easy and flexible.

Increased utilization of telemedicine in cancer patients during the COVID-19 pandemic

Hu T. Huang, PhD[1], Suwei Wang, PhD[1], Irene Dankwa-mullan[1], MD, Gretchen P. Jackson, MD PhD[1,2], Yull Arriaga[1]*, MD, Dilhan Weeraratne, PhD[1]*

[1]IBM Watson Health, Cambridge, MA, USA

[2]Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

The COVID-19 pandemic has significantly strained healthcare systems and disrupted patient care globally. Telemedicine has emerged as a viable alternative to provide quality care while accommodating reallocation of healthcare resources and meeting social distancing guidelines. Cancer patients are especially susceptible to COVID-19 morbidity and mortality. Care of cancer patients during the pandemic has centered on balancing cancer management while reducing the risk of infection. In patients at high risk, it has been recommended to reduce outpatient visits and postpone elective procedures to minimize potential exposure to COVID-19. In this study, we examined the impact of the COVID-19 pandemic on telemedicine utilization and routine outpatient care in cancer patients.

Methods

A retrospective de-identified observational study cohort was selected from the IBM® Explorys® electronic health record (EHR) database to assess telemedicine utilization and outpatient care data in cancer patients during the COVID-19 pandemic (February 01-July 31, 2020), compared with the baseline period of October 1-December 31, 2019. The study cohort included patients of all ages with any type of cancer at any stage. The cohort was stratified into cancer patients with and without a COVID-19 diagnosis, and the service utilization was compared among three periods defined as: October-December 2019 (pre-COVID), February-April 2020 (early-COVID) and May-July 2020 (late-COVID). Telemedicine and outpatient care services include any type of services and only patients with ongoing cancer diagnosis in each period were considered for comparison. The Cochran-Armitage test and two-sample proportions test were employed to examine the trends in usage across these time periods.

Results

A cohort of 2,114,839 cancer patients were identified based on the criteria, and telemedicine utilization and outpatient care of cancer patients during the three periods is shown in Table 1. In cancer patients without COVID-19, telemedicine utilization significantly increased over 60-fold from the pre-COVID to the late-COVID periods ($p<0.0001$). In contrast, the number of cancer patients receiving outpatient care significantly decreased during the three time periods ($p<0.0001$). In cancer patients with COVID-19, telemedicine utilization rate was significantly greater in the late-COVID compared to the early-COVID period ($p<0.0001$), whereas the outpatient utilization rate was significantly less in the late-COVID period ($p<0.0001$).

Table 1: Telemedicine and outpatient care in cancer patients in the COVID-19 pandemic era

Patient Groups	Time Period	Active Cancer Patients	Cancer Patients Using Telemedicine		Cancer Patients Using Outpatient Care	
		N	N	%	N	%
Cancer patients without COVID-19	pre-COVID	2,060,343	3,114	0.15	756,634	36.72
	early-COVID	1,335,878	98,860	7.4	459,848	34.42
	late-COVID	1,348,761	123,975	9.19	444,265	32.94
Cancer patients with COVID-19	early-COVID	5,786	1,176	20.32	3,720	64.29
	late-COVID	9,787	2,403	24.55	5,943	60.72

Conclusion

This study demonstrated a significant increase in the utilization of telemedicine in the high-risk population of cancer patients during the COVID-19 pandemic. In cancer patients who contracted COVID-19, both telemedicine utilization and outpatient visits doubled in comparison to those without COVID-19, which demonstrates intensified demand for clinical services in this population. Increased outpatient care in cancer patients with COVID-19 is likely due to management of the COVID-19 infection itself along with oncologic care that cannot be postponed such as chemotherapy. Finally, a comparison of the early- and late- COVID-19 periods showed increased telemedicine utilization and decreased outpatient visits during the late period. These findings suggest that oncologists embraced telemedicine to adapt their clinical workflows to minimize outpatient visits and protect this vulnerable population.

***Co-senior authors**

Telemedicine Utilization Among Non-Hospitalized Patients with COVID-19

Hu T. Huang, PhD[1], Elisabeth Scheufele, MD, MS[1], Irene Dankwa-Mullan, MD, MPH[1], Gretchen P. Jackson, MD, PhD[1,2], Suwei Wang, PhD[1]
[1]IBM Watson Health, Cambridge, MA; [2]Vanderbilt University Medical Center, Nashville, TN

Introduction

The COVID-19 pandemic has radically transformed the delivery of ambulatory care. Telemedicine visits have seen slow uptake, but during the pandemic, new models to support compensation for virtual care along with relaxation of restrictions on delivering such care across state lines allowed for marked increases in the adoption of telemedicine[1]. This option was particularly advantageous for mildly ill patients with COVID-19, as in-person visits put healthcare workers and other patients at risk of contracting the disease. We studied the sociodemographic and clinical factors associated with utilization of telemedicine to manage patients with COVID-19 in the ambulatory care setting.

Method

We obtained data on COVID-19 patients from the IBM® Explorys® electronic health record (EHR) database, which provides real-world, near real-time, deidentified longitudinal patient-level clinical data from over 330,000 providers and more than 72 million unique patients in the US. The cohort included adult patients (18 years and older) with confirmed COVID-19, who did not have an inpatient or emergency department encounter, from January 1 to July 31, 2020. We assessed the patient characteristics (age, sex, race, and ethnicity), the number and type of encounters, acute symptoms at the time of diagnosis, as well as underlying comorbidities including diabetes mellitus (type 1 and type 2), chronic obstructive pulmonary disease (COPD), cardiovascular disease (CVD), hypertension, cancer, renal disease, liver disease, and hyperlipidemia.

Results

Table 1. Ambulatory Telemedicine Utilization by Comorbidity and Acute Symptoms of Patients with COVID-19

Comorbidity conditions (# patients)	Count (%) of patients using ambulatory telemedicine	Acute symptoms (# patients)	Count (%) of patients using ambulatory telemedicine
Cancer (N=13250)	1883 (14.2%)	Cough (N=3376)	299 (8.9%)
Liver Disease (N=1773)	206 (11.6%)	Fever (N=2192)	138 (6.3%)
COPD (N=2771)	319 (11.5%)	Fatigue (N=733)	55 (7.5%)
Hyperlipidemia (N=14471)	1550 (10.7%)	Diarrhea (N=555)	71 (12.8%)
CVD (N=11529)	1203 (10.4%)	Headache (N=537)	29 (5.4%)
Hypertension (N=17647)	1524 (8.6%)	Vomiting (N=382)	18 (4.7%)
Diabetes (N=8319)	649 (7.8%)	Sore Throat (N=300)	36 (12%)
Renal Disease (N=1636)	110 (6.7%)	High blood pressure (N=113)	3 (2.7%)

A total of 51,076 individuals with confirmed COVID-19 infection were identified, including 52.5% with comorbid conditions and 12.3% showing acute symptoms at the time of diagnosis. 3461 (7.2%) patients used telemedicine services for a total of 5956 telemedicine encounters (mean 1.6 per patient; range 1 to 33). Patients who had telemedicine encounters were slightly younger than those who did not (mean age 47.7 vs 48.8 years, range 18-90 years), and the majority were female (64.6%). The demographic distribution of patients using or telemedicine use by race or ethnicity was Caucasian (8.3%, 1758 of 21295), African American (5.1%, 849 of 16824), Hispanic or Latinx (6.6%, 236 of 3582), and Asian (9.4%, 42 of 449). Patients with symptoms of cough ($p<0.001$) or fever ($p=0.002$) at the time of COVID-19 diagnosis had significantly more telemedicine encounters, compared to those with headache ($p=0.12$) or sore throat ($p=0.09$). Among patients with comorbidities, patients with COVID-19 and cancer had the highest telemedicine utilization (14.2%), followed by patients with liver disease (11.6%), COPD (11.5%), hyperlipidemia (10.7%) and CVD (10.43%).

Conclusion

This real-world evidence study provides some insights into the differential use of telemedicine among patients with COVID-19 in the ambulatory care setting. Telemedicine adoption remained relatively low, likely due to the unknown course of the disease in the early pandemic. Male, African American, and Hispanic or Latinx patients had lower rates of telemedicine usage compared with female, Caucasian, and Asian patients. Patients with comorbidities had high rates of telemedicine utilization, suggesting that this care delivery model was used to protect the patients most at risk for disease complications. These preliminary results suggest ongoing need for programs and policies to promote telemedicine adoption, especially for high-risk patients.

References

1. Mehrotra A, Chernew M, Linetsky D, Hatch H, Cutler D. The Impact of the COVID-19 Pandemic on Outpatient Visits: A Rebound Emerges. [Commonwealth Fund, 2020 May 19, cited 2020 Aug 20] Available from: https://www.commonwealthfund.org/publications/2020/apr/impact-covid-19-outpatient-visits

Quantifying Record Linkage of Research Studies with CMS Claims Data

Vojtech Huser, MD, PhD, Nick Williams, MS, PhD, Craig S. Mayer, MS
[1]National Library of Medicine, National Institutes of Health, Bethesda, MD

Introduction

Researchers often use many sources to assemble a lifetime Electronic Health Record (EHR) that can support longitudinal analyses. Healthcare claims data is often linked with EHR, registry or interventional or observational clinical studies to provide (at relatively low cost) care history or follow up data or otherwise complement study-specific data collection. Since 2019, Center for Medicare and Medicaid Services (CMS) provides a dataset titled: "Projects Conducted Under Research Data Use Agreements (DUAs)" at data.cms.gov. This quarterly updated dataset provides a list of projects that use CMS Virtual Research Data Center (VRDC). We used a simple string search method to identify projects where the study title indicated linkage of CMS claims data to research studies or other datasets[1].

Methods

The CMS project list has four columns: Organization, DUA Category, Project/Study Name and Close Date. We first characterized the listed projects. Next, we selected studies with 'link' in the title and manually reviewed each project. Besides CMS project list columns, we also used a web search engine (Google), if a project title included an acronyms (e.g., SIP 11-043: GA Study on the feasibility of linking the BCCP with GCCR and Medicare') and to find study registry records, result journal publications or other related project websites/reports.

Results

The CMS DUA list contains 5,265 projects of which 2,475 (47%) are ongoing (have no close date). By 'DUA Category', 57.9% of projects use 'Identifiable files' while the remaining projects use 'Limited Data Set'. Considering the 'Organization' field, academic institutions account for 44.8% of projects. The dataset included correct listing of our own active VRDC project. The search strategy for linkage projects identified 38 projects. Three projects were removed in manual review (e.g., 'Leveraging claims data to identify linkages between neonatal abstinence syndrome and long-term patterns of care and health outcomes'). In the remaining 35 projects, five were related second versions of prior completed projects. Three linkage projects included a registry. Examples of linkage projects include (Organization: Title): Duke University PCORNet Medicare linkage and longitudinal follow-up; National Cancer Institute: SEER Medicare data linkage project; Fred Hutchinson Cancer Research Center: Women's Health Initiative and CMS Data Linkage (Full Study); Eli Lilly: A Claims-Linked Comparison of Cognition and Care Changes by Amyloid Status (C5A Study); and West Virginia University and Hospitals: Linking Medicare, WV Medicaid and WV Cancer registry data to study the burden of breast, colorectal, lung and prostate cancers in WV. The project repository at https://github.com/vojtechhuser/project/tree/master/linkage contains the full list of linkage projects, review annotations, the R code for search and analysis, and additional results.

Discussion and Conclusion

Our results clearly demonstrate tens of record linkage projects and prove a growing trend to use claims data as augmenting source for registry or clinical study. Our study fills a gap in quantifying how often record linkage is pursued and catalogues a concrete list of record linkage projects for a single federal data source. Researchers considering linking claims data can use our list to argue in favor of linkage for their study by showcasing tens of other active linkage projects (this was a significant motivation for our analysis). We highly appreciate CMS's open data initiatives and this project shows how secondary analyses of open CMS data can advance clinical research informatics research. Our study is limited by only focusing on CMS VRDC projects and using a simple search strategy. Also, having only the project title limited our manual review (6 linkage project titles were too brief or otherwise difficult to interpret). Addition of project start date (not currently provided) would also allow quantifying the linkage trend over time. This research was supported by the Intramural Research Program of the National Institutes of Health/National Library of Medicine/Lister Hill National Center for Biomedical Communications.

References

1. Didier R, Gouysse M, Eltchaninoff H, et al. Successful linkage of French large-scale national registry populations to national reimbursement data. Arch Cardiovasc Dis. 2020 doi:10.1016/j.acvd.2020.04.006

Assessing the Impact of Telepsychiatry Implementation on Polypharmacy Reduction among Youth Detainees

Humayera. Islam, MS[2,1,4], Abu. Mosa, PhD, FAMIA[1,2,4], Laine. Young-Walker, MD[3]
[1]Department of Health Management and Informatics; [2]Institute for Data Science and Informatics; [3]Department of Psychiatry; [4]Center for Biomedical Informatics, University of Missouri School of Medicine, Columbia, Missouri

Introduction

Telepsychiatry can help overcome the high rates of use of two or more psychotropic medications (polypharmacy) among the youth involved in Juvenile Justice (JJ) [1]. However, few research studies analyzed the impact of telepsychiatry in reducing polypharmacy. The purpose of our study is to evaluate the impact of telepsychiatry services for youth from JJ residential placements on the total psychotropic polypharmacy.

Methods

This study used de-identified prescription data (2013 to 2019) on a population of youth (age 11 to 17 years) serving under the Missouri Department of Youth Services (DYS) and receiving psychiatric care from a telehealth network established with the University of Missouri Department of Psychiatry (MUDP). For each unique patient identifier, a list of prescribed medications (only drugs related to psychiatric conditions) was extracted per unique encounter identifier, which was later mapped to their hierarchical classes using medication ontology. The total polypharmacy (TP) was computed as a sum of the number of medications for each patient per encounter. Z-tests were used to compare the change in TP from the first to the last televisit ("increase", "decrease," and "no change") with TP for the first visit ("No medication", "Exactly one medication", "Two or more medications", and "Three or more medications"), and for specific drug classes like antipsychotics and antidepressants. T-tests were calculated to compare mean TP across gender and race for first and last encounters. R version 3.4.4 (R Foundation for Statistical Computing, Vienna, Austria) was used for the data analysis.

Encounter	Male	Female	t-test (p-value)	White	Black	t-test (p-value)
First Visit	1.61	2.07	<0.0005*	1.74	1.60	0.0434*
Last Visit	1.63	1.86	0.0121*	1.75	1.56	0.0025*
t-test (p-value)	0.71	0.0453*		0.6191	0.2922	

Table 1: Comparison of TP among demographics. * indicate P-values that are significant

Results

Our findings showed that youth patients with two or more medications and three or more medications are more likely to have reductions in total polypharmacy compared to that of patients with one or zero medication (41% vs. 4.41%, p-value<0.00 and 50% vs. 4.41%, p-value<0.00, respectively). Moreover, the rates of antipsychotics usage dropped by 10.1% from the first encounter to that of the last. Hence, our study shows evidence of polypharmacy reduction among the delinquent youth.

		First Visit (%)	Last Visit (%)	Decrease	Increase	No Change	Z- score (p-value)
Medication Usage Count	No medication	90 (7.96)	81 (7.16)	0	46	44	59.1(<0.00)*
	Exactly one medication	500 (44.2)	502 (44.4)	29(5.8)	144 (28.8)	327	90.8(<0.00)*
	Two or more medications	541 (47.8)	548(48.5)	222(41.0)	62 (11.5)	257	120.7(<0.00)*
	Three or more medications	220 (19.5)	208(18.4)	110(50)	22 (10)	88	81.9(<0.00)*
Type of Medications	Antidepressants	624 (55.2)	657(58.1)	134(21.5)	183 (29.3)	814	9.74(0.002)*
	Antipsychotics	402 (35.5)	361(31.9)	140(34.8)	86 (21.9)	905	15.9(<0.00)*

Table 2: Comparison of Change in TP with TP in first visit. * indicate P-values that are significant

Assessing Organizational Context for Implementation

Megha Kalsy [1,2,8], Natalie Kelly[1], Stephane M. Meystre [2,3], Youngjun Kim [3], Bruce E. Bray[2], Dan Bolton [1,2], Mary K. Goldstein [4,5], and Jennifer H. Garvin [1,2,6,7]

[1] IDEAS Center SLC VA Healthcare System, Salt Lake City, UT, USA [2] University of Utah School of Medicine, Salt Lake City, USA [3] Medical University of South Carolina, Charleston, SC [4] VA Palo Alto Health Care System, CA, USA [5] Stanford University, CA, USA [6] Richard L. Roudebush VA Medical Center, Indianapolis, IN, USA [7] The Ohio State University, Columbus, USA [8] VA Northeast Ohio Healthcare System, Cleveland, OH, USA

Abstract: *We studied the context of potential use of an automated quality measurement system in the VA. We used constructs from Implementation Science and Sociotechnical Models to guide our work. We used stakeholder interviews, internal VA documents, and scientific literature to assess context. We identified themes related to sociotechnical dimensions that comprise facilitators and barriers to potential adoption of the automated system.*

Introduction: To study the context of implementing a new tool to automate a congestive heart failure (CHF) quality measurement[1] we used the Promoting Action on Research in Health Services framework[2, 3] and Sociotechnical Model for Health Information Technology[4] to facilitate implementation[5]. **Methods:** We conducted semi-structured interviews with stakeholders, an archival review of VA internal documents using the VA Intranet, and a review of VA-specific scientific literature from 2009-2015. We used an applied thematic analysis[6-8]. We combined the results of these data sources to synthesize findings. **Results:** We conducted 15 stakeholder interviews with 4 key informants and 11 subject matter experts including pharmacists, physicians, advanced practice nurses, physician assistants, and physicians in rural health care facilities. We found themes that informed our potential implementation as described in Table 1.

Table 1. Themes Related to Context of Implementation of an Automated Quality Measure at the VA
VA has an overall hardware infrastructure and architecture that makes implementation of the automated system possible
The VA is using evidence-based HIT to improve clinical and operational workflow and communication
VA has a culture of continuous quality improvement, which is enhanced by its internal organizational factors
VA has the availability of appropriate clinical content in the form of structured, unstructured, and semi-structured data, in VA electronic medical databases, that can be extracted through automated systems
The VA is using HIT as a facilitator to overcome barriers to the automation of performance measures
VA emphasizes use of clinical decision support to improve the quality of care through timely information and advisories
VA encourages the development of clinical tools and extensions to support quality improvement

Discussion: We identified relevant organizational information to inform implementation of an automated quality measure in VA. Performance metrics for CHIEF are described elsewhere[9]. The use of stakeholder engagement based on implementation science and the Sociotechnical Model of HIT coupled with the use of interview, archival, and scientific literature can provide rich data to plan potential implementation of informatics tools in a given organization. **Conclusion:** Our work demonstrated that multiple sources of data can be used in an applied thematic analysis to system design and implementation of HIT for optimal uptake and adoption.

VA Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs, the United States Government, or the academic affiliate organizations. This work was supported by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, IDEAS 2.0 Center, HSR&D project #IBE 09-069, and by HIR 08-374 (Consortium for Healthcare Informatics Research).

References:

1. Meystre S, Kim Y, Gobbel G, Matheny M, Redd A, Bray B et al. Congestive heart failure information extraction framework for automated treatment performance measures assessment. Journal of the American Medical Informatics Association. 2016;24(e1):e40-e46.
2. Kitson A, Rycroft-Malone J, Harvey G, McCormack B, Seers K, Titchen A. Evaluating the successful implementation of evidence into practice using the PARiHS framework: theoretical and practical challenges. Implementation Science. 2008;3(1).
3. Seers K, Rycroft-Malone J, Cox K, Crichton N, Edwards R, Eldh A et al. Facilitating Implementation of Research Evidence (FIRE): an international cluster randomised controlled trial to evaluate two models of facilitation informed by the Promoting Action on Research Implementation in Health Services (PARIHS) framework. Implementation Science. 2018;13(1).
4. Sittig D, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Quality and Safety in Health Care. 2010;19(Suppl 3):i68-i74.
5. Goldstein M. Using health information technology to improve hypertension management. Current Hypertension Reports. 2008;10(3):201-207.
6. Ando H, Cousins R, Young C. Achieving Saturation in Thematic Analysis: Development and Refinement of a Codebook. Comprehensive Psychology. 2014;3:03.CP.3.4.
7. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006;3(2):77-101.
8. Guest G, MacQueen K, Namey E. Applied thematic analysis. Los Angeles: Sage; 2012.
9. Garvin J, Kim Y, Gobbel G, Matheny M, Redd A, Bray B et al. Automating Quality Measures for Heart Failure Using Natural Language Processing: A Descriptive Study in the Department of Veterans Affairs. JMIR Medical Informatics. 2018;6(1):e5.

Development of a Graph Model for the OMOP Common Data Model

Mengjia Kang, MS[1], Jose A. Alvarado-Guzman, MS[2], Luke V. Rasmussen, MS[1], Justin B. Starren, MD, PhD[1]

[1]Northwestern University, Feinberg School of Medicine, Chicago, Illinois; [2]Neo4j, Inc., San Mateo, California

Introduction

Current phenotyping and systems biology research requires not only integration of large volumes of Electronic Health Record (EHR) and multi-omics data, but also capturing the multitudes of relations among the concepts. Graph databases have emerged as a promising technology for such tasks, supporting not only local analysis but also global analysis leveraging graph algorithms like Centrality, Community Detection, Path Finding or Node Embeddings[1]. Unfortunately, EHR data is rarely available in a graph format. While a naïve row-to-node conversion is possible, the resulting graph is typically attribute-heavy, resulting in suboptimal performance. To address this limitation, we developed a modelling method to convert data form the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) to the Neo4j [www.neo4j.com] graph property model.

Methods

The Successful Clinical Response in Pneumonia Therapy (SCRIPT) is a five-year systems biology study that is integrating clinical, transcriptomic, metagenomic and bacterial genomic data to support Machine Learning on host pathogen interaction. Our modeling focused on nine OMOP Standardized Clinical Data Tables (PERSON, PROVIDER, OBSERVATION_PERIOD, VISIT_OCCURRENCE, CONDITION_OCCURRENCE, DRUG_EXPOSURE, PROCEDURE_OCCURRENCE, MEASUREMENT, OBSERVATION) and four Standardized Vocabularies Tables (CONCEPT, DOMAIN, VOCABULARY and CONCEPT_CLASS) which captured the SCRIPT clinical data. Our overall strategy was to encode as much information as possible in the edge topology to take advantage of the intrinsic strengths of the graph database. In general, nominal and categorical data were converted to nodes; foreign keys to edges; and numerical values to node or edge properties as appropriate. We also implemented self-directed relationships RELATED_TO and NEXT on the Concept and VisitOccurrence node separately. The former defines the nature and type of direct relationships between any two Concepts and the later builds up the patient journey.

Results

Our finalized graph property model was implemented using a local installation of Neo4j 4.0.2 Community Edition. It includes 16 types of nodes (entities) and 22 types of edges (relationships) as well as 55 node properties. This model contains on average 3.44 attributes per node. This work is available in both a markdown and Cypher query language format in our GitHub repository, https://github.com/NUSCRIPT/OMOP_to_Graph.

Discussion

Although more data preprocessing is required to load the data into our graph property model than the naïve row-to-node conversion method, previous work has demonstrated that the analytics performance will be greatly improved[2]. This model also reduces redundancy by eliminating the denormalization (Foreign Keys) that is often added to relational databases (e.g. person_id occurs in all other OMOP tables). The model was developed for the SCRIPT project, but the transforms can be applied to other OMOP CDM v5.x databases. Our current and future work will further demonstrate graph analytics examples using the SCRIPT EHR data.

References

1. Fabregat A, Korninger F, Viteri G, Sidiropoulos K, Marin-Garcia P, Ping P, et al. (2018) Reactome graph database: Efficient access to complex pathway data. PLoS Comput Biol 14(1): e1005968. https://doi.org/10.1371/journal.pcbi.1005968
2. Alvarado-Guzmán JA, MS, Keren I, MS https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=resources:jose_alvarado_rd2gd_ohdsi_submission_2017.pdf

The Potential in Standardized Process for Dysphagia Rehabilitation after Cerebrovascular Diseases

Shogo Kato, Ph.D.[1], Satoko Tsuru, Ph.D.[2], Makoto Ide, MD, Ph.D.[3,4], Akira Shindo, MD[5], Naohisa Yahagi, MD, Ph.D.[1,2], and Shu Yamada, Ph.D.[1]
[1]Keio University, Kanagawa, Japan; [2]The University of Tokyo, Tokyo, Japan; [3]St. Mary's Healthcare Center, Fukuoka, Japan; [4]University of Occupational and Environmental Health, Fukuoka, Japan; [5]Ooguno Hospital, Tokyo, Japan

Introduction

Japan has been a super-aged society as more than 28% of the population is 65 years or older in 2019. Rehabilitation is an important treatment that affects the prognosis. Standardization of rehabilitation processes is difficult because there are few quantitative indicators and they depend largely on the individual therapists rather than medicines and/or instruments. As a result, there are differences in processes and outcomes among individual therapists and hospitals.

In this study, we aim to develop a standard process for rehabilitation after cerebrovascular diseases. Then, we aim to computerize the operation for dysphagia rehabilitation using Patient Condition Adaptive Path System (PCAPS), which is a model and/or system to structure clinical knowledge based on patient condition to implement it into hospitals.

Method

The inherent/empirical knowledge of medical professionals (Physicians, Therapists, Nurses) was visualized and structured, focusing on responsible lesions, dysfunctions, and required examinations, through discussion using matrix format among four hospitals in Japan. We designed and improved rehabilitation processes for dysphagia, language disorder, basic motion, and work activity. We organized assessments and interventions needed for dysphagia, and identified the components of assessments, interventions, and relationships between assessments and interventions.

We elaborated the process, assessments and interventions for dysphagia rehabilitation based on PCAPS. PCAPS structures clinical knowledge by both "Clinical Process Chart (CPC)," which describes the overall flow, and "Unit Sheet (US)," which describes the details. We computerized the operation for dysphasia rehabilitation on "PCAPS-Administrator," which is an existing computer application to administrate processes in PCAPS format, and implemented it retrospectively into an acute stage hospital in Japan to validate the comprehensiveness of the contents.

Results

14 responsible lesions, 14 dysfunctions, and 64 examinations were structured for cerebral infarction and cerebral hemorrhage, and the relationships between them were expressed in two types of matrix format. Rehabilitation processes were structured as CPCs, focusing on acquisition and recovery process for each dysfunction. Assessments and interventions needed for dysphagia in USs was structured by 111 assessment items and 140 intervention items. The relationships between assessments and interventions were structured by 198 records.

Rehabilitation processes were identified to have patient condition-oriented structure, and the operation was rationally computerized on PCAPS-Administrator. The standard process for dysphagia rehabilitation running on PCAPS-Administrator was operated in the acute stage hospital retrospectively to record over 100 cases of clinical process during 21 months, and the comprehensiveness of the contents was confirmed, because all cases were recordable.

Conclusion

It becomes possible to operate rehabilitation process standardized based on patient conditions. By accumulating standardized assessment data, it would be possible to evaluate the effects of interventions by outcomes such as "food form," which is the form of the meal the patient can eat, and/or process indicators such as "progress and/or speed on CPC," in addition to compare and verify the transition of patient conditions. It was still difficult and a next issue to operate the standardized process in actual clinical practice because it takes too much time.

A Pilot Study on An In-Depth Comparison Between Nurse and Physician Health Terminology from a Non-Professional Perspective.

Leen Khatib, Haleh Vatani, MS[1], Barbara Di Eugenio, PhD[1], Carolyn Dickens, PhD, RN[1], Pamela Martyn-Nemeth, PhD RN[1], Karen Dunn Lopez, PhD RN[2], Amer K Ardati, MD[1], Andrew D Boyd, MD[1]

[1]University of Illinois at Chicago, Chicago, Illinois; [2]University of Iowa, Iowa City, Iowa

Introduction

Healthcare professionals rely on interdisciplinary collaboration and teamwork to provide the ultimate care for the patient [1,2]. Patients' healthcare and health literacy can also be enhanced by allowing patients to access their medical records in order to promote more active participation in their health. Previous studies have found that providing a patient with access to their discharge summaries resulted in the patient becoming more informed about their health[3]. However, most discharge summaries only contain information from the physician's perspective. In fact, nurses tend to spend the majority of a hospital stay with the patient and are in charge of conducting many treatments which are often essential for a patient to continue at home, therefore, it is important that their documentation be accessible by patients as well.

By creating a personalized discharge summary that includes both nurse and physician notes, patients can become better informed about their health. However there exist large discrepancies between nurse and physician house healthcare terminology which could result in difficulty navigating through the different terms. The purpose of this study was to identify the discrepancies in healthcare terminologies between nurse and physicians from a nonprofessional's perspective.

Methodology

Healthcare terminologies were extracted from University of Illinois Health documents of 2 patients admitted for heart failure problems. Physician terminology was manually extracted from physician discharge summaries. Nursing health terminologies were extracted from the following documents: Nursing Diagnosis, Nursing Notes, Nursing Care Plan from a Cerner Millennium EHR and transcripts of verbal Nursing Shift Handoffs. A list of all healthcare terms for each source, and all 5 lists were compared manually for similar meanings.

Results

For patient 1, physicians and nurses shared 14 terms out of a complete list of 150 (9.3%). Physician term overlap in the nursing documents above were respectively: 8, 4, 2, and 7. For patient 2, physicians and nurses shared 5 out of 97 terms (5.2%). Physician term overlap in nursing care plans and recording transcripts was respectively: 2, 4.

Conclusions

Although each document analyzed was about the same patient, both patients 1 and 2 had less than 10% of terms shared between doctors and nurses in their own respective documentation. These results indicate there is considerable variation in the terms used by nurses and doctors regarding the same patient. 13.46% and 8.97% of terms extracted from patient 1 and 2 respectively were only documented during nursing handoffs. It is important that this issue be addressed alongside the notion that providing patients with personalized discharge summaries will lead to greater long-term benefits between a patient and their health.

References

1. Roussi K, Soussa V, Lopez KD, Balasubramanian A, Keenan GM, Burton M, Bahroos N, DiEugenio B, Boyd AD. Are we talking about the same patient?. Studies in health technology and informatics. 2015;216:1059.
2. Boyd AD, Lopez KD, Lugaresi C, Macieira T, Sousa V, Acharya S, Balasubramanian A, Roussi K, Keenan GM, Lussier YA, Burton M. Physician nurse care: A new use of UMLS to measure professional contribution: Are we talking about the same patient a new graph matching algorithm?. International journal of medical informatics. 2018 May 1;113:63-71.
3. Woods, S. S., Schwartz, E., Tuepker, A., Press, N. A., Nazi, K. M., Turvey, C. L., & Nichol, W. P. (2013). Patient Experiences With Full Electronic Access to Health Records and

Automated Category Alignment Applied to Different De-identification Annotation Schemata

Youngjun Kim, PhD[1], Paul M. Heider, PhD[1], Stéphane M. Meystre, MD, PhD[1]
[1]Medical University of South Carolina, Charleston, South Carolina, USA

Introduction: The annotation of electronic health record clinical notes is laborious. When applying natural language processing (NLP) to clinical notes, annotated text availability is limited by confidentiality protection requirements. Even when annotated text is available, annotation schemata (i.e., annotation categories and their organization) are mostly unique, making significant manual customization efforts a requirement. Successful annotation adaptation can enable effective reuse of existing text annotations. As a use case for this adaption problem, we focus on text de-identification, which involves detecting and hiding personally identifiable information (PII). This study aims to improve generalization across different annotation schemata with automated PII category alignment. We present a preliminary assessment of category alignment across four publicly available data sets (2006 i2b2[1], 2014 i2b2[2], 2016 N-GRID[3] shared tasks, and PhysioNet[4]).

Methods: We created a statistical model to optimally map pairs of PII categories defined for two different corpora. First, we trained a Bi-LSTM (bidirectional long short-term memory network)-based sequence labeling model with source training data. The Bi-LSTM model predicted the label of each word token. We collected all the findings of the Bi-LSTM model from the target training data and aligned them with the reference PII concepts. We then calculated the probability of each source PII category mapping to each of the target PII categories. If a source PII category had the highest probability of being aligned with one of the target categories, we assigned the source category to that category of target data. For example, the phrases classified as '*City*' by the 2014 i2b2 model were most frequently mapped to '*Location*' type phrases in the 2006 i2b2 test set.

Table 1. Category alignment with the 2014 i2b2.

2014 i2b2	2016 N-GRID	2006 i2b2	PhysioNet
Age	Age	Age	Age
Date	Date	Date	Date
Profession	Profession		
Patient	Patient	Patient	Relative
Doctor	Doctor	Doctor	Doctor
User name	User name	ID	
Hospital	Hospital	Hospital	Location
Organization	Organization	Doctor	Location
Street	Street	Location	Location
City	City	Location	Location
State	State	Location	Location
Country	Country	Location	
Zip	Zip	Location	
Location Other	Organization		Patient
Phone	Phone	ID	Phone
Fax	Fax	Phone	
Email	Email		
URL	Profession	Hospital	Location
Medical record	Phone	ID	
Biometric ID		ID	
ID number	License number	ID	Other

Results: Table 2 shows the performance of the Bi-LSTM model trained with the 2014 i2b2 training data, before and after PII category conversion. Precision (P), recall (R), and F_1-scores (F_1) were calculated with token matching where each PII term was evaluated on a per-token basis. We used paired t-tests to measure statistical significance. There was almost no difference in the 2016 N-GRID data due to the category definition shared with the 2014 i2b2. However, applying category conversion significantly improved the performance on the 2006 i2b2 and PhysioNet data (with P values of $<.00001$).

Table 2. Results of the 2014 i2b2 Bi-LSTM model.

	Before conversion			After conversion		
Test set	P	R	F_1	P	R	F_1
2014 i2b2	97.9	97.2	97.6	97.9	97.2	97.6
2016 N-GRID	86.8	84.3	85.5	86.8	84.3	85.6
2006 i2b2	78.4	72.6	75.4	90.3	83.3	86.7
PhysioNet	30.2	41.3	34.9	53.4	71.6	61.1

Conclusion: This study showed that our category alignment method could provide an efficient and convenient solution for combining different text collections with heterogeneous semantic annotations. Our future research involves integrating predictions from multiple de-identification models for more accurate alignment.

References

1. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. JAMIA 2007;14(5):550–63.
2. Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task track 1. J. Biomed. Inform. 2015;58:S11–S19.
3. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks track 1. J. Biomed. Inform. 2017;75:S4–S18.
4. Douglass M, Clifford GD, Reisner A, Moody GB, Mark RG. Computer-assisted de-identification of free text in the MIMIC II database. Computers in Cardiology, 2004: IEEE, 2004:341–44.

Using Natural Language Processing to Predict ICU Transfer in Hospitalized COVID-19 Patients

Phillip Kinney[1], Amara Tariq, PhD[2], Hari Trivedi, MD[3], Judy Gichoya, MD, MS[4], Imon Banerjee, PhD[5]

[1]Georgia Institute of Technology, Atlanta, GA; [2]Emory University School of Medicine, Atlanta, GA; [3]Emory University School of Medicine, Atlanta, GA; [4]Emory Winship Cancer Institute, Druid Hills, GA; [5]Emory School of Medicine, Atlanta, GA

Introduction

Many hospital systems, including Emory, have experienced a resurgence in COVID-19 cases since states have re-opened, with potential to overwhelm the healthcare systems due to insufficient resources. Today, hospitals are facing difficult decisions on how to allocate ICU beds and ventilators, and how to reorganize elective patients to optimize and ensure continued care for all patients. To mitigate the burden on the healthcare system, while also providing the best possible care for COVID-19 patients, efficient diagnosis and prognosis of the disease is needed. While AI can provide prediction information, current models[1] use only the structured components of Electronic Medical Records (EMRs) focusing on a set of pre-selected clinical information and omitting the comprehensive clinical history from the modeling framework. We developed an AI tool that predicts ICU transfer for COVID-19 patients 72 hours in advance using only *unstructured* free-text medical notes such as history and physical, progress notes, discharge summaries, and nursing notes.

Methods

With the approval of Emory IRB, we collected 1,688 of COVID-19 confirmed patients' data (55.1% female, 19.4% White, 48.6% African American) where a confirmed COVID-19 diagnosis was defined as either a positive SARS-CoV-2 RNA detection test or a diagnosis code for COVID-19. Each patient was labeled as Self-Isolation, Hospitalized or ICU based on their clinical timeline. Self-Isolation patients were sparse in number, so the Self-Isolation and Hospitalized classes were joined into a single Non-ICU class. We retrieved the free-text clinical notes for each patient from their COVID encounter up to a year before diagnosis, with a mean of 1,576 ± 2029 notes per patient. Notes within 3 days of the outcome and re-edited ED notes were removed to prevent prospective data leaks from improving model performance. Ultimately 381 unique patients remained for training and validation, comprising 221 Hospitalized, 17 Self-Isolation, and 143 ICU patients. The notes were embedded, word-by-word, into 300-dimension GloVe vectors[2] that had been pre trained on a corpus of Wikipedia articles; any words not available in the embedding map were replaced by zero-vectors of the same dimensionality. An average was taken across every embedded word from every note such that a single vector encompassed information from the patient's entire note record. These vectors were used as inputs into a single-layer neural network (NN) classifier model with RELU activation function on the hidden layer and Softmax activation function on the output layer. The NN model was trained and its predictive performance validated using 10-fold cross-validation.

Results

Across 10-fold cross-validation, the model resulted: 90.15% sensitivity (95% CI: 85.98% - 94.32%); 83.57% specificity (95% CI: 78.28% - 88.86%); 83.91% positive predictive value (95% CI: 77.02% - 90.80%); 87.41% accuracy (95% CI: 84.79% - 90.03%); and 86.86% area under the receiver operating characteristics curve (95% CI: 84.28% - 89.44%) at predicting ICU admission.

Conclusion

Our work demonstrates the feasibility of using AI based natural-language processing models to identify patients at risk of severe deterioration 72 hours in advance. Since the model only uses clinical notes, the tool is easily transferable to a different setting and could help medical staff prioritize medical attention to patients most at-need.

References

1. Wynants L, Van Calster B, Bonten MM, Collins GS, Debray TP, De Vos M, Haller MC, Heinze G, Moons KG, Riley RD, Schuit E. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. bmj. 2020 Apr 7;369.
2. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 Oct (pp. 1532-1543).

Data-Specific Training for Detecting Reports of Medication Intake on Twitter

Ari Z. Klein, MA, PhD[1], Graciela Gonzalez-Hernandez, MS, PhD[1]
[1]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Introduction

Scaling Twitter data for pharmacoepidemiology[1] requires automatically detecting users reporting that they have actually *taken* a medication mentioned in their tweet. The Social Media Mining for Health Applications (#SMM4H) shared tasks have addressed this challenge; however, the fact that the classifiers' performance generally declined between 2017[2] and 2018[3] has led us to examine the assumption that models for this task would generalize to all medications. Building on our preliminary analysis[4], this study assesses whether performance improves by training classifiers on tweets that mention the same types of medication for which the classifiers would be deployed for observational studies.

Methods

The #SMM4H corpora for this task comprise tweets that mention medications and are manually annotated as *intake*, *possible intake*, or *no intake*[5]. Incidentally, the types of medication in the tweets were similar across the training and tests sets in 2017, but largely different in 2018. To experiment with medication-specific training, we split the 5000 tweets in the 2018 test set into 80% (4000 tweets) and 20% (1000 tweets) sets to train and evaluate SVM and BERT classifiers. Then, we used these 1000 test tweets to evaluate the classifiers trained on the 2018 training set (17,773 tweets). Finally, we added the 2018 training set (17,773 tweets) to the 4000 tweets. The classifiers were evaluated based on the micro-averaged F_1-score for the "intake" and "possible intake" classes.

Results

Based on a training/test split of the 2018 test set, SVM and BERT classifiers performed better using a much smaller training set (4000 tweets) than using the 2018 training set, with the micro-averaged F_1-score of the SVM classifier improving from 0.41 to 0.55, and the BERT classifier improving from 0.56 to 0.60. When we added the 2018 training set to the 4000 tweets, the performance of the SVM classifier (0.51) was actually lower. The performance of the BERT classifier (0.64) did improve with the additional 4000 tweets. Based on a basic post-classification feature analysis, n-grams that are highly informative for distinguishing the three classes are different for the 2018 training and test sets.

Conclusion

Models for detecting medication intake on Twitter do not necessarily generalize to all medications. A BERT classifier achieved the best performance not merely with larger training data, but when the training data represented nuances in how users express whether or not they have taken specific medications. This study can advance the use of Twitter data for observing medication exposure among populations for whom traditional sources of data are limited.

References

1. Golder S, Chiuve S, Weissenbacher D, Klein A, O'Connor K, Bland M, Malin M, Bhattacharya M, Scarazinni LJ, Gonzalez-Hernandez G. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. Drug Saf. 2019;42(3):389-400.
2. Sarker A, Belousov M, Friedrichs J, Hakala K, Kiritchenko S, Mehryary F, Han S, Tran T, Rios A, Kavuluru R, de Bruijn B, Ginter F, Mahata D, Mohammad SM, Nenadic G, Gonzalez-Hernandez G. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (#SMM4H)-2017 shared task. J Am Med Inform Assoc. 2018;25(10):1274-1283.
3. Weissenbacher D, Sarker A, Paul M, Gonzalez-Hernandez G. Overview of the third Social Media Mining for Health (#SMM4H) shared tasks at EMNLP 2018. In: Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task; 2018 Oct 4; Brussels, Belgium. Association for Computational Linguistics; 2018. p. 13-16.
4. Klein AZ, Sarker A, O'Connor K, Gonzalez-Hernandez G. An analysis of a Twitter corpus for training a medication intake classifier. AMIA Jt Summit Transl Sci Proc. 2019:102-106.
5. Klein AZ, Sarker A, Rouhizadeh M, O'Connor K, Gonzalez G. Detecting personal medication intake in Twitter: an annotated corpus and baseline classification system. In: Proceedings of the BioNLP 2017 Workshop; 2017 Aug 4; Vancouver, Canada. Association for Computational Linguistics; 2017. p. 136-142.

Maturity in Enterprise Data Warehouses for Research Operations: Initial Model and Pilot Results

Boyd Knosp, MS[1], Thomas R. Campion, Jr., PhD[2]

[1]Institute for Clinical & Translational Science, University of Iowa, Iowa City, IA; [2]Clinical & Translational Science Center, Weill Cornell Medicine, New York, NY

Introduction

Enterprise data warehouses for research (EDW4R) have become a required part of a robust translational research enterprise [1]. All hubs of the Clinical Translational Science Award (CTSA) funded by the National Institutes of Health (NIH) National Center for Advancing Translational Science (NCATS) have adopted some form of an EDW4R that curates and delivers electronic patient data to inform clinical studies. While there is broad adoption of EDW4Rs [2], little is known about how individual CTSA hubs have implemented EDW4R activities, limiting the effectiveness of informatics for biomedical research within the CTSA community. Maturity models applied to research IT have "shown the potential to inform a range of leadership stakeholders within institutions in a number of ways, including…defining organizational best practices for research IT support within academic medicine" [3]. Building on our prior work exploring EDW4R operational practices at CTSA hubs [4], we have created a maturity model for EDW4R operations that may be used to assess and guide an institution's roadmap for optimizing EDW4R operations.

Methods

We conducted 34 semi-structured interviews with informatics leaders responsible for EDW4R activities at CTSA hubs. Topics included data governance, service management, workforce, relationship with enterprise IT, research access and outreach, and EDW4R metrics. Based on analysis of interview data, we created a maturity index[4,5,6] for EDW4R operations.

Results

The maturity index for EDW4R is shown in figure 1. There are six categories for EDW4R operations and each category has a list of maturity anchor statements [5] Each statement that is true for an institution results in a higher level of maturity. To determine the level of maturity in each category, institutions provide a Likert scale assessment for each survey. An overall EDW4R operational maturity score is calculated as an average across all categories.

Maturity Index for Enterprise Data Warehouse for Research Operations – each bullet is a "maturity anchor statement"

Data governance • Our governance structure considers both clinical and research data requests • We have a high-level committee that reviews external agreements regarding accessing data from the EDW4R. • We have a team that reviews and prioritizes data requests to the EDW4R. • We have a team that defines what data goes into the EDW4R • We have a team that engages with the IRB, compliance and legal to define policies regarding requests for using data from the EDW4R. • We manage requests for access to our EDW4R with guidance from our IRB.	**Workforce** • One or more positions on our EDW4R team are shared with enterprise IT. • We have identified training programs as pipelines to fill open positions on our EDW4R team. • We have project managers on our EDW4R team. • Our EDW4R team includes faculty domain experts who assist with EDW4R services. • We have one or more staff whose duties include processing requests for data from our EDW4R. • We have one or more staff whose duties include aggregating and managing the data stored in our EDW4R.
Service management • Our EDW4R services are listed as part of our enterprise IT services. • Our IT helpdesk knows refer research requests for clinical data to our EDW4R team. • We have a standard format for submitting data requests. • We have a written description of the services available to access our EDW4R.	**EDW4R relationship to Enterprise IT** • The EDW4R is part of the overall EIT Strategic planning process. • Our CRIO collaborates closely with other C-suite leaders. • Our EDW4r data team is integrated into our Enterprise IT organization. • EDW4R services are listed as part of th Enterprise IT service catalog. • Our EDW4R group works closely with Clinical data warehouse teams. • Enterprise IT teams know to refer research data requests to our EDW4R team.
EDW4R access & outreach • We have researcher-available documentation on what data is available in our EDW4R. • We offer a service that establishes a population specific data mart that is periodically updated. • We have regular orientation courses in accessing data in the EDW4R. • We have a variety of methods for enabling users of different levels of data expertise to access data in our EDW4R. • We require CITI or other relevant training prior to providing accessing our EDW4R. • We have self-service tools for exploring a de-identified copy of the EDW4R. • We provide regular trainings for EDW4R services. • We are able to generate reports from our Electronic Health Record for research requests.	**EDW4R metrics** • We provide data quality assessments for a research network such as PCORnet or ODHSI. • We provide information to describe our EDW4R based on the NCATS common metrics. • We also track metrics that are used for strategic planning. • We track response times for research requests for patient data. • We track the outcomes (publications, grants…) resulting from research requests for patient data. • We keep track of the number of research requests we received for patient data.

Figure 1. EDW4R operations maturity index

Discussion

This model, based on broad community input, establishes a set of best practices for EDW4R operations and provides a metric that might be used by individual institutions or across a consortium of institutions to understand maturity.

References

1. Obeid JS, Tarczy-Hornoch P, Harris PA, Barnett WK, Anderson NR, Embi PJ, et al. Sustainability considerations for clinical and translational research informatics infrastructure. J Clin Transl Sci. 2018 Oct;2(5):267–275.
2. Obeid JS, Beskow LM, Rape M, Gouripeddi R, Black RA, Cimino JJ, et al. A survey of practices for the use of electronic health records to support research recruitment. J Clin Transl Sci. 2017 Aug;1(4):246–252.
3. Campion TR Jr, Craven CK, Dorr DA, Knosp BM. Understanding enterprise data warehouses to support clinical and translational research. J Am Med Inform Assoc. 2020 Jul 17;ocaa089. doi: 10.1093/jamia/ocaa089.
4. Knosp B, Barnett W, Anderson N, Embi, P. Research IT maturity models for academic health centers: Early development and initial evaluation. J Clin Transl Sci, 2(5), 289-294.
5. Fraser P, Moultrie J, Gregory M. The Use of Maturity Models/Grids as a Tool in Assessing Product Development Capability. Vol. 1, 2002. doi:10.1109/IEMC.2002.1038431.
6. Grajek, Susan. "The Digitization of Higher Education: Charting the Course." *Educause Review* (December 12, 2016 2016

Leveraging Electronic Health Records Data for Predicting Alzheimer's Disease Progression

Sayantan Kumar; Inez Oh, PhD; Aditi Gupta, PhD; Albert M. Lai, PhD; Philip R.O. Payne, PhD
Institute for Informatics, Washington University School of Medicine, St. Louis, MO

Introduction

Alzheimer's disease (AD) is the most common form of dementia, a set of progressive neurodegenerative disorders associated with progressive memory loss, cognitive impairment, and general disability. AD-related brain pathology, which includes the accumulation and deposition of amyloid-β peptide and tau protein, begins almost 10-20 years before the onset of dementia symptoms. Identifying individuals with early AD brain pathological changes can lead to preventive therapeutic interventions to delay disease progression. Current gold-standard diagnosis of AD requires expensive and/or invasive procedures such as neuroimaging or sampling of cerebrospinal fluid for biomarker testing. In this research, we investigated the use of retrospective analysis of electronic health records (EHR) data, collected routinely during outpatient visits, as a low-cost, non-invasive method to predict whether patients in mild AD stage will progress to moderate/severe AD by their next outpatient visit.

Methods

Outpatient clinical data between June 1st 2013 and May 31st 2018 were extracted from the EHR of Barnes-Jewish Hospital, a large tertiary-referral academic medical center in St. Louis, MO. Longitudinal data from 1595 successive visit-pairs from 900 patients were eligible for inclusion, where each visit-pair had mini-mental state exam (MMSE) recorded at both visits and a mild AD diagnosis at the initial visit. All visit-pairs for which MMSE ratings declined from >=20 at the initial visit to < 20 at the successive visit were categorized as the moderate/severe AD group (n = 486, mean age = 78 +/- 8.5 years, male = 58 %). All other visit-pairs (MMSE >=20 at both visits) were categorized as the mild AD group (n = 1109, mean age = 76 +/- 8 years, male = 40%). Patient information included 35 clinical features with demographics (age, gender, and race), vital signs, cognitive test scores, medication types and comorbidities (ICD10-CM codes). Missing data values were filled by mean, median and mode imputation for continuous, ordinal and nominal variables respectively. XGBoost, a scalable and interpretable decision tree based ensemble model was used and its performance was evaluated using 5-fold cross validation. To gain insight into model interpretability, the TreeSHAP summary plot was calculated showing both the feature importance (weights) and their impact on the model predictions.

Results

Figure 1 represents the Area Under the Receiver Operating Characteristics curve (AUROC) with a mean AUROC value of 0.804 along with the top features used in predictive modeling. The F1 scores of the mild AD and moderate/severe AD group are 0.66 and 0.80 respectively. The SHAP summary plot shows that cognitive test scores have the maximum importance with high positive SHAP score indicating greater impact of the feature on the model positively predicting moderate/severe AD in the next visit.

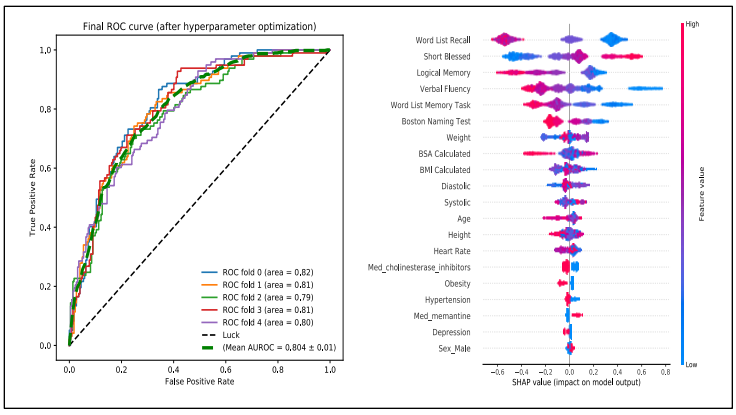


Figure 1: AUROC performance (left) over all the 5 folds and SHAP Summary plot (right). In the SHAP plot, each dot in a row represents a patient's visit-pair. The positivity (x-axis) of the SHAP score indicates the feature is more important for predicting AD. The y-axis represents feature importance or weight. Color of the dot represents value of the feature for that visit.

Conclusion

Our research examines the feasibility of predicting the progression of AD using retrospective analysis of readily available EHR data and also provides clinicians with insight into potential risk factors for the disease, with the ultimate goal of providing personalized disease treatment for each patient in a minimally-invasive and cost-effective manner. Possible reasons for the low AUROC and F1 score values of the AD group include the imbalanced distribution of the moderate/severe AD and mild AD classes and the missing feature values in a significant number of visits. Future analysis will include adding features like biomarkers and genetic factors, performance comparison with other with supervised classifiers, and extending the proposed approach to apply it to other disease progression modeling.

References

1. Ferri CP, Prince M, Brayne C, et al. Global prevalence of dementia: a Delphi consensus study. The lancet 2005;**366**(9503):2112-17

Evaluation of Multiple Myeloma Clinical Decision Support Tool Value When Populated with Community Health System Data

Steven E. Labkoff, MD, FACP, FACMI, FAMIA[1], Kathy Giusti, MBA[1], Paul Giusti, MBA[1], Ryan Wilcox, MD[2], Derrick Haslem, MD[2], Daanish Hoda, MD[2], Kerry Rowe, PhD[2], Gail Fulde[2], Jesse Gygi, MHA, MPH[2], Brad Bott, MBA[2], Ben Smith[3]; [1]The Multiple Myeloma Research Foundation, Norwalk, CT, [2]Intermountain Healthcare, Salt Lake City, UT, [3]IQVIA, Plymouth Meeting, PA

Introduction

Clinical decision support (CDS) technology has the potential to improve health outcomes by offering physicians an informational resource to support review and application of best pratices.[1] The Multiple Myeloma Research Foundation (MMRF) and Intermountain Healthcare (IMH) conducted a study to assess the suitability of a single health system's data for a myeloma-specific CDS tool that visualizes treatment pathways, and to assess the effort needed to support a CDS program.[2] This research is part of a longer-term effort to explore how CDS technology can help:

- increase awareness of and apply treatment guidelines by visualizing pathways for specific MM patient cohorts
- improve understanding of treatment variation within health systems
- improve outcomes research by showing relationships between treatments and outcomes

Methods

IA12 data from the CoMMpass study[3] was used to create a CDS tool prototype. These data were aggregated into state and transition maps to identify nodes and pathways with corresponding outcomes, including response, progression-free survival (PFS), and overall survival (OS). Intervening patient states were displayed using Sankey diagrams.

We also tested if EMR data from a community health system (i.e., IMH) could support such visualization. The team designed a study protocol and obtained IRB approval. Inclusion criteria included patients with active MM between January 2016–June 2018; adult aged 18 years to 89 years at diagnosis of active or smoldering MM. An IMH-specific data dictionary was assessed for variable importance, quantity, and ease of acquisition. IMH then manually abstracted prioritized structured (eg: labs) and non-structured (eg: notes) data for use in the tool.

Results

Ninety-six of an initial 146 patients meeting eligibility criteria had sufficient data usable for the study, reflecting 44 unique drug combinations across 9 lines of therapy. The tool was able to associate and visualize all patients and their clinical states and transitions to their outcomes. Clinical data was typically complete (99% of the time), including key clinician-derived data, such as ECOG scores (78%) and treatment response (99%). 569 person-hours were required to conduct the abstraction activity on 96 cases, averaging 5.9 hours/patient.

Discussion

The IMH portion of the study supports the hypothesis that a community health system can provide sufficient high-quality information to power a CDS tool with priority features. Only 65% (96/146) of the initial study group had usable data because some patients had received partial care outside of the IMH integrated delivery network (IDN) leaving associated data inaccessible. Initial biostatistical analysis suggests that roughly 750-1000 complete patient records would be required for statistically significant outcomes research with granularly stratified cohorts.

The MMRF is currently recruiting 5-7 additional large IDNs to obtain the patients to power more generalizable functionality.

References

[1] McKie PM, Kor DJ, Cook DA, Kessler ME, Carter RE, Wilson PM, et al. Computerized advisory decision support for cardiovascular diseases in primary care: a cluster randomized trial. Am J Med [Internet]. 2019 Dec 18 [cited 2020 Mar 5]. Available from: https://doi.org/10.1016/j.amjmed.2019.10.039

[2] Garcelon N, Burgun A, Salomon R, Neuraz A. Electronic health records for the diagnosis of rare diseases. Kidney Int [Internet]. 2020 Jan 14 [cited 2020 Mar 5]. Available from: https://doi.org/10.1016/ j.kint.2019.11.037

[3] Christofferson A, Nasser S, Aldrich J, Penaherrera D, Legendre C, Benard B, et al. Integrative analysis of the genomic landscape underlying multiple myeloma at diagnosis: an Mmrf Commpass analysis. Blood. 2017 Dec 7; 130 (Supplement 1): 326.

The Informatics, Business and Logistic Challenges of Launching an Integrated Direct-to-Patient Registry During a Pandemic: The MMRF CureCloud Experience

Steven E. Labkoff, MD, FACP, FACMI, FAMIA[1], Michele Likens[1], Shaadi Mehr, PhD[1], Michael Andreni, Sergey Miron[2], Ben Lawlor[2], Karen Dietz, JD[1], Daniel Auclair, PhD[1], Hearn J Cho, MD, PhD[1], Leon Rozenblit, JD, PhD[2]
[1]The Multiple Myeloma Research Foundation, Norwalk, CT, [2]Prometheus Research, an IQVIA company, New Haven, CT

Abstract:
Multiple Myeloma is a hematologic cancer of plasma cells producing nonsense proteins causing end-organ damage. Launched in July 2020 to advance translational research and precision medicine, the Multiple Myeloma Research Foundation launched the CureCloud Direct-to-Patient Registry (CC-DTC). This 5000 patient registry will aggregate 8 disparate data sets. We describe an array of major challenges to implementation and operations in multiple domains including complications due to the COVID-19 pandemic. We report on the challenges encountered, as well as solutions during the implementation and launch of the CureCloud.

Methods:
In late 2017 the MMRF embarked on a follow-up plan on its 2011 CoMMpass[1] program for patients with newly diagnosed multiple myeloma. Between 2017-2020, the CC-DTP was conceived, architected, constructed, and launched. With an ambitious goal to create a longitudinal registry by aggregating data from 5000+ patients in 8 different data types, the MMRF team began to create this longitudinal, 10-year journey focusing-in on patients with multiple myeloma and smoldering myeloma. Using techniques from other informatics registry efforts[2,3,4] the CC-DTP began its journey. The program engaged an array of contractors to sort out the mechanics and informatics challenges associated with such an undertaking. As it took shape, the management team began to realize that unlike a typical registry that generally aggregates 1-2 data types, the work- and dataflow challenges proved to be far more complex than anticipated. After IRB approval, a pilot engagement platform was constructed and tested, the team realized that a far more comprehensive system (a rewrite) was needed to deal with many of the challenges met in the pilot. Challenges encountered were in a variety of domains including, but not limited to: Informatics, medico-legal, regulatory, social, and the complexities that arose as a result of COVID-19.

Discussion
A project of this complexity presents a myriad of privacy and confidentiality challenges - specifically around data in transit and data at rest. A comprehensive privacy and security review workstream was initiated and executed including end-to-end encryption for all data in the registry. Fully 22% of hospitals and clinics refused our e-Signatures and refused to deliver the requested data. There were multiple excuses including non-certification of the eSignature, refusal to provide for research, and simply refusing any requests not on their letterhead. We are moved to a DocuSign certified eSignature solution. Practicing medicine across state lines became a complex legal challenge when the decision was taken to return data to patients. The NGS assay had to be CLIA-certified, and a licensed clinician (in the same state as the patient) had to order the assay - and receive the results before the patient. The CC-DTP registry ecosystem is complex. It initially engaged 14 different commercial and academic partners, later winnowed to eight. All but two of the remaining vendors were private-sector organizations, with MMRF informatics team taking on program and vendor management functions, and the Prometheus team taking on the role of software developers and general informatics advisors.

COVID-19 presented an array of unforeseen challenges. The research lab we partnered with (The Broad Institute, Cambridge, MA) had to stop all CureCloud work in March as they converted to 100% focus on COVID. Our phlebotomy partner (EMSI) went out of business due a pandemic-related business failure 3 days prior to launch. A new phlebotomy company was found in under 5 weeks and integrated into the system in an additional 30 days.

Despite the enormous array of challenges from multiple domains, the CC-DTP opened its doors on July 14, 2020. In the first month of operations, 849 patients came to the site to attempt to register, 424 screened-in, and 181 completed the enrollment screener, contributed their data to the program, and engaged their clinicians to order the 70-gene NGS panel, and are fully enrolled in the registry. Despite the encountered challenges, the success of the registry appears to be on a solid trajectory.

Perioperative Workflow Management System and Improvement in First Case On-Time Start Rates in the U.S. Department of Veterans Affairs

David LaBorde, M.D., M.B.A.[1], Amy Green, B.S.N.[1]
[1] Document Storage Systems, Inc., Juno Beach, FL, USA

Abstract

While the operating theater can generate significant revenue for healthcare provider organizations, building, staffing and operating them is costly. As such, making full use of available time is key. When the first case of the day starts late, this wastes a valuable resource. Herein we report the experience of a facility that increased their first case on-time start rates.

Introduction

Running operating theaters has a high fixed cost. It has been estimated that, excluding physician costs, the cost of operating room time ranges from $15 to $20 per minute and that at least half of this amount is fixed overhead costs.[1] Thus, maximizing the utilization of available operative time is fiscally responsible.[2] In addition, for facilities with operating theaters that run at or near capacity, maximizing operating theatre utilization also improves access to care for patients. First case on-time start (FCOS) rates are a process metric often tracked in operating theaters given the impact late first cases can have on cases scheduled to follow in the same operating room later the same day. Herein we report an analysis of FCOS rates before and after the implementation of a perio-operative workflow management system at a single, large, 1a (high complexity) U.S. Department of Veterans Affairs (VA) medical center.

Methods

A peri-operative workflow management improvement and analytics software system (*PeriOp Manager, LiveData Inc., Cambridge, MA*) was procured by the facility. The system was integrated into the system of record with an integration technology utilized enterprise wide by the VA (*Integration Framework, DSS Inc., Juno Beach, FL*). The system went live at the facility in May of 2018. Slightly less than two years after implementation and go-live, this small retrospective analysis was undertaken. FCOS rates collected as a part of routine health care operations were obtained for the historical period three months prior to system go live (this period included May 2018, the month system went live) and for a period three months after the month the system went live. A one-tailed paired t-test was used to compare the average baseline FCOS rates to the average FCOS rates after go-live. In the retrospective analysis, the null hypothesis was that there would be no difference in the average FCOS rate after system go live; the alternative hypothesis was that the difference would be greater than zero. Type I error was permitted to be no greater than 0.05.

Results

During the period analyzed, 1,111 weekday cases were completed (559 Pre-Go-Live or PrGL, 552 Post-Go-Live or PoGL). Of these 1,111 cases, 409 were first cases (208 PrGL, 201 PoGL). On average 68% of first cases started on time, 72% of cases started less than 15 minutes late (72% PrGL, 70% PoGL), 16% of cases started 16 to 30 minutes late (16% PrGL, 16% PoGL) and 12% of cases started > 30 minutes late (11% PrGL, 14% PoGL). A total of 35.3 hours were lost to first case late starts (20.3 PrGL, 15.0 PoGL). The baseline monthly FCOS rates for March, April and May 2018 were 49%, 55%, 69%, respectively; post-go-live FCOS rates for June, July and August 2018 were 80%, 71% and 83%, respectively. The average pre-implementation FCOS rate was 57.7%±10.3 (SD) as compared to the average post-implementation which was 78.0%±6.2 (SD); this difference was statistically significant ($p = 0.03$) and the null hypothesis was refuted.

Table 1: t-Test: Paired Two Sample for Means

	*Pre-Go-Live**	*Post Go-Live*
Mean	57.7	78.0
Variance	105.3	39.0
Observations	3	3
Pearson Correlation	0.452	
Hypothesized Mean Difference	0	
df	2	
t Stat	-3.79	
P(T<=t) one-tail	0.03	

** Includes May, the month of go-live*

Discussion

In this single facility retrospective analysis, FCOS rate improved in the time period that temporally followed deployment of a software intervention targeting workflow improvement. Further work could aim to examine these observations and sustainment over a longer timeframe; a prospective analysis would also be of potential value.

References

1. Macario A. What does one minute of operating room time cost? Journal of Clinical Anesthesia (2010) 22, 233-236.
2. Chapman WC, Luo X, Doyle M, Khan A, Chapman WC, Kangrga I, Martin J, Wellen J. Time is money: can punctuality decrease operating room cost? J Am Coll Surg (2019) 230(2):182–189.

CoviDash-SM: A public COVID19 dashboard for social media data-based research and surveillance

Sahithi Lakamana, MS[1], Mohamed Al-Garadi, PhD[1], Yuan-Chi Yang, PhD[1], Abeed Sarker, PhD[1]

[1]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA;

Introduction

The COVID19 pandemic is one of the worst on record, and due to the enormity of the crisis, it is necessary for us to utilize available resources to monitor and control the spread of the coronavirus. There are many efforts to use traditional sources of information (*e.g.*, emergency department data), but new-age sources such as social media (SM) are largely overlooked. SM is of particularly high-utility at this time as it has become the primary mode of communication for many people since the 'lockdown' guidelines went into effect. Curating knowledge from this noisy source is, however, difficult due to the complex nature of text and the computational infrastructure required to process the massive volume of data. As part of our COVID19 response at Emory University, we are developing a dashboard that visualizes various real time statistics derived from SM chatter via natural language processing (NLP) and machine learning methods. Our objectives for this poster are to illustrate how various types of data can be accessed through the dashboard, and to receive feedback about possible future extensions.

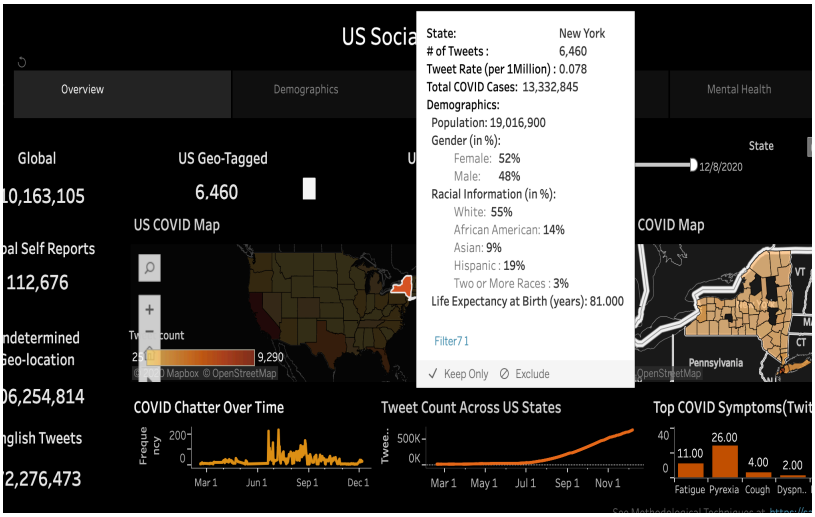


Fig 1. Sample tab of the COVID19-social media dashboard showing state-level information. Full dashboard: https://sarkerlab.org/covid_sm_data_bundle/

Methods

Data for the dashboard is drawn from Twitter and Reddit starting from January 1, 2020. Data is being continuously collected since May 8—through the COVID19 firehose stream for Twitter; and the PRAW API for Reddit—and earlier data were obtained from publicly available sources.[1,2] For Twitter, over 500 COVID19-related keywords are used for data collection (see cited research), and further information is mined from users automatically classified to be COVID19 positive. Several NLP modules are employed on the streaming data to extract symptoms and compute their distributions, perform sentiment and topic analysis, track approved an unapproved treatment, and perform topic analysis. For Reddit, data is collected from the /r/Coronavirus subreddit, and COVID19 users are identified by their self-reported *flair.* Following collection, similar NLP methods compute statistics from the data for visualization on the dashboard. Two dashboards—(i) Global and (ii) United States are available for viewing.

Results and Discussion

The dashboard was created using the Tableau software package and is available at [URL]. Between 8th May (when our streaming data collection commenced) and 8th December, our classification approach discovered 112,676 self-reports. Countries with the highest volumes of chatter are the United States, Brazil, Japan, India, and the United Kingdom. 893 COVID19-positive users were detected from Reddit. The dashboard visualizes geolocation-based (state-level in the US; nationally otherwise) self-reports of COVID19 diagnoses and symptoms, mental health conditions, topic-specific sentiments and emotions, and frequently discussed possible treatments for COVID19 (including unapproved and fraudulent substances). In the near future, we will integrate contact tracing and substance use information to the dashboard. We will also incorporate feedback from the AMIA Informatics Summit participants.

References

1. Tekumalla R, Banda JM. *A Large-Scale Twitter Dataset for Drug Safety Applications Mined from Publicly Existing Resources*.; 2020. doi:arXiv:2004.03688.
2. Chen E, Lerman K, Ferrara E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Heal Surveill*. 2020;6(2):e19273. doi:10.2196/19273

Research Recruitment with the Patient Portal and Clinical Data Warehouse

Adam M. Lee, MBA[1], Stephanie J. Deen[2]
1 University of North Carolina, Chapel Hill, NC 2 UNC Health, Chapel Hill, NC

Introduction

The North Carolina Translational and Clinical Sciences (NC TraCS) Institute and UNC Health established a process framework for patient recruitment utilizing the Epic MyChart™ patient portal within UNC Health's Epic Electronic Health Record (EHR) implementation. This framework is open to researchers at the University of North Carolina - Chapel Hill offering three aims 1) provide an additional recruitment method and channel for research engagement and recruitment. 2) Leverage computable phenotypes for mass identification of patients within the EHR. 3) Ensure regulatory compliance, system governance, and patient privacy encapsulates the process framework.

Methods

NC TraCS offers three services using the patient portal recruitment framework to researches looking to recruit using MyChart. All services are routed through IRB, School of Medicine, and UNC Health's and governance processes to ensure regulatory compliance and patient privacy. Additionally, all services must utilize a computable phenotype that identifies and extracts patients from the clinical data warehouse.

The first service, dubbed *Foundation,* utilizes foundational features and functions in Epic, such as the study management component, and leverages the CDW for controlled cohort identification. A study coordinator reviews the cohort, selects patients one-by-one, and sends recruitment requests. The EHR fully contains the whole recruitment process and updates records as patients accept or deny these requests using MyChart.

The second service is called *Message*. This service shifts from Foundations from being driven by a study coordinator to being processed by an honest broker. NC TraCS uses computable phenotypes to identify populations, and then a bulk message process is used to send the entire cohort population study-specific language. These messages are similar to an email message, but limited to the patient portal. While the recruitment messages may contain links to external websites or study contact information, these actions, such as clicking a link, are not tracked within the system.

The third service, *Integrate*, builds upon the Message functionality. Instead of untraceable third-party web links, study-specific URLs are embedded in the MyChart message and directs patients to REDCap, a research data capture tool. Researchers use the REDCap integration to capture and record study interest, then can link out to additional resources. This option allows complete cohort management and linkages between external study identifiers, the patient's medical record, and third-party study websites.

Results

Nine studies have utilized this framework and services, engaging 29,435 patients. Effectiveness as been shown to promising, one Principle Investigator reported 69 newly consented patients after the first week of using *Message*, this was of a batch of 4,116 messages.

Table 1. Patient Counts by Service

Service	**Patients**	**Studies**	**Patients per Study**
Foundation	1,961	4	27; 40; 49; 1845
Message	17,915	4	541; 4116; 4892; 8366
Integrate	9,559	1	9559

Discussion

The main discussion points revolve around consent and selection bias. UNC Health opts-in patients for research during the consent for the patient care process, with no global opt-out. Methods exist for patients to turn off *Foundation* messages. However, this is a communication preference and not an opt-out. The process framework does not recommend using patient portal recruitment as the sole recruitment method, as enrollment for UNC Health's MyChart is around 40% of patients and limits recruitment opportunities for those without patient portal access.

Conclusion

Overall, TraCS and UNC Health established a repeatable, sustainable process framework in which researchers can engage patients via a trusted and secure channel for research recruitment opportunities and enhancing the UNC research informatics pipeline.

Syndromic surveillance for COVID19 from Reddit using multi-platform lexicons

Abimbola Leslie, MPH[1] Sahithi Lakamana, MS,[2] Abeed Sarker, PhD[2,3]
[1]Laney Graduate School, Emory University, Atlanta, GA 30322
[2]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA 30322
[3]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30322

Introduction

In the United States, the National Syndromic Surveillance Program (NSSP)—a collaboration between the Centers for Disease Control and Prevention (CDC), state and regional health departments, and federal government, academic and private partners—conducts syndromic surveillance to obtain early warnings for public health concerns, such as an outbreak. The NSSP's COVID response focuses on emergency department (ED) data to identify potential outbreaks.[1] While ED-based surveillance data has been effective in the past, COVID19 presents challenges (*e.g.*, mildly symptomatic patients; lack of immediate confirmatory testing) that warrant the exploration of additional sources for syndromic surveillance. Social media, which has become a particularly important channel for communication during COVID19, has been shown to contain large amounts of COVID19-related chatter, which can be captured and analyzed in real-time. To effectively utilize social media data, it is crucial to develop and evaluate methods and resources that can be applied to data from distinct social networks. In this study, we utilized a COVID19 symptom lexicon from Twitter,[2] expanded using a small amount of Reddit data, to discover and compare, via natural language processing methods, automatically detected COVID19 symptoms reported publicly on Twitter and Reddit.

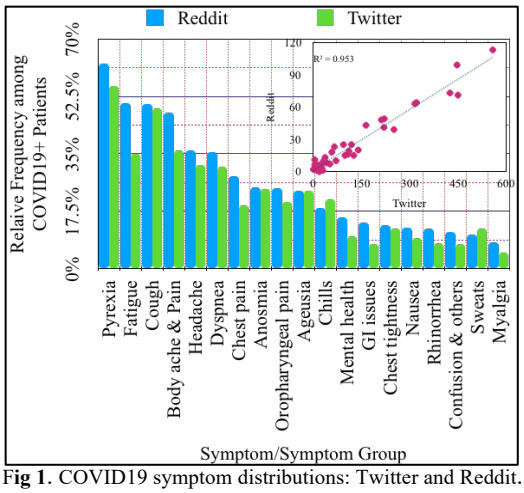


Fig 1. COVID19 symptom distributions: Twitter and Reddit.

Methods

We collected Reddit data from the */r/coronavirus* subreddit. Within the subreddit, posts can be marked with a *flair*, and the flair '*tested positive*' represents posts from users who have themselves tested positive. We randomly selected a small sample of these posts and manually annotated the COVID19-specific symptoms reported [62 posts; 1 annotator (AL), 1 validator (AS)]. We combined these symptoms (32 unique expressions; 277 symptoms; 27 negated symptoms) with those from a Twitter lexicon.[2] After the creation of the meta-lexicon, we attempted to use it to discover all reported symptoms by self-identified COVID19 positive users on Reddit (unlabeled users). We matched the lexicon entries with possible symptoms presented by the user using a lexical similarity function (Levenshtein ratio) and a semantic similarity function (Word2Vec vector similarities). Similarities above a predefined threshold are considered to be matches. Negations were detected using a customized version of NegEx and all symptoms occurring within the scope of a negation were removed. We defined the scope as: up to 3 terms following the negation, unless a period character ('.') or another negation occurs. The distribution of symptoms obtained were statistically compared to known symptom distributions from Twitter to assess the feasibility of Reddit for COVID19 syndromic surveillance.

Results and Discussion

The Reddit data consisted of 8,435 posts from which 893 COVID19-positive users were detected, based on the flair information they voluntarily provided. Symptom distributions between Twitter and Reddit had significant and high correlation (r^2=0.95; Figure 1). Twitter lexicon, without combining with Reddit, obtained F_1-score of 0.71, suggesting room for improvement by adding additional lexicon entries. Notable differences were *fatigue* and pain-related symptoms, with Reddit users consistently reporting pain at higher rates. Reddit users appear to consistently report higher numbers of symptoms compared to Twitter users, resulting in significantly higher reporting rates for individual symptoms (p=0.0076; two-tailed paired T test). However, there was no statistically significant difference in the mean number of symptoms reported per person (4.94 *vs.* 5.55; p=0.0503). While Twitter has been widely used for surveillance work, Reddit has been scarcely used. Unlike Twitter, Reddit does not provide geolocation information, but the discussions are considerably richer in information, and have the potential for use in syndromic surveillance and contact tracing. Our work shows that the symptom lexicon developed is largely portable across the networks, and that Reddit and Twitter data may complement each other. A multi-network syndromic surveillance approach over social media data has the potential of complementing and benefiting existing syndromic surveillance efforts.

References

1. National Syndromic Surveillance Program. *NSSP Supports the COVID-19 Response*. https://www.cdc.gov/nssp/covid-19-response.html. [Accessed: 08/25/2020]
2. Sarker A, Lakamana S, Hogg-Bremer B, Xie A, *et al. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource*. [online: 2020 Jul 4]. J Am Med Inform Assoc. 2020. doi:10.1093/jamia/ocaa116.

Measuring NICU Healthcare Workers' Time Spent on Collaborative Activities Through EHR Audit Logs

Patrick Li[1], BobChen[2], Wael Alfrifai, MD[2], Daniel France, Ph.D.[2], You Chen, Ph.D. [2]

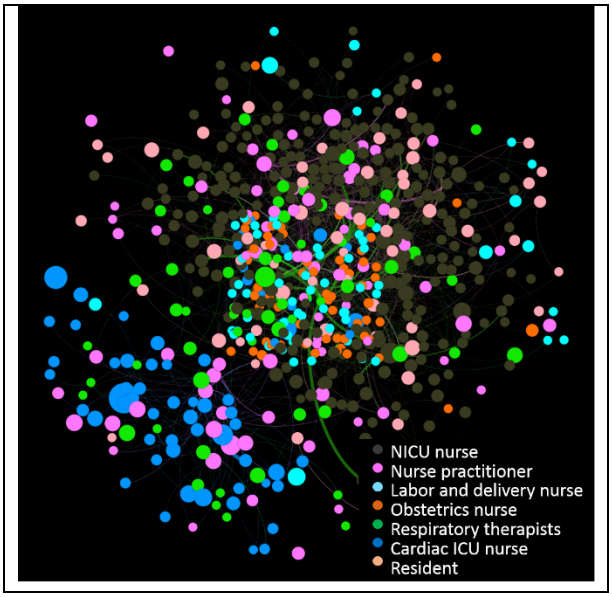
[1]University of Pennsylvania, Philadelphia, PA; [2]Vanderbilt University Medical Center, Nashville, TN

Description of the Problem

Healthcare workers (HWs) widely use electronic health record (EHR) systems as a collaborative platform to share health information in the neonatal intensive care unit (NICU)[1]. The time spent on the EHR activities, including those individual or collaborative ones, could potentially impact a HW's workload. However, few studies systematically examine activities in EHRs, and time spent on those activities. Thus it is hard to understand those activities, especially collaborative ones, and measure their relationships with clinician workload. The objective of this study is to develop an informatics framework to estimate times of collaborative activities in EHRs through the lens of audit logs.

Methods

We gathered 4 months of EHR audit log data from a large academic medical center NICU. The data include 2,840,249 actions performed by 3,303 HWs to the EHRs of 382 NICU patients. We defined an interval as a consecutive action sequence performed by a HW to EHRs of a patient; and a collaborative session as a set of overlapped intervals performed by different HWs to EHRs of the same patient. We define intervals that are part of a collaboration session as collaborative ones, and those which are not part of as an individual. Using audit log data, we measure time spent on each individual and collaborative interval, and who performed the intervals, to create an intermediary data set. Next, we use this data set to measure the proportion of collaborative intervals over all intervals and the proportion of EHR time a HW spends on the collaborative intervals, which we call collaboration intensity. The one-way analysis of variance is used to test the differences in the collaboration intensity between HWs with different specialties. Gephi is used to depict the collaborative network and collaborative intensity of each HW.



Discussion of Results

We identified 15,367 collaborative sessions for 2,361 HWs. The statistical test results show there exist differences in the collaborative intensity ($p<1.29e-11$) between the 13 investigated specialties (e.g., NICU nurses, respiratory therapists, diagnostic radiology technologists), which have the largest amount of time spent in EHRs. The left figure shows a subnetwork consisting of 662 HWs coming from the top 7 specialties, which have the largest number of HWs involved in the collaborative activities. The size of the node shows the strength of the collaborative intensity. NICU nurses are the most active in the EHR collaboration and they spread the whole network.

Conclusion

Leveraging audit logs to examine collaborative activities, time spent on those activities, and who works with whom to complete them, can assist healthcare organizations in understanding teamwork and their efficiency. Results learned from our developed informatics framework can assist HCOs in understanding collaboration occurring in EHRs, which can potentially optimize collaboration in the NICU.

References

1. Chen Y, Lehmann CU, Hatch LD, et.al. Modeling Care Team Structures in the Neonatal Intensive Care Unit through Network Analysis of EHR Audit Logs. Methods of information in medicine. 2019 Nov;58(4-05):109.

Applying deterministic and probabilistic record linkage approaches to quantify residential history in the electronic health record

Xuan Lin[1], Sharmistha Guha, PhD[1], Michael Valancius, MS[1,2], Kay Jowers, PhD[1], Laura Richman, PhD[1], Jerry Reiter, PhD[1], Christopher Timmins, PhD[1], Nrupen A. Bhavsar, PhD[1]
[1]Duke University, Durham, NC; [2]University of North Carolina, Chapel Hill, NC

Introduction:

Risk for chronic health conditions is often dependent on long term exposure to environmental pollutants. Residential context is one of the important loci of exposure but one that is not fixed. Therefore, residential history, defined as a person's current and past residential addresses, is necessary to quantify the impact of long-term pollution exposure on health. However, most electronic health record (EHR) systems only record a patient's current and former residential address when they have an encounter in the health system. This can result in incomplete information on residential history, potentially misclassifying exposure and introducing bias when we assess the impacts of environmental hazards on health. The aim of this study was to use deterministic and probabilistic linkage methods to develop a reproducible approach to link third party residential mobility data with EHR data.

Methods:

Data Sources: We used EHR data from the Duke University Health System (DUHS) to identify 3,181 patients who were diagnosed with pancreatic cancer from 2014-2019. Residential history data from 2006-2019 was obtained from InfoUSA which provides household level residential mobility data (including first/last name of three family members, residential address, family identifier [family ID]). The dataset includes approximately 200 million household units from the United States. This approach was further tested in 42,892 DUHS patients diagnosed with chronic obstructive pulmonary disease (COPD) in 2019. Data Linkage: We used deterministic as well as probabilistic linkage approaches to match patient ID from EHR data with family ID in the InfoUSA data. Deterministic linkage matched records in EHR data with InfoUSA data by shared keys including first name, last name, and address. Probabilistic linkage was implemented using R package 'fastLink'[1], which calculates the string similarities of linkage fields based on the Fellegi-Sunter method[2]. Pairs of records were regarded as probable linkages if string similarities were above a chosen threshold of 0.8, and manual review was used to identify false positives. Probabilistic linkage detected matches that were missed through deterministic linkage due to alternative spellings, spelling mistakes and spacing errors. These errors in data collection and missingness are argued to be inherent to some extent, given the fact that certain identifiers have more discriminatory power than others do[3]. The linkage procedure was first developed using pancreatic patients and then validated with COPD patients. Residential History: After InfoUSA family ID was linked to name and address from the EHR, residential history was quantified by tracking family IDs across prior years of InfoUSA data. Past addresses from the EHR were used to address discrepancies if they were recorded in EHR but was not found in InfoUSA. We compared demographic and neighborhood socioeconomic status (i.e., the Agency for Healthcare Research and Quality [AHRQ]) neighborhood index) among patients we were and were not able to link.

Results:

Our data linkage approach was able to match 89% of pancreatic cancer patients from the EHR with residential mobility data in infoUSA. Deterministic linkage matched 85% of pancreatic cancer patients and probabilistic linkage approaches added an additional 4%. When we extended these approaches to COPD patients, a total of 83% of COPD patients were linked to InfoUSA via deterministic approaches; probabilistic linkage further increased the matching rate by 1%. Patients for whom we were unable to link EHR data with InfoUSA data were more likely to be male, younger, and Black. There were no appreciable differences in the blockgroup neighborhood SES of linked and unlinked patients.

Table 1: Characteristics of linked and unlinked patients

	Pancreatic Cancer		COPD	
	Linked	Unlinked	Linked	Unlinked
Age, years (Mean)				
>65	73	66	50	33
25-65	27	32	47	49
18-25	0	1	3	17
Male (%)	52	50	37	41
Race (%)				
Caucasian/White	69	57	69	56
Black/African American	21	26	26	31
Asian	1	2	1	2
2 or more Races	1	2	1	2
Other	8	13	3	9
AHRQ nSES score (median)	51	50	51	51

Conclusion:

We developed data linkage approaches that were able to quantify residential history for a large proportion of patients across two clinical areas. A small proportion of patients were unable to be linked and tended to be male, younger, and Black. Quantifying residential history may allow for more accurate classification of prior exposure.

References

1. Ted E, Ben F, Kosuke I. FastLink: fast probabilistic record linkage with missing data. 2020; Version 0.6.0 [R]. Available from: https://CRAN.R-project.org/package=fastLink.
2. Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association. 1969; 64(328): 1183-1210.
3. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? a simulation study. Journal of Biomedical Informatics. 2015; 56: 80-86.

Comparison and Analysis of Concordance for Two Popular Geocoding Methods Applied to Electronic Health Record Data

Selah Lynch MS[1], Jessica R. Meeker MPH[4], Emily Schriver MS[2], Andy Cruz[2], Nebojsa Mirkovic PhD[2], Kehinde Oyekanmi BS[3], Eugenia South MD[3], Danielle L. Mowery PhD[1,4], Mary Regina Boland PhD[1,4]
[1]Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA,
[2]Data Analytics Center, University of Pennsylvania Health System, Philadelphia, PA, USA
[3]Emergency Medicine & [4]Department of Biostatistics, Epidemiology & Informatics,
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Abstract

We compared two methods of geocoding (Google API and ArcGIS) using 91,166 unique addresses from the Penn Medicine electronic health record (EHR). We found high agreement between methods; 88.6% of addresses yielded coordinates within 200ft of each other.

Introduction

Geospatial information greatly enhances the possibilities of medical research with the EHR. For example, neighborhood-level variables such as education level, poverty, crime, and pollution are all determinants that can profoundly influence health, but are not well documented in the EHR.[1] Geocoding, the process of converting text-based addresses into geographic coordinates, is a crucial step to studying social determinants of health. Furthermore, discrepancies in geocoding results may lead to contradictory policy decisions. While geocoding technologies are widely available, their performance is not well understood.[2]

Methods

For this IRB-approved pilot study, we selected EHR emergency department encounters that occurred from October 2011 to November 2014, that contained a Philadelphia zip code, and that had been previously geocoded using Google API from our EHR system. From these encounters, we selected four address fields: street address, city, state, and zip code. We performed preprocessing on the address field to strip off superfluous information, i.e., apartment or floor number and geocoded them using ArcGIS. We then compared the two distinct geocoding methods: Google API versus ArcGIS in terms of coverage, concordance, and location of geocoded addresses.

Results

Of the encounter, 323,315 encounters fit our criteria and had valid address information; 91,166 of these were unique addresses. Using the Google API, the 91,166 unique addresses were geocoded: 89,282 (97.9%) were successfully geocoded and 1,884 (2.1%) were not matched. In comparison, using ArcGIS, 88,494 (97.1%) were successfully geocoded and 2,672 (2.9%) were not matched. In some cases, the coordinates returned by the two methods differed. Overall, we found that 86,630 (95.0%) addresses were successfully geocoded by both Google API and ArcGIS, but the coordinates were sometimes different. Regarding concordance across methods, we observed that 80,778 (88.6%) coordinates were within 200 ft of each other. We manually reviewed the discordances in the geocodes and determined that the Google API autocorrects street name and zip code errors, which could explain higher coverage, but also have introduced other errors. We determined that most of the addresses that produced large geocode disagreements were addresses with invalid street numbers and geocodes that had been determined by interpolation.

Conclusion

Following pre-processing steps, we observed high coverage and concordance between geocoding methods. When considering either method, Google API or ArcGIS, for geocoding addresses to conduct an epidemiological study, researchers should consider differences in coverage as a result of variable approaches taken by each system which could have important implications for reproducibility.

References

1. Marmot M. Social determinants of health inequalities. Lancet. 2005;365(9464):1099–104.
2. Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. Ann Epidemiol. 2006 Nov;16(11):842–9.

Leveraging the ACT Network for Covid-19 Research

Doug MacFadden, MS[1]; Griffin Weber MD, PhD[1]
[1] Harvard Medical School, Boston, MA

Abstract
ACT is a federated query tool that searches the clinical data of more than 130 million patients at 47 CTSA institutions across the country. With the rapid increase of COVID-19 cases in March 2020, ACT decided to adapt the network to support COVID-19 research. We designed, implemented and tested changes to the ACT ontology to include COVID-19 related concepts and to the data import processes to ensure up-to-date data on patients with COVID-19.

Introduction
Launched in 2017, the ACT network allows investigators to phrase queries based upon a medical ontology, execute the queries across the federated network, and receive aggregate results from each site indicating the number of matching patients[1,2]. With the rapid rise of COVID-19 cases in March 2020, ACT leadership decided to focus ACT efforts on improvements that would enable it to query the clinical data of patients with COVID-19.

Methods
We made four changes to the ACT network to support COVID-19 research: (1) The first and most important improvement was developed by the Data Harmonization working group as additions to the query ontology for rapidly changing COVID-19 coding. These were vetted with domain experts and given to the ACT test network sites for trial implementation. Query ontology additions went through multiple iterations as different approaches were implemented and evaluated. Through this process we identified a necessary new feature needed within the ontology – derived terms. These are query terms that do not relate directly to EMR codes but rather contain multiple codes and Boolean logic to allow users to query COVID-19 concepts such as stage and severity of the illness. (2) Historically, sites in the ACT network refreshed their data (from their EHR to ACT database) monthly. To provide more recent data, which is critical since COVID-19 rates and standard of care change rapidly, the test network sites were asked to increase data loading frequency to twice weekly. In many cases this required a complete revamp of a site's ETL process. (3) The ACT Governance working group developed changes to documentation, agreements and processes that allow for the ACT network to be used for research. (4) Finally, the network operations group monitored progress, validated implementation and provided the most recent status of each site to users within the ACT query application itself for convenience.

Results
Each of these improvements were fast tracked, fully implemented in our test network and validated for production use in a rapid 8-week period. We began production network rollout of the COVID-19 improvements in June of 2020. We shared the final ontology with other informatics groups, national and international, to assist their COVID-19 efforts.

Conclusions
The combined improvements to the ACT network transformed it into an effective COVID-19 research tool available to all researchers at the 47 participating CTSA sites. Investigators can query both patients with COVID-19 as well as any of the other 130M patients in the ACT network for control groups (e.g., negative COVID-19 test results, other infectious diseases, etc.). In a retrospective of this overall effort, we now have a template for future changes to the ACT network to support any new disease area. This will be a critical new capability for ACT going forward.

References
1. Shyam Visweswaran, Michael J Becich, Vincent D'Itri, et al. Accrual to clinical trials (ACT): a clinical and translational science award consortium network. JAMIA open. 2018 Oct;1(2):147-52. PMID: 30474072
2. Andrew J McMurry, Shawn N Murphy, Douglas MacFadden, et al. SHRINE: enabling nationally scalable multi-site disease studies. PLoS One. 2013;8(3): e55811. PMID: 23533569; PMCID: PMC3591385.

COVID-19 Research: Messy Data, Consequent Pitfalls and Lessons Learned

Tanja Magoc, PhD[1], Gloria Lipori, MT, MBA[2], Jennifer Myles, PhD[2], Scott Sortino[2], Christopher A. Harle, PhD[1]
[1]University of Florida, Gainesville, FL; [2]UFHealth, Gainesville, FL

The COVID-19 pandemic has prompted researchers to seek electronic health record (EHR) data to predict disease spread, understand health outcomes, and develop and test interventions. However, researchers' ability to use EHR data has been challenged by COVID-19's novelty, and thus the lack of available data standards and inconsistent use of existing standards. Thus, it is important to share lessons learned in using EHR data for COVID-19 research. In this study conducted at University of Florida Health, we describe trends in diagnosis and laboratory terminology standard usage in EHR data in the early stages of the pandemic, and related lessons learned for using these data in COVID-19 research. We analyzed a de-identified research dataset extracted from the UF Health's enterprise data warehouse. The dataset included EHR information on all known COVID-19 positive patients, all patients tested for COVID-19, and patients with COVID-like symptoms recorded between January 1, 2020 and July 22, 2020.

When the World Health Organization declared a public health emergency in January of 2020, there was no COVID-19 International Classification of Diseases (ICD) diagnosis code. The ICD-10 code, U07.1, was made available for emergency use on April 1, 2020. In our analysis, we found that only about one-third of patients with a record of a positive laboratory test for COVID-19 also had diagnosis code of U07.1 in their EHR. Reasons for this inconsistency included laboratory results recorded before April 1, current (as of July 22) inpatients or recently discharged patients whose discharge diagnoses were not finalized, and patients whose laboratory tests were conducted at our institution, but for whom no other EHR information was recorded.

We identified records containing a U07.1 diagnosis code for patients where no COVID-19 laboratory test was recorded (approximately 2,000 patients) or where recorded tests had negative results (approximately 1,600 patients). Many of these cases may be attributed to the early stage of the pandemic when laboratory tests were not widely available. This example also suggests that U07.1 was used regularly to identify suspected COVID-19 or intent to test for COVID and not a definitive diagnosis. Thus, researchers should be cautious in identifying COVID infection using ICD-10 codes.

Relative to ICD diagnosis codes, we have found laboratory test results to be more reliable indicators of Coronavirus infection. However, tests for COVID-19 have been developed by various companies. Each of these tests have been assigned a Logical Observation Identifier Names and Codes (LOINC) code. Given patient surges and supply limits, healthcare organizations, including our own, have used test kits from multiple companies and outsourced some testing to commercial labs without knowing the exact test that was used. In these cases, we found that laboratory result records remain without a LOINC code for extended periods of time. Thus, using LOINC codes to identify COVID-19 positive patients in EHR data may lead to patient undercounts.

Aggregating data from multiple sites has also been used in analysis of COVID-19. Our organization covers three campuses, each of which includes one or more hospitals and affiliated clinics. Each campus developed its own protocols and processes for COVID-19 testing. These processes varied based on the availability of tests, the presumed positivity rate of the local population, and other factors. As a consequence, the presence of a COVID-19 test in the EHR for patients across campuses, especially in the early stages of the pandemic, may be the result of very different clinical decisions or protocols. When conducting research on COVID-19 test data, it's important to understand variation across location and time of testing protocols and processes.

Lastly, due to COVID-19's effect on hospitals' capacity, and the novelty of the disease some patients are being tested, hospitalized, or otherwise served by health systems from which they have not typically sought care in the past. At our organization, almost one-quarter of COVID-19 positive patients have no EHR data at our institution prior to January 1, 2020. This may limit the comprehensiveness of EHR data for conducting COVID-19 research.

In conclusion, COVID-19 has highlighted both new and well-known limitations of using EHR data in research. Thus, researchers using EHR data to study COVID-19 should be cautious in their analyses and conclusions, and thoughtful in understanding the processes by which COVID-related data are generated in EHRs.

Modeling Alzheimer's Disease by Combining Knowledge Extracted from Biomedical Literature with Biomedical Ontologies

Scott A. Malec, PhD, Sanya B. Taneja, MS, Kailyn F. Witonsky, BS, C. Elizabeth Shaaban, PhD, MPH, Helmet T. Karim, PhD, Arthur S. Levine, MD, Steven M. Albert, PhD, MS, Paul W. Munro, PhD, Richard D. Boyce, PhD
University of Pittsburgh, Pittsburgh, PA USA

Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disease with a significant and growing socioeconomic burden without effective preventive or therapeutic agents. The multitude of complex and varied neurobiological mechanisms contribute to the difficulty in developing effective interventions. Routinely collected observational data such as clinical notes may contain clues concerning risk-modifying targets for AD. However, to distinguish associations from genuine causal relationships from such data, we need to identify confounders (common causes of the exposure and outcome), mediators (intermediate variables between the exposure and outcome), and colliders (common effects of the exposure and outcome), as per Figure 1. This knowledge tells us whether or not to adjust for a variable: adjusting for a collider or mediator will induce bias. However, such causal knowledge is complex, the literature is vast but incomplete, and machine reading may be inaccurate.

Methods

We developed a pipeline to refine knowledge mined from the literature to infer causal variables using a knowledge graph (KG), a graph-theoretic formalism for processing the semantics of computable knowledge. After consulting with a health sciences librarian to construct a Pubmed query to scope the literature, we extracted and harmonized outputs from two machine reading systems, INDRA[1] and SemRep[2]. Next, we combined the extracted knowledge with a robust ontology-based KG framework developed by computational biologists[3]. Finally, to search for variables linking depression and AD, we transformed standard epidemiological definitions of causally relevant variables into queries that apply Dijkstra's shortest path algorithm to search the KG. To gauge plausibility, we compared our output with papers proposing pathways linking depression and AD[4,5].

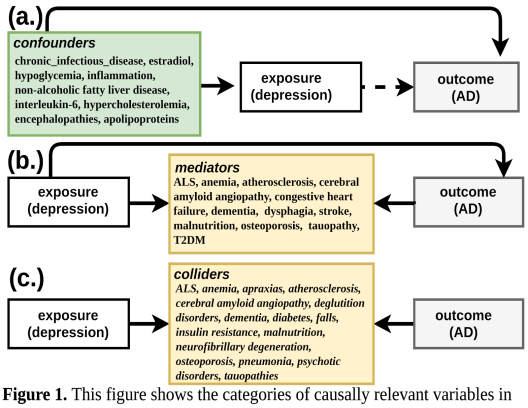


Figure 1. This figure shows the categories of causally relevant variables in terms of directed acyclic graphs, or DAGs, where the arrows indicate a causal dependency between variables. We have included sample confounders, mediators, and colliders from our depression and AD use case.

Results

A total of 13,365 PubMed-indexed articles were returned from PubMed. 226,997 subject-predicate-object triples were extracted by the machine readers, including 10,020 unique UMLS concepts. 2504 concepts were mapped to the merged ontologies in PheKnowLator. Variable search methods identified 126 confounders, 18 mediators, and 28 colliders. Curiously, related conditions involving hypersensitivity to blood glucose levels, e.g., hypoglycemia and T2DM, were identified in all three categories of causal variables.

Discussion and Conclusion

The many identified confounders, mediators, and colliders confirm the complexity of third-factor variables. The existence of problematic variables that fulfill multiple causal roles strongly suggests the value of a combined machine-human strategy. The next steps include a review by subject-matter experts of the variables and to use the KG-derived adjustment sets to help answer causal questions about AD from EHR-derived data.

References

1. Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK. From word models to executable models of signaling networks using automated assembly. Mol Syst Biol [Internet]. 2017 Nov 24 [cited 2020 Aug 14];13(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5731347/
2. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 2020 May 14;21(1):188.
3. Callahan TJ, Tripodi IJ, Hunter LE, Baumgartner WA. A Framework for Automated Construction of Heterogeneous Large-Scale Biomedical Knowledge Graphs. bioRxiv. 2020 May 2;2020.04.30.071407.
4. Ownby RL, Crocco E, Acevedo A, John V, Loewenstein D. Depression and Risk for Alzheimer Disease. Arch Gen Psychiatry. 2006 May;63(5):530–8.
5. Butters MA, Young JB, Lopez O, Aizenstein HJ, Mulsant BH, Reynolds III CF, et al. Pathways linking late-life depression to persistent cognitive impairment and dementia. Dialogues Clin Neurosci. 2008 Sep;10(3):345–57.

Loss of diagnostic data due to claim form limitations

Craig S. Mayer, MS[1], Nick Williams, Ph.D[1], Vojtech Huser MD, Ph.D[1]
[1]National Library of Medicine, NIH Bethesda, MD

Introduction

The Virtual Research Date Center (VRDC) provides Medicare claims data for institutional and professional claims. Institutional claims are submitted by facilities, mainly hospitals, while professional claims are submitted by healthcare professionals. Institutional claims are filed on form UB-04 with a maximum of 25 diagnosis (Dx) slots, while professional claims are filed on form CMS-1500, with 12 Dx slots. In contrast, Electronic Health Record problem lists have no such restrictions on the amount of diagnostic data able to be recorded. From a data scientist perspective, our hypothesis was that the lack of Dx slots on claim forms (different for each form) might be arbitrarily limiting the volume of diagnostic data and the removal of such restrictions may provide valuable additional diagnostic data.

Methods

We analyzed the density of diagnostic data in 2016 inpatient and outpatient Medicare data by splitting it into 4 categories, taking into account visit setting (inpatient [IP; dark shade] or outpatient [OP; light shade]) and claim type (institutional [Inst; red color] or professional [Pro; blue color]). For professional claims, we used place of service (POS) to properly classify claims as IP-Pro and OP-Pro (specifically IP-Pro: POS code 21 (inpatient hospital), OP-Pro: POS codes 11 (office) and 22 (on-campus outpatient hospital). These codes cover the most used POS codes. For all four categories, we analyzed the usage of Dx slots by calculating the percentage of claims in each category using each slot position. We also calculated what percentage of claims used all available slots allowed by the form ("hitting the wall", shown by vertical dotted lines in Exhibit 1b) and at what diagnosis slot, claim submitters naturally run out of diagnostic information they wish to provide (point where less than 2% of claims utilize a hypothetical "last" Dx slot, we refer to this as "hitting the floor).

Results and Conclusions

Type	# of Claims (%)	# of Dx Events (%)	Avg. slots used
IP-Inst	11.3 M (1.3)	162.5 M (6.4)	14.4
OP-Inst	172.7 M (20.6)	576.7 M (22.8)	3.34
IP-Pro	112.7 M (13.7)	335.9 M (13.3)	2.9
OP-Pro	543.2 M (64.5)	1,453.2 M (57.5)	2.68

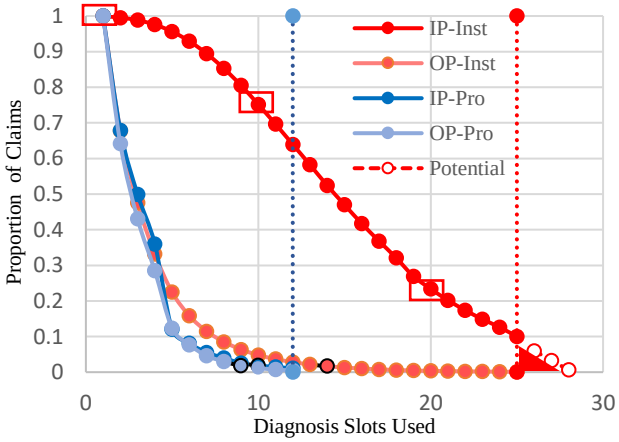


Exhibit 1. a) Claim metrics b) Proportion by slots graph.

We analyzed 842 855 146 claims with a total of 2 528 263 266 diagnostic events. By count of diagnoses, OP-Pro is the largest (57.5%) but the average number of Dx slots utilized is the fewest (2.68) (see Exhibit 1a). The graph in exhibit 1b shows how the proportion of claims (y axis) decreases based on the amount of diagnosis slots used (x axis). IP-Inst claims use of slots drops at the slowest pace: looking at the red boxes the graph shows that 100% of IP-Inst claims use at least one diagnosis slot. It decreases to 75.2% that use at least 10 diagnosis slots and further decreases to 23.4% using at least 20 slots. For the three categories that hit the floor, alternative thresholds can be considered; see points with black outline in the graph (e.g., OP-Inst: 14 slots of 25). 10.01% of IP-Inst claims are hitting a wall (use all available slots). Removal of UB-04 form restrictions may be in the interest of data scientists (based on our prediction [points with red outline] and may add 1.1 million diagnostic events [red triangle]). This, however, would not be in the interest of healthcare plans (case severity impacts reimbursement levels [an IP specific phenomenon]).[1] As future work, we plan to analyze this density and limits phenomenon in non-US data and examine the impact of length of stay on IP Dx density (including multi claim stays).

References

1. Compare and Contrast Physician and Outpatient Facility Coding (AAPC). [cited 2020 Mar 12]. Available from: https://www.aapc.com/blog/29346-compare-and-contrast-physician-and-outpatient-facility-coding/

Decision Support to Engage Consumers in the Selection of Health Benefits

Mollie M. McKillop, PhD, MPH[1], Bedda L. Rosario[1], PhD, Anita M. Preininger, PhD[1], Carla Huff[1], Nawshin Kutub, PhD[1], Gretchen Purcell Jackson, MD, PhD[1,2]
[1]IBM Watson Health, Cambridge, MA,
[2]Vanderbilt University Medical Center, Nashville, TN

Introduction

Understanding health insurance options is important during health plan selection as consumers often lack health insurance literacy.[1] While decision support tools exist, they are often limited in their functionality beyond expected cost information. Benefits Mentor (BM) is a decision-support tool which promotes health insurance literacy through interactive features to help consumers select a plan that best fits their needs. This includes a conversational agent that uses natural language processing to answer users' questions about health plans, an embedded glossary that contains over 35 medical and health benefit terms, and video content that explains options and benefits. The system uses employees' prior claims data to estimate utilization and inform plan choice. In this exploratory study, we describe user engagement and satisfaction with the tool during annual health plan selection.

Methods

Health plan decision support was deployed at 12 diverse employer organizations in the US including insurance, manufacturing, pharmaceutical, banking, technology, higher education and government sectors. A retrospective, post-implementation study examined usage data and satisfaction ratings by employees during annual health plan enrollment from 8/2019 to 12/2019. After users completed or exited their session, we measured factors related to engagement and satisfaction including (1) time, (2) number of steps/transitions in a conversation or 'turns' (with one turn representing one question-response pair), (3) number of glossary requests, (4) number of videos the user accessed, (5) number of videos viewed, in whole or in part, and (6) average willingness to recommend the system, from 0 (unwilling) to 10 (very willing) using an online survey. The study was exempt from human subjects' research review per the Western Institutional Review Board.

Results

The usage rate among employees eligible to use the tool was about 23%. There were 37,793 unique users from 12 employer organizations with one or more conversational sessions during the study period, resulting in a total of 52,148 sessions. We identified 31,291 unique users that had a single conversation with the system, representing 82.8% of all sessions. Users averaged 25 turns per conversation with the conversational agent. Average willingness to recommend the tool was 8/10; most conversations included at least one glossary request and video view (Table 1).

Table 1. Engagement with Benefits Mentor features across all sessions (N=52,148)

Measure	Minimum	Maximum	Median (First Quartile, Third Quartile)
Length of conversation (minutes)	0	94	3.27 (1.22, 6.87)
Number of transitions in a conversation	1	266	20 (11, 34)
Number of glossary requests	1	10	1 (1, 2)
Number of videos user exposed to	1	5	1 (1, 1)
Number of videos the user saw	0	4	1 (1, 1)
Willingness to recommend	0	10	8 (7, 10)

Conclusion

Deployment of health plan selection decision support resulted in high levels of engagement, with avid use of conversational agent functionality and multimodal information sources and high satisfaction scores. This study is limited by only examining users. Additional ongoing work is investigating long-term user satisfaction with health plan choices, literacy, and cost savings with non-users as a comparison group. We are also exploring user experience.

References

1. Tipirneni R, Politi MC, Kullgren JT, et al: Association Between Health Insurance Literacy and Avoidance of Health Care Services Owing to Cost. JAMA Netw Open 1:e184796, 2018.

Answering Common COVID-19 Questions with Conversational Technology

Mollie M. McKillop, PhD, MPH[1], Brett R. South, PhD. MS[1], Anita M. Preininger, PhD[1], Gretchen Purcell Jackson, PhD, MD[1,2]
[1]IBM Watson Health, Cambridge, MA;
[2]Vanderbilt University Medical Center, Nashville, TN

Introduction

The novel severe acute respiratory syndrome virus 2 (SARS-COV-2) has infected millions worldwide, resulting in significant mortality from coronavirus disease 2019 (COVID-19). With the potential for severe clinical complications from this new pathogen, information regarding the disease and prevention is needed. Stakeholders including public health departments and local governments have taken steps to help communities control the spread of disease through curation and dissemination of information related to COVID-19. Providing accurate and timely information is challenging, especially in the face of rapidly evolving science. To address this, automated information delivery with the use of conversational agents, which use natural language processing to answer user questions, shows promise.[1]

System Description

Watson Assistant (WA) is a platform for building a natural language conversational interface into any application, device, or channel such as a website or automated voice system, which can be trained to include customized information related to a specific language, locale, or organization. The core natural language capabilities of WA include: (1) understanding content, (2) classifying topics, (3) retrieving information from a knowledge base, and (4) generating natural language responses. When a user enters common questions about COVID, WA interprets the question to identify the *intent* (target of a user's query) and match it to an internal list of intents and *entities* (for example, a drug or condition) that answer the question. WA can dynamically search, identify, and abstract information on a daily basis from unstructured documents and trusted websites, leveraging evidence-based sources such as guidance from the United States (U.S.) Centers for Disease Control and Prevention (CDC). To provide the latest information, WA treats the user input as a search query. It finds information that is relevant to the query from an external data source, such as the CDC, and returns it to the user. For COVID-19 conversational agents built with WA, content validation was performed by a team of clinical and public health experts.

Usage

Between March 30, 2020 and June 22, 2020, WA for COVID was used by 51 organizations in 15 countries. The types of organizations implementing the tool were primarily government (N=27), hospitals and public health departments (N=16), employers (N=6), and health plan providers (N=2). WA for COVID was used in the U.S. and Canada (N=33), Europe (N=12), Asia Pacific (N=3), Latin America (N=2), and the Middle East (N=1). Most organizations deployed a web-based textual conversational approach (N=49); two organizations used voice technology.

Table 1. Usage Metrics from March 30, 2020 to June 22, 2020.*

Organization Type	Total Number of Messages	Number of Messages (Mean)	Conversational Turns (Mean)
Government (N=18)	2,937,299	163,183 (min = 406; max = 1294866)	2.96 (min = 1.75; max = 7.46
Hospitals, Public Health Depts. (N = 5)	6,923	1,385 (min = 617; max = 28665)	3.29 (min = 2.07; max = 6.76)
Employer (N=3)	4,794	1,598 (min = 448; max = 2979)	2.52 (min=2.08; max=2.90)
Health Plan (N=1)	2,643	2,643 (min = 2643; max = 2643)	2.06 (min = 2.06; max = 2.06)

*Available data provided by 27 participating organizations

Conclusion

We demonstrate the ability of a wide variety of organizations including governments, employers, providers, and payers to use conversational technologies to provide current information related to COVID-19. The WA platform enabled implementation of a set of conversational agents for a wide variety of use cases. Usage data show demand for and adoption of these technologies during a rapidly-evolving public health crisis. Our ongoing research is investigating user satisfaction and experience with COVID-19 conversational agents.

References

1. Miner AS, Laranjo L, Kocaballi AB: Chatbots in the fight against the COVID-19 pandemic. npj Digital Medicine 3:1-4, 2020

Combining Weather and Pollution Indicators with Insurance Claims on Identifying and Predicting Asthma Prevalence and Hospitalizations

Divya S. Mehrish[1], Monica Sharma[2], Laurent Hasson[2], J. Sairamesh (Ramesh), PhD[2]
[1]Stanford University, Palo Alto, CA; [2]CapsicoHealth, Palo Alto, CA

Introduction

Asthma impacts 1 in 12 children and 1 in 13 adults in the U.S., causing unplanned hospitalizations due to unhealthy and hazardous environmental conditions. Our study is one of the first to combine pollutants and weather conditions with insurance claims to correlate and predict asthma prevalence rates to target and alert at-risk patients[1, 2].

Study Design and Data

Our data includes de-identified CMS hospital inpatient data for both pediatric and adults patients (CMS Medicare and Medicaid) and Medical Claims summaries (CMS 2013-2017) as well as daily county-level temperature, precipitation, and pollution data for 2017 (EPA and NOAA). We focus on the seven counties in southwestern Pennsylvania surrounding Allegheny where 10% of adults and 10% of school-aged children live with asthma[3]. We examine over 200,000 patient records alongside county-specific minimum, maximum, and average values for NO_2, $PM_{2.5}$, PM_{10}, SO_2, and O_3. Our aim is to understand how asthma prevalence, hospitalization, and readmission rates correlate with seasonal weather and pollution data. Given patient demographics, clinical history, and environmental exposures, we use logistic and linear regression models to predict the probability of hospitalization and readmission.

Predictive Results

Our results show that asthma prevalence, asthma-related hospitalizations, and hospital readmissions are negatively correlated with external temperature (Figures 1, 2, and 3). Figures 4 and 5 show the ROC metric at 78% accuracy in predicting readmission and hospitalization risk. We find that the most predictive pollutant measures are $PM_{2.5}$ and NO_2. Hospitalization rate is negatively correlated with $PM_{2.5}$ and positively correlated with NO_2.

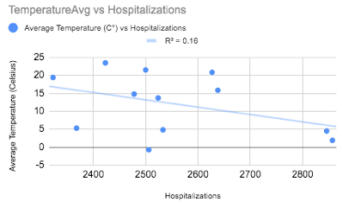


Figure 1. Avg. Temp. vs. Hosp.

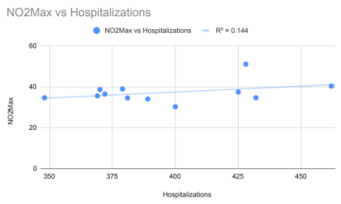


Figure 2. Max. NO_2 vs. Hosp.

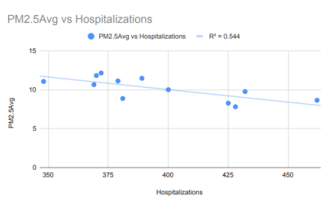


Figure 3. Avg. $PM_{2.5}$ vs. Hosp.

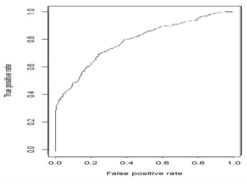


Figure 4. ROC (Readmission Risk)

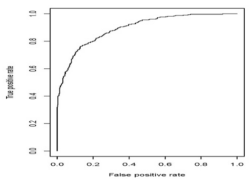


Figure 5. ROC (Hospitalization Risk)

Conclusion: Our study using CMS claims and external weather data shows that winter months (when temperature is low and pollution levels are moderate) correlate strongly with higher asthma rates and hospitalizations. Pollutants get trapped closer to the earth's surface due to winter smog and other factors. Our risk models, which combine clinical information with external indicators, can alert both hospitals in regions with large at-risk populations as well as vulnerable patients living in areas with unhealthy levels of pollution that can exacerbate asthma.

References

1. Patel, Molini M, and Rachel L Miller. "Air pollution and childhood asthma: recent advances and future directions." *Current opinion in pediatrics* vol. 21,2 (2009): 235-42. doi:10.1097/mop.0b013e3283267726
2. Soyiri IN, Sheikh A, Reis S, et al. Improving predictive asthma algorithms with modeled environment data for Scotland. BMJ Open 2018;8:e023289. doi:10.1136/ bmjopen-2018-023289
3. Hacker, Karen, Brink LuAnn, et al. 2015-2016 Allegheny County Health Survey

A Pilot Evaluation of the Performance of MetaMap for Processing Clinical Actionable Genomics Texts

Omika A. Merchant*, Shreya S. Tellur*, Xia Jing, MD, PhD[1]**

[1]Department of Public Health Sciences, College of Behavioral, Social, and Health Sciences, Clemson University, Clemson, SC, 29634, USA; * OAM and SST share the co-first authorship; **, corresponding author

Introduction

Natural Language Processing (NLP) can facilitate information processing efficiently. MetaMap, an NLP application, identifies key medical concepts from texts and maps them into the Unified Medical Language System (UMLS). Both MetaMap and UMLS, popular NLP tools and resources, are developed and maintained by the National Library of Medicine. One application of NLP is in precision medicine to extract clinically actionable genomics information automatically. Variations in MetaMap's different configurations and their combinations create significant changes in the outputs of MetaMap, which present a challenge among MetaMap users[1]. We conducted this pilot evaluation to compare the performances of MetaMap's configurations to set up a methodology foundation to seek the most optimal results in processing clinically actionable genomics texts in the future.

Methods

A medical professional manually marked relevant phrases in three input texts about actionable genomic information on cystic fibrosis to create gold standards, i.e., the hypothesized ideal value. We tested 17 Behavior configurations, individually, with the default settings: "Relaxed Mode", UMLS 2018AA, USABase to test if the selected configuration makes MetaMap perform better than the default settings (without any selected configurations). We compared each output to their golden standard and deemed them as exact, similar, or incorrect[2]. Then, precision, recall, and F-measure ($\beta = 0.33$) were calculated for each of the 17 configurations. F-measure represents MetaMap's ideal performance in considering both its precision and recall in processing the input texts. We deemed configurations as relevant, if the F-measure was ≥50%, or irrelevant, if the F-measure was <50%. The top 3 most relevant configurations were then combined and applied to all three texts to compare the MetaMap performance versus the default settings. The NLP evaluation principles were followed[3].

Results

Table 1 shows the top 8 most relevant configurations of MetaMap, under Behavior configurations, based on its corresponding F-measures by using the three input texts. According to the F-measures, Unique Acronym/Abbreviation Variants Only, Use Word Sense Disambiguation, and Allow Large N are the top three configurations. When combining these three configurations, the corresponding F-measures for the three input texts are much more effective/relevant than either the default settings or individual configurations.

Table 1 Resulting F-measures of each Behavior Configuration for Three Input Texts

Use Word Sense Disambiguation	65.40%	83.44%	76.90%
Allow Large N	65.50%	83.44%	63.70%
No Derivational Variants	58.90%	83.44%	63.70%
Enable NegEx	51.70%	83.40%	47.00%
No Text Tagging	51.70%	95.41%	51.20%
Composite Phrases	71.00%	69.52%	51.20%
Threshold (600)	65.50%	58.01%	63.70%
Default Settings (No Configurations)	39.30%	30.00%	71.00%
Combined Top Three Configurations	91.00%	94.00%	84.00%

Irrelvant F-measure: < 50%

F-Measures of Default/Combined Configurations

Conclusions

This study aims to evaluate MetaMap's[3] performance based on its Behavior configurations. We evaluated these 17 configurations, individually, and compared their corresponding performances to the default settings. The two authors (OM and ST) assessed the results based on the same metrics, independently first and obtained the agreement later. Our results show that the top three most relevant Behavior Configurations: Unique Acronym/Abbreviation Variants Only, Use Word Sense Disambiguation, and Allow Large N, have the highest F-measures and, therefore, perform better than the default settings in processing the selected clinically actionable genomics texts. Because, we only have one clinical annotator and three input texts in the pilot evaluation a more generalizable conclusion about MetaMap needs a more comprehensive test, including a diverse range of input texts to represent all biomedical domains. 17 Behavior configurations in MetaMap were tested for this pilot study; the evaluation of the permutations and combinations of all 17 configurations have not been conducted. We also consider automating the evaluation processes. Nevertheless, our methodology is a beneficial addition in evaluating MetaMap configurations and potentially improving NLP processing performance. In the future, we plan to use our limitations to guide future explorations in this project. (This work was partially supported by NIH under Award Number R15LM012941 and P20 GM121342).

References

1. Demner-Fushman, Dina, et al. "MetaMap Lite: An Evaluation of a New Java Implementation of MetaMap." Journal of the American Medical Informatics Association, vol. 24, no. 4, Oxford University Press, July 2017, pp. 841–44, doi:10.1093/jamia/ocw177.
2. Ruggieri, A. P., et al. "Representation by Standard Terminologies of Health Status Concepts Contained in Two Health Status Assessment Instruments Used in Rheumatic Disease Management." Proceedings - AMIA Symposium, 2000, pp. 734–38.
3. Friedman, C, and G Hripcsak. "Evaluating natural language processors in the clinical domain." Methods of information in medicine vol. 37,4-5 (1998): 334-444.

Impact of the ARX Anonymization Tool for Biomedical Data – A Review

Thierry Meurers, M.Sc.[1,2], Marco Johns, M.Sc.[1,2], Felix Wirth, M.Sc.[1,2], Fabian Prasser, Ph.D.[1,2]
[1]Berlin Institute of Health, Anna-Louisa-Karsch-Straße 2, 10178 Berlin, Germany
[2]Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany

Introduction

The ARX Data Anonymization Tool is a comprehensive open source software for anonymizing structured individual-level health data. Anonymization or de-identification can be performed to mitigate privacy risks when data is used for secondary purposes or shared with third parties. However, little has been published about real-world applications of anonymization tools, often because of legal uncertainty regarding the validity of anonymization processes. As a first step towards bridging this gap, we investigated the impact of ARX based on publicly available information.

Method

We searched Google for "ARX" or "arx.deidentifier.org" (the URL of the project website) in combination with one or more of the following terms: "anonymization", "data", "tool", "deidentification" and "privacy". Additionally, we searched for all scientific articles citing one of the 20 original papers published by the ARX development team using Google Scholar. We reviewed the 389 search results as well as the 435 scientific articles and selected references to ARX from three different kinds of resources: (1) guidelines and reports by authorities, (2) international research projects and (3) research data management guidelines and software. Finally, we report download numbers tracked via the project website and indicators for interest in the software provided by the development platform GitHub.

Results

We found 14 different reports and guidelines by national and international authorities referring to the software. They were released between 2015 and 2019. The institutions publishing these resources, their countries of origin and the titles of the publications are shown in Figure 1. In addition, we found 17 projects funded by the European Horizon 2020 program, which produced deliverables making use of or referring to ARX. We found five different academic research data management toolboxes mentioning ARX and five large software products, including SAP HANA Data Anonymization and the KNIME data analytics platform, that utilize the software or one of its core technologies. The graphical anonymization tool was downloaded 24,852 times and the programming library 3,741 times since the initial release in 2012. On GitHub the project has 346 stars, 164 forks, 20 contributors and is under active development receiving frequent updates.

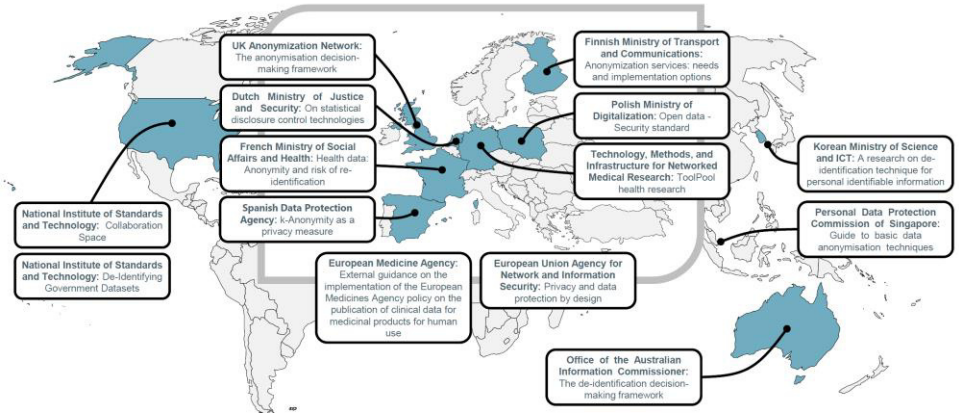


Figure 1. Guidelines and reports by national and international authorities mentioning ARX

Conclusion

We found that ARX is used for various purposes and is internationally recognized not only in the academic field, but also by authorities and government agencies, commercial enterprises and specialists from various fields. The download numbers suggest that the graphical user interface is an important aspect contributing to its success. Further information is available on the project website (https://arx.deidentifier.org).

Health Care Actors' Roles to Improve Quality with Clinical Data Exchanges

Sandra Mitchell, RPh, MSIS, FASHP[1], Gay Stahr, BS[1], Todd Turner[2], Jeffery E. Anderson, MD[3]

[1]Veterans Health Information Exchange (VHIE), J.P. Systems, Inc., Clifton, VA, [2]Todd Turner, Program Manager, Enterprise Solutions Management, VHIE, Clinical Informatics and Data Management Office (CIDMO), Veterans Health Administration (VHA); [3]Jeffery E. Anderson, MD, MS, Director, VHIE, CIDMO, VHA

1. Introduction

Clinicians, analysts, and patients alike are often preoccupied by the technical or administrative processes of interoperability in health care; though intended to be helpful, these processes often distract health care actors from interoperability's primary goal: improving patient care. True interoperability grants each clinician access to the same complete and high-quality patient data across the health care ecosystem. To achieve true interoperability, every actor in the health care ecosystem must understand that they have a role to play in improving clinical data quality that facilitates research and better care decisions for meaningful patient experiences and outcomes.

The Veterans Health Information Exchange (VHIE) Clinical Data Quality Management Program contributes to achieving optimal interoperability through its program of 1) continuous monitoring of the clinical quality of the nation's largest exchange of Health Level Seven (HL7) production messages, 2) educating health care actors of the team's discoveries and exposing opportunities on how these actors can play a part in attaining clinical data that is fit for purpose across the ecosystem, and 3) participating in improving national standards that influence interoperable exchanges of information.

2. Methodology – Mixed Approach

Analytics and metrics provide the cornerstone to build the data quality story for each specific organization. The C-CDAs exchanged are the primary insight stream to develop quantitative assessments and trends.

Clinician-facing issues are collected through semi-structured interviews conducted by VHIE Community Coordinators nationally across VA's health care system. These front-line users detail their usability challenges with specific data exchange partners. Additionally, a community-based participatory approach is used to engage with commercial health care organizations (e.g., software vendors and/or health care organizations) to gain understanding into pain points.

3. Evaluation Results

Vendor/organization findings are presented by the VHIE Clinical Data Quality Team with a clear focus on impactful data quality challenges. For example, in one Health Information Exchange's Allergy domain, 46.2% of data is missing a code and/or code system. In one Medications domain, 55.9% of data is missing a code and/or code system, immunizations are misplaced in the CCD-A, and fields are mis-aligned in format. Both of these examples show huge issues in data quality. The collaborative presentation includes Community Coordinator input, engages and develops consensus plans to resolve issues, and in the example above, reviews the domain code/code system mapping tables for specific stakeholders as a starting point. As a deliverable, a workbook of de-identified data issue examples (at the C-CDA document level) provide an easy tool for the source health care team to identify roles that can address specific data issues.

Too often the data exchange is a technical success, but clinical usability fails, leading to clinician frustration with the reality of missing, miscoded, or misplaced external data. Our research has also shown that vendors face challenges with multiple standards documents (e.g., standards, companion guide, implementation guide), leading to differences in implementations.

4. Conclusion

Informatics is the field responsible for developing the technology to collect, store, and use health care data and information with the purpose of improving patient care. However, when poor clinical data content is exchanged, even the most sophisticated cutting-edge technology will not be able to live up to expectations. Reasons for failure are often complex, and resolution must involve a spectrum of health care actors working together with a focus on the quality of the clinical content exchanged, not just a successful technical exchange. We can achieve this with education, collaboration, and involvement; everyone has a part to play. We can collectively shape the future of healthcare.

5. Attendee's Take-away Tool:

The entire health care interoperability ecosystem and every health care actor within it is responsible for the clinical quality of data.

Listening to your Data: Analyzing Medical Concept Mentions in Longitudinal Clinical Notes

Asher Moldwin, Dina Demner-Fushman, MD, PhD, Travis R. Goodwin, PhD
U.S. National Library of Medicine, Bethesda, MD, USA

Introduction

Computerized processing of clinical texts is important for tasks such as Disease Prediction and Clinical Decision Support.[1] Unlike the structured data included in medical records, clinical notes are a written description of any medically relevant information included by the physician. Because of their unstructured nature, clinical notes do not provide a consistent list of variables that can be easily analyzed in aggregate. For this reason, downstream applications often rely on medical concepts from the Unified Medical Language System™ (UMLS)[2] extracted from clinical notes rather than the full text of the notes themselves. While concepts have been used for medical NLP for many years, there has been little work exploring how the distribution of concepts changes over the course of the average patient's hospital admission. In this poster we provide several statistics about how concepts vary throughout a patient's stay using in the MIMIC-III critical care database. Specifically, we provide information about the time-distribution of concepts during a hospital admission as well as data about how different concepts correlate with specific patient outcomes.

Data

We used the MIMIC-III critical care database. MIMIC-III includes de-identified patient information from 46,520 patients comprising 58,976 distinct hospital admissions including 2,082,294 clinical notes (giving an average of 35.3 notes per admission).

Methods

As in Goodwin et al. (2020),[3] we represent hospital stays as a discontinuous sequence of "snapshots" of the patient's clinical picture, where each snapshot is represented by the set of concepts extracted from any notes provided on the same calendar day. These snapshots include 8 to 1216 medical concepts, with an average of 82 and standard deviation of 86.3. We analyzed (1) how the distribution of concepts changes over time, and (2) any correlations between concepts and outcomes including 25 disease phenotypes and mortality. Specifically, we calculated smoothed Pointwise Mutual Information (PMI), Fisher's Exact Test and χ^2.

Table 1: Correlation between UMLS concepts and Mortality and Disease Phenotype, using smoothed PMI

Mortality Prediction			Phenotype Prediction		
Concept	**Δ**	**PMI**	**Concept**	**Phenotype**	**PMI**
Blood in the esophagus	48 h	1.95	Repair of trachea	Other upper respiratory disease	2.72
Mesenteric Lymphadenitis	48 h	1.83	Pharyngeal Carcinoma	Other upper respiratory disease	2.64
Radiography of kidney-ureter-bladder	48 h	1.83	Mucin 5AC	Other upper respiratory disease	2.64

Results

The average Jaccard similarity between adjacent notes in the same admission was 26% and between the first and last note in each admission was 12%. Table 1 displays the three most-correlated UMLS concepts for mortality prediction and phenotype prediction.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

References

1. Demner-Fushman D, Chapman WW, Mcdonald CJ. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics 2009;42:760–72.
2. Humphreys BL, Mccray AT, Lindberg DAB. The Unified Medical Language System. Yearbook of Medical Informatics 1993;02:41–51.
3. Goodwin TR, Demner-Fushman D. A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision. Journal of the American Medical Informatics Association 2020.

The Patient Voice in Clinical Practice: Results of Qualitative Interviews with Patients Completing Patient-Reported Outcomes

Therese A. Nelson, AM, LSW[1], Faraz S. Ahmad, MD[1,2], Martha-Margaret Cotton, MPH[2], Kristina Davis, MSN, MPH[2], Leilani Lacson, MPH[1], Ryan Merkow, MD[1,2], Luke V. Rasmussen, MS[1], Nan E. Rothrock, PhD[1], Justin B. Starren, MD, PhD[1]
[1]Northwestern University Feinberg School of Medicine, Chicago, Illinois; [2]Northwestern Medicine, Chicago, Illinois

Introduction

The use of Patient-Reported Outcome (PRO) measures in the clinical setting can help clinicians track patient symptoms and function over time, elevating the voice of patients in their own care.[1,2] However, given their more recent introduction into clinical care, it is unclear how patients view PRO questionnaires and why patients often fail to complete PROs in advance of their clinical appointments. This poster presentation will share the patient view of PROs, the challenges they identified, and a prioritized list of recommendations.

Methods

The Electronic Health Record (EHR) Access to Seamless Integration of PROMIS (EASI-PRO) consortium consists of nine universities integrating PROs into EHRs. To learn about patient viewpoints, EASI-PRO researchers conducted 23 patient interviews across five clinics at one site. Transcripts were reviewed to examine patient experiences regarding PRO completion, reactions to PRO questions, and physician interaction.

Results

Patients noted a number of barriers to PRO completion, including a lack of patient portal access, email overload resulting in difficulty identifying important messages, confusion between PROs and healthcare satisfaction surveys, challenging physical health, and technical factors. Patients described their experience interpreting email prompts and advised how to make PRO requests more likely to be answered. Patients expressed confusion regarding the purpose of PROs and how they would be used and voiced a desire to learn how results would impact their clinical care.

Patients reported that PRO measures themselves were generally understandable but they described some difficulty with interpreting answers and matching those answers to the nuance of their experience. PRO length and content were appropriate. Comments demonstrated the importance of selecting PRO measures that are highly relevant to the patient population. Patients reported that completing PRO measures can result in feelings of introspection and gratitude.

Patients expressed a strong desire for quick communication of concerning scores even outside of appointment times and voiced their hope that physicians would utilize PRO results to enhance their care. Many patients assumed that the physician would take their PRO results into account and use results to prepare for their medical appointments.

Conclusion

In our study, most patients were willing to complete PROs, but barriers to completion hampered their response. In this poster presentation, we will present practical recommendations to address barriers, such as revising the call center script, creating descriptive messages, setting tablets at maximum time-out, communicating expected PRO completion time, framing PROs as a way to give providers more information before they start a face-to-face visit so that the time is maximized, informing patients about the purpose of PROs and their role and importance in clinical care, and refraining from using the word "survey." Recommendations also focus on patient desires concerning use of PROs in patient care, encouraging clinicians to acknowledge PRO completion and use in the clinical setting.

References

1. Greenhalgh J, Dalkin S, Gooding K, Gibbons E, Wright J, Meads D, Black N, Valderas JM, Pawson R. Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care [Internet]. Southampton (UK): NIHR Journals Library; 2017. Available from: https://www.ncbi.nlm.nih.gov/books/NBK409450
2. Patient-reported outcomes. National Quality Forum [Internet] [cited 2009 Oct 16]. Available from: https://www.qualityforum.org/Projects/n-r/Patient-Reported_Outcomes/Patient-Reported_Outcomes.aspx

BRATsynthetic: A Software Tool for the Generation of HIPAA Safe Harbor Compliant Replacement Text

Tobias O'Leary, B.S.[1] and John D. Osborne, PhD[1]
[1]University of Alabama at Birmingham, Birmingham, Alabama, United States

Abstract

Realistic synthetic replacement text for personal health information removed from clinical text may not be provided by de-identification software or in manually curated clinical data sets. This is a problem for machine learning and natural language processing algorithms that rely on the availability of realistic text for training. We present BRATsynthetic, an open source tool that generates realistic synthetic text replacements.

Introduction

The creation of application relevant, realistic, synthetic text by computer algorithm remains an open problem in Natural Language Processing (NLP), even with recent advances made by models such as GPT-3[1]. One application is the creation of synthetic data to replace personal information (PI) from text removed due to privacy, ethical or legal considerations. In the United States, the governing legal framework for privacy is Health Insurance Portability and Accountability Act (HIPAA), which under Section 164.512 of the Privacy Rule describes a "safe harbor" method that details categories of data elements for removal. While software exists to remove these elements[2], de-identification software may not replace identified PI with synthetic text or be unavailable to the broader community. This is unfortunate, because redacted text (such as text replaced simply by the category type of the PI removed) does not resemble the original text, making it difficult to train machine learning algorithms. The challenge of creating high quality de-identified synthetic text for use with machine learning may lead even large tech corporations to skip the process entirely, leading to PI leaks[3]. To address this issue, we have developed software to create synthetic data from the widely used BRAT[4] annotation software.

Method

BRATsynthetic utilizes regular expressions and the open source library faker to convert BRAT .ann formatted files and the associated text file to generate realistic text for 24 categories of PI from the I2B2 2014 challenge[5], with the addition of a TIME category. For each category we generate a range of data using regular expressions and basic text analysis to create synthetic data matching the format of the original text.

Results

Software is available on github at https://github.com/uabnlp/BRATsynthetic A preliminary qualitative evaluation reveals the generation of realistic text substitutions in the same format as the original text. For example, an original date of "june 2nd" will be replaced by a random date in a similar format such as "october 5th". Similarly, a hospital abbreviation (UAB) is replaced by another hospital abbreviation (BMC) from a replacement list. Some errors are still present, for example substitution of profession "cook" for "carpenter", when the person was injured in a restaurant.

Conclusion

In conclusion, we have released a new software tool to generate synthetic text covering all I2B2 2014 de-identification task categories. A formal evaluation of the software and an assessment of its impact for training is pending.

References

1. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. eprint arXiv:200514165. 2020.
2. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc. 2017;24(3):596-606.
3. Copeland R. Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans. The Wall Street Journal. 2019 Nov. 11, 2019.
4. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii Ji. BRAT: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics; Avignon, France: Association for Computational Linguistics; 2012. p. 102–7.
5. Stubbs A, Kotfila C, Uzuner O. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. J Biomed Inform. 2015;58 Suppl:S11-9.

The Application of Social Network Analysis for COVID-19 Nosocomial Infection Control

Mina Ostovari PhD, Claudine Jurkovitz MD MPH, Lee Pachter DO, Marci Drees M.D., MS, DTMH, FACP, FSHEA, David Chen MD MPH
Christiana Care Health System, Newark, DE

Introduction

The highly infectious clinical syndrome Covid-19 is caused by the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2). The virus can be transmitted from person to person through droplets in close contacts[1] [2]. As COVID-19 can spread quickly, the traditional contact tracing methods of contacting cases and identifying contacts could become cumbersome. In this study, we propose an approach using the application of network analysis on electronic health records to identify those healthcare workers (HCWs) with higher risk interactions with COVID-19 patients and other HCWs they interacted with (contacts). Our findings can help inform hospitals' future decision makings regarding the allocation of resources, safety protocols, and contact tracing in case of future pandemics.

Methodology

We used electronic health records to develop a small-world contact network of HCWs within a hospital system to facilitate contact tracing and infection control efforts. The study period was between April 1st to June 30th, 2020, when there were surges in the number of COVID-19 patients in the hospital. The patient population included those who were not initially hospitalized for COVID-19 but had to go to COVID isolation sometime after their admission as they were tested positive. We identified all HCWs who cared for these patients before they went into isolation and flagged them as having high-risk contacts with patients. For each HCW, we identified the last day they cared for a patient before the patient went into COVID isolation. We identified other patients of the HCW in the next 14 days after their last high-risk contact and all other health care workers of those patients. We only considered HCWs who had definite or potential contacts with patients. To distinguish different types of contacts, we categorized health care workers into groups (such as clinicians, administrative). We extracted the clinical events related to each group from the electronic health records. A physician and a resident went over the clinical events separately and categorized them into potential, definite, or no contact with patients. We removed the administrative staff who had no potential or definite contacts with patients. We generated a network of HCWs based on patient-sharing relations. The nodes represented HCWs, and the edges represented patients shared between them. The edge weights represented the number of patients shared between health care workers. We identified important HCWs in the network using centrality measures. We used the Louvain community detection algorithm to identify groups of HCWs more densely connected. We also identified the central HCWs in each community using the centrality measures.

Results

During the study period, 863 patients went into COVID isolation sometime after their admission. The total number of patients used for generating the network was 21,060. The network was undirected, weighted, and connected with 3932 nodes (health care workers), 367,803 edges (patient shared between health care workers), and a network density of 0.047. Healthcare workers shared an average of 2 patients (average edge weight). We calculated degree, weighted degree, closeness, and betweenness centrality measures for the network nodes. Based on degree and weighted degree, the central node was a respiratory staff with direct high-risk contact with COVID-19 patients (patients who were not initially admitted for COVID-19 and went into COVID isolation sometime after their admission). Based on the betweenness centrality, the central node was a doctor with no high-risk contact with COVID patients. Based on the closeness centrality, the central node in the network was dialysis staff with high-risk contacts with COVID patients. The Louvain community detection algorithm detected 8 communities in the network with ranges from 132 to 1421 nodes. The network modularity was 0.4. We calculated centrality measures for all nodes in each community. Respiratory staff, patient care technicians, and registered nurses were identified as central nodes in the communities.

Conclusion

This study approach could inform the policies regarding infection control procedures and enhance contact tracing efforts in hospitals.

References

1. van Doremalen, N. *et al.* Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *N. Engl. J. Med.* **382**, 1564–1567 (2020)
2. Chan, J. F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* **395**, 514–523 (2020)

Chronicles of Nationwide Health Information Exchange (HIE) Told through Data Profiling of the Veterans HIE Audit Log

Eric Pan, MD, MSc, FAMIA, Nelson Hsing, ScD, Omar Bouhaddou, PhD, Nathan Botts, PhD, Bharathi Vedula, MS, Jim Malpass, Jeffrey Anderson, MD, Jonathan Nebeker, MD
Veteran Health Information Exchange, Clinical Informatics and Data Management Office, Department of Veterans Affairs, Washington, DC

Problem Statement and Methodology

Given its nationwide footprint, the Department of Veterans Affairs (VA) exchanges health information with a large number of organizations and consequently has a unique opportunity to experience and share most issues affecting interoperability. The history of these exchanges is recorded in the Veterans Health Information Exchange (VHIE) log files, including details about the patient, sender organization, user requesting the information, description of the documents requested as well as the documents themselves, purpose of use of these documents, and date and time. The primary use of the auditing information is to report an accounting of disclosures (a privacy requirement). The audit data is also transferred to an analytics platform where it is leveraged to monitor operations (system response time), assist in measuring the value of different features (such as value of pre-fetching data ahead of an appointment), assess data quality (completeness and relevance), and evaluate the impact of HIE on clinical and financial outcomes.

This study profiles the 10-year data in the audit logs of VHIE to chronicle the story of nationwide health information exchange and draw lessons learned. Several studies have used HIE audit logs to draw conclusions on utilization and potential benefits of HIE [1]. To our knowledge, none have provided a detailed description of these audit logs whose schema can be critical for effective HIE monitoring and evaluation.

Results

There is an **exponential growth of nationwide record exchange**, especially in the last four years. Robust technology platforms with scalable architecture (hub and single on-ramp). Veterans **patient** population served with these exchanges has been limited, due to a strict consent requirement. This was addressed with a recent policy change (MISSION Act). Patients have information with one external private sector organization (78%), two (17%), and three or more (5%). The **Purpose of Use** is 99% treatment, 0.5% benefits determination and 0.5% other. Expansion will require configurable policy engines. The **document types** are the CCDs (53%), progress notes (24%), consultation notes (7%), radiology studies (4%), and others (12%). A third (34%) are unstructured documents. More modular data packages like FHIR resources may be needed. Healthcare staff or clerical **users** represent 66% and healthcare professionals 20%, which is expanding based on better integration between EHR and HIE. Unfortunately, the current audit logs do not tell the whole story. They miss essential elements, such as an indication of the actual use of the data and the user-perceived value of the information at the time of use. Extended auditing covering end-to-end workflows are needed.

Conclusion

Health information exchange is the current way providers keep each other informed about their shared patients. Audit logs of these HIEs record the exchange transactions. Profiling the 10-year VA HIE audit logs tell the story of development and usage characteristics of nationwide HIE. The story shows exponential growth in the volume of transactions, and a growing number of users and roles who are exchanging a diverse set of document types, for many purposes. Compliance with standards is variable and actual end-user usage is missing from the story. Maintaining and extending these log files are critical for analyzing existing exchange patterns and forecasting future HIE needs and trends.

References

1. Devine E, Totten A, Gorman P, Eden K, Kassakian S, Woods S, Daeges M, Pappas M, McDonagh M, Hersh WR. Health information exchange use (1990-2015): a systematic review. EGEMS 2017 Dec 7;5(1):27

Communication Efficient Distributed Tensor Factorization based on Local SGD for Collaborative Health Data Analysis

Zhangyi Pan, BS[1], Jian Lou, PhD[1], Li Xiong, PhD[1], Jing Ma, MS[1]
[1] Department of Computer Science, Emory University, Atlanta, GA, US

Abstract

Federated tensor factorization has been proposed recently to analyze massive medical data. Existing methods either introduce high communication overhead or lead to inferior results. We propose a communication efficient approach called LocalTF, where the local sites only communicate after several iterations of local updates. Experiments on real medical datasets verify the accuracy improvement and communication reduction of LocalTF.

Introduction

Nowadays, more and more institutions store clinical histories in the form of Electronic Health Records (EHRs). Recent studies proposed federated tensor factorization to combine EHR data from different institutions for tensor analysis. Current federated methods either introduce high communication overhead [1] or lead to inferior results due to auxiliary penalty [2]. In this research, we propose localTF, a communication efficient and privacy preserving distributed tensor factorization approach based on Local SGD.

Methods

Suppose we have a three-order tensor with modes of patients, procedures and diagnoses. Since our algorithm is a distributed one, we have to compute phenotypes while protecting the privacy of the patients in local sites. As a result, the local sites only send procedures and diagnoses to the global server. We set a batch size b and we only globally communicate after b local updates and then add Gaussian noise. After global communication, the server sends the updated information back to the local sites and local sites continue their local updates. The process continues until convergence.

Results

We report the factorization accuracy under Root Mean Square Error (RMSE). LocalTF achieves the best accuracy, and CP-ALS ranks the lowest. CP-ALS has the best stability. We use the mortality prediction task to measure how useful the phenotypes are. We evaluate the prediction result with the AUC score. CP-ALS has the best AUC score, and LocalTF ranks the second. For communication costs, both DPFact and LocalTF significantly reduce communication cost compared with methods based on Distributed SGD, and LocalTF has a slightly lower cost than DPFact.

Conclusion

LocalTF is a useful tensor factorization method that guarantees strict privacy and reduces communication costs. Our model outperforms the state-of-the-art federated tensor factorization method and also achieved comparable results when compared with centralized methods. Future works include the evaluation of the methods on more datasets (e.g. higher order) and evaluate the quality of the phenotypes with guidance of domain expert.

References

1. Kim Y, Sun J, Yu H, Jiang X. Federated Tensor Factorization for Computational Phenotyping. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 887–895. Available from: https://doi.org/10.1145/3097983.3098118.

2. Ma J, Zhang Q, Lou J, Ho JC, Xiong L, Jiang X. Privacy-Preserving Tensor Factorization for Collaborative Health Data Analysis. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 1291–1300. Available from: https://doi.org/10.1145/3357384.3357878.

Social determinants of health and opioid dependence and overdose

Apoorva M. Pradhan, MD, MPH[1], Tim Oates, PhD[2], Fadia T. Shaya, PhD, MPH[1,3]
[1]University of Maryland School of Pharmacy, Pharmaceutical Health Services Research Department, Baltimore, MD, USA; [2]University of Maryland Baltimore County, Department of Computer Science and Electrical Engineering, [3]Institute for Clinical and Translational Research, Baltimore, MD, USA

Introduction: Negative impacts of opioids continue to rise, with little published evidence on how they are affected by social and demographic factors[1]. It is further unclear whether a diagnosis of opioid use disorder (OUD) is a strong predictor of opioid overdose (OD)[2]. The purpose of this study is to evaluate the association between social and demographic factors and the likelihood of developing an OUD or having an overdose (OD), in patients who are prescribed opioids and are covered under a commercial insurance plan.

Methods and findings: The study design is a longitudinal retrospective cohort in IQVIA Pharmetrics data covering a 9 year period from 2007-2015. The index date was established as the date of the first opioid prescription, and OUD and OD were identified using ICD-9 codes. We built simple and multiple logistic regression models, adjusting for confounders, to assess the association between social and demographic factors and the risk for either developing an OUD or having an overdose event, and we then examined the association between OUD and OD.
Findings: The study population includes 927,395 individual patients with at least one opioid prescription (Figure 1).

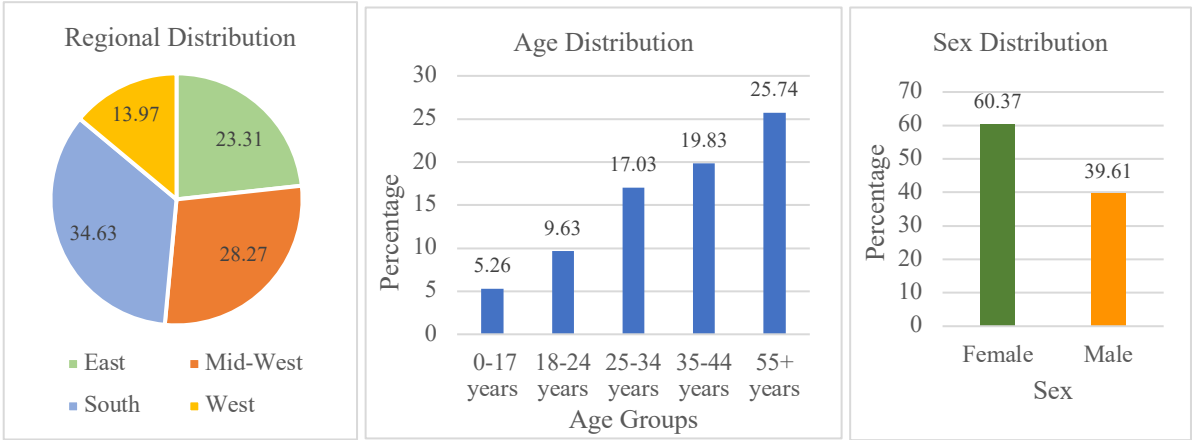


Figure 1: Demographic distribution of patients who received at least one opioid prescription in 2007-2015

We found that middle age (OR 7.09, CI 2.23-22.49), and residing in the southern region (OR 1.99, CI 1.32-3.01) were associated with a higher likelihood of OUD, whereas women (OR 0.71, CI 0.56-0.89) had a lower likelihood of an OUD. For OD, we found that southern region of residence was associated with a lower likelihood of OD (OR 0.09, CI 0.01-0.54), while the presence of OUD was associated with an exponential increase in the risk of OD (OR >999.9, CI 3.70- >999.9)

Conclusion: In this commercially insured population, we found that the likelihood of OUD is lower for women, and it is higher in middle age patients and in those residing in southern regions. The fact that there was no association between any socio-demographic factor and OD except place of residence and OUD, may suggest an over-reporting or over diagnosis of OUD in this commercially insured geographical population. These results support the call for ongoing dissemination and implementation research in opioid and clinical practice systems.

References

1. Dufour R, Mardekian J, Pasquale M, Schaaf D, Andrews GA, Patel NC. Understanding predictors of opioid abuse: Predictive model development and validation. Am J Pharm Benefits. 2014 Sep 1;6:208–16.
2. Glanz JM, Narwaney KJ, Mueller SR, , Gardner EM, Calcaterra SL, Xu S, et al. Prediction model for two-year risk of opioid overdose among patients prescribed chronic opioid therapy. J Gen Intern Med [Internet]. 2018 Oct;33(10):1646–53. Available from: https://doi.org/10.1007/s11606-017-4288-3

Development and Alpha Testing of Specifications for a VTE/Major Bleeding Electronic Clinical Quality Measure (eCQM)

Avery Pullman, BS[1], Troy Li, BS[1], Ania Syrowatka, PhD[1,2], Alexandra Businger[1], MPH, Michael Sainlaire, MS[1], David Bates, MD, MSc[1,2], Patricia Dykes, PhD, RN[1,2]
[1]Brigham & Women's Hospital, Boston, MA, [2]Harvard Medical School, Boston, MA, [3]

Background

The purpose of this study is to develop an electronic Clinical Quality Measure (eCQM) that uses electronic health record (EHR) data to assess both the venous thromboembolism (VTE) and major bleeding rates for patients following elective primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) at the clinician group level for adults 18 years and older. This measure is needed because THA/TKAs are the most common implant surgeries performed on Medicare beneficiaries[1] and are increasingly performed for younger and more active recipients[2]. However, patients undergoing these procedures are at risk of developing VTE, which includes deep vein thrombosis and pulmonary embolism. Studies have estimated that about 5% of patients undergoing THA/TKA develop VTE without anticoagulants, which, when overprescribed, can cause major bleeding and death[3]. The proposed measure aims to assess the balance between over-prescribing (bleeding) and under-prescribing anticoagulants (VTE) following surgery to assist providers in finding the appropriate prescribing regime.

Results

Denominator Statement: Patients, 18 years of age or older, undergoing an elective primary THA and/or TKA procedure who did not meet any exclusion criteria (*Figure 1*).

Numerator: Patients, 18 years and older, who had a major bleeding and/or VTE event occur from the date of the THA/TKA procedure to 35 days after the procedure.

Alpha testing: The target population consisted of 17,374 THA/TKA patients from January 2016 - December 2019 at Mass General Brigham (MGB). Exclusion analysis (*Figure 1*) and results (*Figure 2*) are provided below.

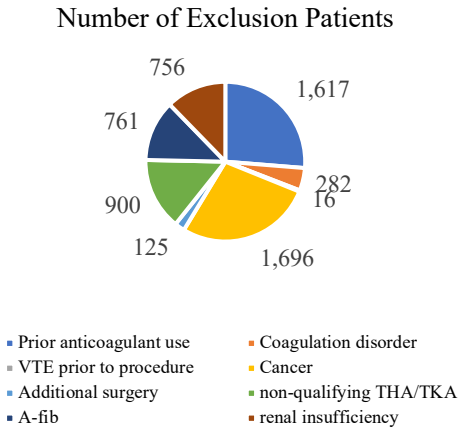


Figure 1: Exclusion analysis for 2016-2019 MGB data.

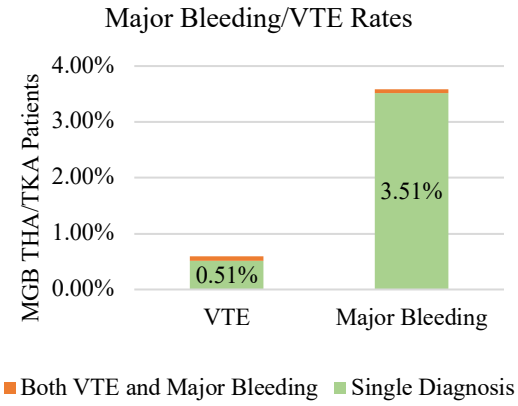


Figure 2: 2016-2019 MGB Major Bleeding and VTE rate.

Conclusions and Next Steps

This study reports on the initial testing of a THA/TKA VTE/Major Bleeding eCQM, demonstrating an MGB Major Bleeding and VTE rate of 3.51% and 0.51%, respectively, producing an overall rate of 4.02%. A risk-adjustment will be performed to determine the overall MGB risk-standardized rate as well as to determine the risk-standardized rates across the six individual MGB provider groups. Additional testing is underway at a geographically distant site using a different EHR system to confirm the feasibility of implementing this eCQM on another health system. Once validated, this VTE/Bleeding eCQM will enhance the Quality Payment Program by providing a performance measure capable of analyzing clinician performance in an area where there is currently a measurement gap.

References

1. CMS (2016). Medicare Program, Federal Register November 15, 2016.
2. CMS, Comprehensive Care for Joint Replacement Model, CMS.gov. https://innovation.cms.gov/initiatives/CJR. Updated 2/11/2019.
3. Lieberman JR and Heckmann N. Venous Thromboembolism Prophylaxis in Total Hip Arthroplasty and Total Knee Arthroplasty Patients: From Guidelines to Practice. J Am Acad Orthop Surg. 2017; 25:789-798.

A Novel Solution for Clinical Knowledge Management using Collibra

Aravind Rajagopalan, Rajan Chandras, Manana Gagnidze, Rod Aminian, Meg Ferrauiola, Devin Mann, MD, Jonathan Austrian, MD
Medical Center Information Technology, NYU Langone Health, New York, NY

Problem Description

Clinical Decision Support (CDS) provides clinicians with referential information to make the most informed clinical decisions at the point of care[1]. CDS tools are incorporated into electronic health records (EHRs) in many forms: alerts, order sets, and checklists. To deploy the most up-to-date, accurate and high-performing CDS at the right point of care, healthcare systems should establish governance over the CDS lifecycle — receipt of new requests, management of CDS inventory, and periodic review of CDS tools[2]. Our CDS committee had previously performed these governance functions using a variety of disconnected organizational systems including Excel (database), Trello (tracking governance process), and PowerPoint (project intake). Consequently, the process was inefficient, requiring manual data entry, duplication of effort, and reconciliation among systems. Furthermore, these processes resulted in compromised data integrity and lack of data transparency to the CDS community. These challenges have been described at other institutions[3]. Our CDS governance committee sought to find a novel solution that would streamline CDS lifecycle workflows and overcome these challenges. Ideally, this solution would not be a niche product to CDS and could help solve other health IT challenges to maximize the return on investment.

Solution

We initiated the project to implement Collibra's Data Governance Platform, a web-based software, for CDS in August 2019. The CDS committee collaborated with the software development team to design the database, model the CDS governance workflows within Collibra, and create the integrations with the Epic EMR and Tableau. Ultimately, the initiative went into production in July 2020 and supplied the framework for all CDS stakeholders to perform the CDS lifecycle in one location.

Our revised workflow begins with new CDS alert requests submitted through Collibra. Because Collibra is integrated with our enterprise active directory, all authorized users could submit a request. Once an initial request is completed, the CDS committee reviews the request within Collibra. Collibra provides customized views to CDS requesters and committee members to track progress of individual requests to ensure they are properly completed. A CDS governance member then approves or rejects the request and that decision is automatically communicated via email to the operational owner. The CDS team updates the metadata for the new CDS in Collibra. Following implementation in the EHR, on a nightly basis, key CDS metadata documented in the EHR is uploaded automatically to the corresponding CDS record in Collibra.

CDS chairs can also initiate periodic review of existing CDS alerts within Collibra. The operational owner would receive an email notification to review the alert metadata. From Collibra, they may review an alert's performance by selecting a link that automatically opens that CDS's dashboard in Tableau. To ensure an accurate, up-to-date knowledge base, owners and CDS implementers are responsible for updating important artifacts directly in Collibra such as screenshots of the alert or decision trees. Upon completion, the owner approves the updates in Collibra and concludes the review.

Thus far, Collibra has been successful in mitigating the challenges associated with our previous CDS lifecycle. Since July 2020, we have used Collibra to publish 91 new alerts. We hope to iterate more improvements in the future and translate our success with Collibra Data Governance to manage other knowledge assets at our institution.

References

1. Bates DW, Kuperman GJ, Wang S et al. Ten commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality. JAMIA 2003;10(6):523-530, p. 523.
2. Ash JS, Sittig DF, Dykstra R, Wright A et al. Identifying best practices for clinical decision support and knowledge management in the field. In Medinfo 2010 - Proceedings of the 13th World Congress on Medical Informatics. PART 1 ed. IOS Press. 2010. p. 806-810. (Studies in Health Technology and Informatics; PART 1).
3. Sittig DF, Wright A, Osheroff JA et al. Grand challenges in clinical decision support. Journal of Biomedical Informatics. 2008 Apr 1;41(2):387-392. https://doi.org/10.1016/j.jbi.2007.09.003

System Architecture and Design of a US Department of Veterans Affairs Hospital Information System Integrated System for Patient Safety at Care Transitions

Michele L. Redding, B.S.N.[1], Christina M. Alvaro, B.A.Sc.[1], David V. LaBorde, M.D., M.B.A.[2]
[1]Document Storage Systems, Juno Beach, FL, USA, [2]Iconic Data, Inc., Norcross, GA, USA

Abstract

Care transitions are vulnerable to communication failures. When such failures occur, patients are at risk for potentially preventable medical errors (PMEs). Prospective, multi-center studies have, demonstrated standardized verbal and written handoff reports can reduce PMEs. Herein we describe the architecture of a U.S. Department of Veterans Affairs (VA) hospital information integrated system for improved patient safety at care transitions usable across mobile and desktop devices.

Introduction

Preventable medical errors can cause patient harm and even death. Reported results from prospective, multi-center clinical research have shown that interventions targeting the reduction in communication failures during the transfer of the professional responsibility for patients can decrease PMEs[1]. Herein we report the system architecture of a mobile and desktop device capable VA hospital information system integrated platform for standardized communications during care transitions.

Methods

Requirements were established and a Veterans Health Information Systems and Technology Architecture (VistA) remote procedure call (RPC) application programming interface was developed utilizing the Massachusetts General Hospital Utility Multi-programming System (MUMPS). Additional integration was accomplished via a Health Level Seven International v2.4 (HL7) interface developed using a VA wide integration technology (Integration Framework, Document Storage Systems, Inc., Juno Beach, FL). The integrated solution (*Patient Case Manager, Iconic Data Inc., Norcross, GA*) was then tested in a VistA sandbox environment prior to subsequent testing and deployment in a VA facility.

Results

Provider teams create lists of patients under their care in the solution via automation or manually. Structured data elements including patient demographics, patient location information, visit data, provider information, test results, vitals, code status, orders, clinical charting, and other key data is pulled in real time from VistA. This data is complemented with team annotations entered by providers used for internal team communications. Various projections of this information can be generated by the system. In particular standardized reports to aid discussions at care transitions can be generated and populated with key data pulled in real-time from the system of record. Clinical documentation memorializing the participants in the care transition, when the care transition occurred, and the content communicated during the care transition can be generated automatically and entered into the formal medical record, if desired. Data and care transition reports can be accessed with appropriate credentials via desktop workstations and / or mobile devices on the VA network.

Discussion

The safety of care transitions is improved when commination failures are minimized. In addition, information available at the time of care transitions that participants are not aware of can lead to delayed care and potentially care decisions made without the benefit of new data. The system and architecture herein reported has the potential to reduce the incidence of communication failures and the frequency with which decision making might occur without the most current information. This in turn can potentially reduce PMEs. Future work could aim to understand and quantify the impact of such VA systems.

References

1. Starmer AJ, Landrigan CP, et al. Changes in medical errors with a handoff program. N Engl J Med. 2015;372(5):490-491.

Characterizing Effects of Air Pollution Exposure in Diabetic Patients Affected by COVID19

Naomi O. Riches, PhD, MSPH[1], Ramkiran Gouripeddi, MBBS, MS[1], Willard Dere, MD, FACP[1], Adrian Payan-Medina[1], Julio C. Facelli, PhD, FACMI[1]
[1]University of Utah, Salt Lake City, UT, USA

Introduction

Several risk factors impact COVID-19 severity, such as type 2 diabetes mellitus (T2DM), which worsen outcomes of COVID-19 infection, increasing serious complications and mortality rates.[1] Another such risk factor is long-term exposure to air pollution (AP), which has been found to increase COVID-19 mortality via chronic systemic inflammation and increased susceptibility to the virus. It was found that one $\mu g/m^3$ increase in PM2.5 concentrations increased mortality risk from COVID-19 by 8%.[2] However, to our knowledge, there have not been any studies assessing the combined effect of AP and T2DM on COVID-19 outcomes. Unsupervised machine learning methods, such as k-means clustering, have been used to successfully categorize AP data with health outcomes,[3] and is the method we choose for the current analysis.

Methods

County-level cumulative COVID-19 cases and deaths, from March 11, 2020, when WHO declared COVID-19 a pandemic, to July 24, 2020, were downloaded from Johns Hopkins University Coronavirus Resource Center (https://coronavirus.jhu.edu/us-map). Case-fatality rates were then calculated. T2DM county-level prevalence per 1000 among adults 20 years and older was downloaded from the US Centers for Disease Control and Prevention (CDC, https://www.cdc.gov/diabetes/data/index.html). Daily AP data (NO_2, SO_2, O_3, $PM_{2.5}$, and PM_{10}) was obtained from the US Environmental Protection Agency (EPA, https://aqs.epa.gov/aqsweb/airdata/download_files.html), and a 5-year mean level was calculated for each US county. All variables were matched by county. Where multiple EPA monitoring stations existed per county, an average was used. The number (k) of the k-means clustering was unknown prior to analysis. Therefore, the Elbow method combined with Davies-Bouldin indexing was used establish the k which maximizes the Euclidian distance between clusters and minimizes the distance within clusters.

Table 1. Mean Covid-19 case-fatality rates, diabetes prevalence, and air pollution components by cluster

cluster	(n) per cluster	Case-fatality	Diabetes prevalence	Population size	ozone (ppb)	PM2.5 (ug/m3)	SO2 (ppb)	NO2 (ppb)	PM10 (ug/m3)
2	1654	0.01	7.42	47350	38.96	3.85	0.60	2.88	12.03
7	544	0.01	8.80	271240	31.88	6.78	0.21	3.04	21.77
9	2312	0.02	8.82	114104	32.32	4.84	0.73	2.55	11.20
5	3768	0.03	8.32	611837	29.40	7.17	0.33	6.18	14.03
3	1798	0.03	7.70	1633961	36.07	7.71	0.40	11.18	21.56
4	1088	0.03	8.15	2493009	29.51	9.77	0.84	14.95	28.83
6	680	0.04	7.56	1702061	26.99	8.11	0.35	14.75	18.40
8	2564	0.04	9.66	1353007	29.75	8.42	0.42	11.48	17.01
1	4088	0.04	10.44	741043	28.23	9.13	0.86	7.73	18.42

Results

Nine clusters resulted from using k-means for COVID-19 case-fatality rates, diabetes prevalence, and concentrations of air pollution constituents at the county level. Mean values for each cluster are seen in Table 1.

Conclusion

Case-fatality rates were positively related to T2DM, $PM_{2.5}$, and NO_2, but negatively related to ozone. k-means clustering proved to be a useful tool in assessing this multidimensional data. Future research will include additional variables that explore how social determinants of health impact clustering outcomes.

References

1. Guo W, Li M, Dong Y, et al. Diabetes is a risk factor for the progression and prognosis of COVID-19. *Diabetes/Metabolism Research and Reviews.* 2020;n/a(n/a):e3319.
2. Wu X, Nethery RC, Sabath BM, Braun D, Dominici F. Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study. *medRxiv.* 2020:2020.2004.2005.20054502.
3. Kioumourtzoglou MA, Austin E, Koutrakis P, Dominici F, Schwartz J, Zanobetti A. PM2.5 and survival among older adults: effect modification by particulate composition. *Epidemiology (Cambridge, Mass).* 2015;26(3):321-327.

Comparison of Algorithms for Identifying HIV-Positive Individuals from Electronic Medical Records in a Large Multi-site Database

Jessica P. Ridgway, MD[1], Junlan Zhou, MS[1], Eleanor Friedman, PhD[1], John Schneider, MD[1]
[1]University of Chicago, Chicago, IL

Introduction

As electronic medical record (EMR) data are increasingly used in clinical research among HIV-positive individuals, accurately identifying HIV-positive patients from EMR data is paramount. Others have developed EMR algorithms for identifying HIV-positive patients at single centers,[1,2] but these algorithms have not been validated externally. The objective of this study was to develop and compare EMR algorithms to identify HIV-positive individuals in a multi-center EMR database.

Methods

We collected deidentified EMR data from the Chicago Area Patient-Centered Outcomes Research Network (CAPriCORN) for all patients with either a diagnosis code (ICD9 or 10) for HIV or an HIV viral load test result. CAPriCORN is a clinical data research network with linked data from EMRs for over 10 million patients from diverse healthcare systems in Chicago, including academic medical centers, community hospitals, and clinics.[3] For each patient, we collected demographics, encounters, diagnosis codes, antiretroviral therapy (ART) regimens, and laboratory test results. We compared algorithms for identifying HIV-positive individuals.

Results

40,718 patients from 7 healthcare systems were included. 13,280 patients had a positive laboratory test for HIV (i.e., confirmatory HIV antibody, p24 antigen, or HIV viral load > 20 copies/mL). 24,041 patients had at least one encounter with a diagnosis code for HIV, and 17,787 patients were prescribed ART, excluding ART regimens used for Hepatitis B or pre-exposure prophylaxis (Figure 1). Table 1 describes the number of patients identified by algorithms combining laboratory results, diagnosis codes, and ART prescriptions.

Figure 1: Venn Diagram of Patients Identified as HIV-Positive from Different EMR Data Types

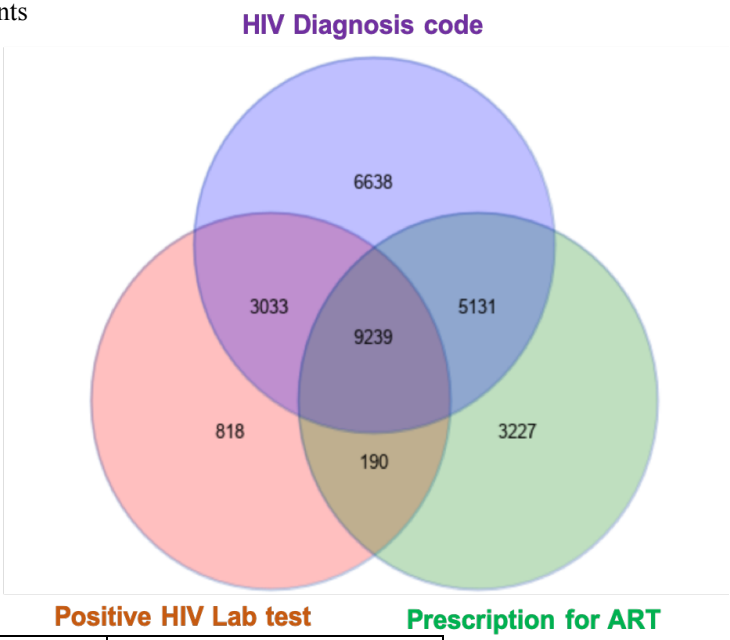


Table 1: EMR Algorithms for Identifying HIV-Positive Patients

Algorithm	Number of patients
Positive HIV Laboratory test OR Diagnosis code for HIV AND receiving ART	18,411
Positive HIV Laboratory test OR Diagnosis code for HIV AND at least 2 viral load tests performed	18,684
Positive HIV Laboratory test OR Diagnosis code for HIV AND at least 2 viral load tests performed OR Diagnosis code for HIV AND Receiving ART	20,937

Conclusion

EMR algorithms that combine laboratory results, administrative data, and ART prescriptions detect more HIV-positive patients in a large multisite EMR database than use of a single data type alone.

References

1. Paul DW, Neely NB, Clement M, et al. Development and validation of an electronic medical record (EMR)-based computed phenotype of HIV-1 infection. *J Am Med Inform Assoc.* 2018;25(2):150-157.
2. Goetz MB, Hoang T, Kan VL, Rimland D, Rodriguez-Barradas M. Development and validation of an algorithm to identify patients newly diagnosed with HIV infection from electronic health records. *AIDS Res Hum Retroviruses.* 2014;30(7):626-633.
3. Kho AN, Hynes DM, Goel S, et al. CAPriCORN: Chicago Area Patient-Centered Outcomes Research Network. *J Am Med Inform Assoc.* 2014;21(4):607-611.

Usability and Impact of a Care-Management System in Holistically Delivering Services to Vulnerable Populations

Rubina F. Rizvi, MD, PhD[1], Tiffani J. Bright, PhD[1], Courtney VanHouten, MA[1], Mollie McKillop, PhD, MPH[1], Suwei Wang, PhD, MS[1], Shira Alevy, EdM[1], David Brotman, MS[1], Megan Sands-Lincoln, PhD, MPH[1], Barbie Robinson[3], Carolyn Staats[3], Gretchen P. Jackson, MD, PhD[1,2] William J. Kassler, MD, MPH[1]

[1]IBM Watson Health, Cambridge, MA, USA, [2]Vanderbilt University Medical Center, Nashville, Tennessee, USA, [3]County of Sonoma, CA, USA

Introduction

Providing comprehensive services to vulnerable populations is complex and requires effective and efficient collaboration across various safety net agencies. In 2018, Sonoma County implemented the *Accessing Coordinated Care and Empowering Self Sufficiency* (ACCESS) initiative to help residents receive well-coordinated care.[1] An interdepartmental multi-disciplinary team (IMDT) was established to streamline care delivery through an integrated care-management and coordination system comprised of Connect 360, a data hub, and Watson Care Manager (WCM), a cloud-based platform to present data from multiple social-service sources.[2] This study explored and analyzed the usability and impact of WCM's usage on care coordination from the end-users' perspectives.

Methods

In this mixed-methods study, 3 weeks of data collection took place shortly after the 2019 Kincaid Fire, a crisis that exacerbated the needs of vulnerable populations in Sonoma County. IMDT meetings were observed, and WCM end users were surveyed and later interviewed by four experts (RR, CVH, MSL, TB). The Technology Acceptance Model (TAM)[3] was administered to assess WCM's perceived usefulness (PU) and perceived ease of use (PEOU). Survey data were summarized with descriptive statistics; interview and observational data were analyzed qualitatively using thematic analysis, directed by a grounded theory approach.

Results

Approximately nine hours of IMDT meeting observations were conducted, and eight WCM end users were interviewed. Participants were mostly female (n=6), between 31 to 62 years of age, with undergraduate (n=4) or graduate-level (n=4) degrees. The reported level of interaction with WCM varied across end users, most having an intermediate technology skill level. TAM response details are provided in Figure 1. The median TAM score for PU, PEOU and combined PU+PEOU was 5 (range: 2-7), measured on a 7-point Likert scale (1=extremely disagree and 7=extremely agree). No association was observed between TAM scores and level of interaction, education and self-rated technology skills. From observational and interview data, three principal themes were identified for the role of WCM: supporting data integration and collaboration, facilitating data sharing and reporting, and helping to efficiently and effectively execute tasks.

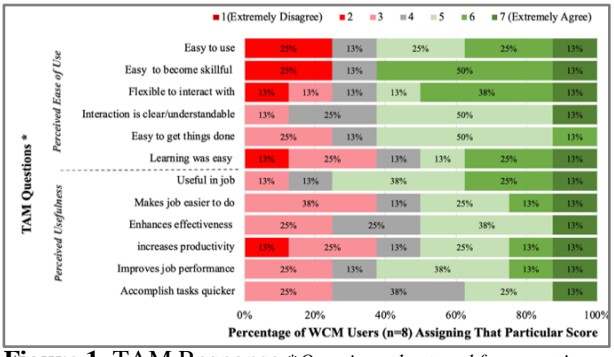


Figure 1. TAM Response **Questions shortened for reporting.*

Conclusion

WCM serves as an essential anchoring tool that facilitates IMDT meetings by bringing together previously siloed data into a single, shared, and secured platform. TAM scores suggested general positive perceived usability and impact. Further research is needed to measure population impact of the adoption of this tool.

References

1. Robinson B, Staats C. Access Sonoma: enabling multi-discipline teams to deliver safety net services to county residents with complex needs. Think 2019; San Francisco, CA.
2. Snowdon J, Robinson B, Staats C, Wolsey K, Strasheim T, Kassler W, et al. Empowering caseworkers to better serve the most vulnerable with a cloud-based care management solution. Applied Clinical Informatics. 2020.
3. Lewis JR. Comparison of four TAM item formats: effect of response option labels and order. Journal of Usability Studies. 2019;14 (4).

A COVID-19 Drug Repurposing Network from Integrated Text Mining and Semantic Data Mining

Karen E. Ross, PhD[1], Chuming Chen, PhD[2], Julie Cowart, MS, PSM[2], Sachin Gavali, BDS[2], and Cathy H. Wu, PhD[1,2]

[1]Georgetown Univ. Medical Center, Washington, DC; [2]Univ. of Delaware, Newark, DE

Background: As medical researchers respond to the COVID-19 global health emergency, the COVID-19 literature is rapidly expanding (~80,000 articles returned in a PubMed search for "COVID" in December 2020). Computational approaches that can automatically distill key information from these articles and integrate it with relevant information from curated biological databases are essential to gain insight into COVID-19 etiology, diagnosis and treatment. Knowledge graphs (networks) are a powerful way to represent such diverse biological information and generate novel hypotheses. Several efforts are underway to investigate COVID-19 using knowledge graphs, including the COVID-19 Disease Map (PMID:32371892), INDRA (PMID:29175850) and KG-COVID-19 (PMID:32839776). In this study, we constructed a COVID-19 drug repurposing knowledge graph based on mining of literature and databases, taking advantage of semantic web technologies (RDF and SPARQL) to streamline data integration.

Methods: Entities of interest (proteins, variants, diseases, drugs) mentioned in the COVID-19 Open Research Dataset literature (CORD-19; https://arxiv.org/abs/2004.10706) were annotated using PubTator (PMID:31114887), converted into RDF triples and stored in a Virtuoso triplestore. Drug, molecular target, and therapeutic indication information from DrugBank (PMID:29126136) was also converted into RDF triples and added to the triplestore. We used federated SPARQL queries across several endpoints such as UniProt (PMID:29425356), Protein Ontology (PMID:27899649) and DisGeNET (PMID:27153650) to retrieve and integrate relations among genes, genetic variants, and diseases. We queried STRING (PMID:30476243) for protein-protein interactions, iTextMine (PMID:30576489) for text-mined relations among genes, variants, and effects on disease/drug response, and the COVID-19 Therapeutic Information Browser (TIB; https://covidtib.c19hcc.org/) for drug-COVID-19 co-mentions in the literature using their respective web interfaces. Custom scripts were used to combine relations from different sources into node/interaction and node property files for visualization in Cytoscape.

Results: Using the most frequently mentioned eight proteins and nine genetic variants in the CORD-19 corpus as seeds, we constructed a network with 15 proteins, 93 variants, 139 diseases, 32 drugs, and 7 disease outcome/drug response terms, connected by ~490 relations of the following types: protein-variant, variant-disease, disease-drug, disease-outcome, drug-protein (target), drug-drug response, and protein-protein (interaction). The network provides a wealth of information about drug, gene, and disease relationships in the context of COVID-19. For example, Tumor Necrosis Factor (TNF; Figure 1) is implicated in the hyper-inflammatory response, known as the "cytokine storm", seen in some patients with severe COVID-19. TNF is included in the network because two of the most frequently occurring genetic variants in the CORD-19 corpus (rs1800629, rs361525) lie in the regulatory region upstream of the TNF gene. Of the drugs that target TNF, several (infliximab, golimumab, and adalimumab) are frequently mentioned in the CORD-19 corpus (Figure 1, drugs with red borders). Moreover, one of these drugs, infliximab, is being tested as a potential COVID-19 therapy in a clinical trial. However, other TNF-targeted drugs (etanercept, certolizumab pegol) are not mentioned. All of these drugs are recommended for use in a similar set of auto-immune diseases (blue nodes), suggesting they may have similar immune modulatory effects. Thus, based on our network, etanercept and certolizumab pegol may be drugs of interest to investigate in the context of COVID-19. This work has demonstrated the value of integration and knowledge graph representation of biological data to gain insight into a rapidly evolving biomedical problem. Network construction is highly automated to facilitate rapid updates as new information is published. Our future plans include systematic evaluation of the network to identify other drugs that have properties in common (e.g., molecular target, therapeutic indications) with promising anti-COVID-19 drugs and dissemination of the network through mechanisms such as an API and SPARQL endpoint.

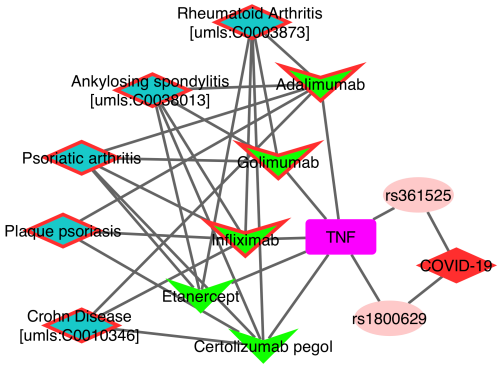


Figure 1. TNF-associated drugs, variants, and diseases in the COVID-19 network.

Validating synthetic data derivatives using adapted inpatient cascade of care for Type II diabetes mellitus

Irene Ryan[1], Dr. Cynthia Herrick [2], MD, MPHS, Dr. Randi Foraker, PhD, MA, FAHA, FAMIA[1]
[1] Center for Population Health, Institute for Informatics, St. Louis, Missouri; [2] Department of Medicine, Washington University School of Medicine, St. Louis, Missouri

Introduction: More than 30 million Americans have Type I or Type II diabetes, and recent studies show population-level diabetes care hasn't improved over the last decade.[1] Synthetic data may allow clinicians and informaticians to evaluate, in real-time, the "cascade of care" of diabetes. The "cascade of care" framework defines diagnosis, linkage to care, and HbA1c control for population-level diabetes monitoring. This study uses the data synthesis platform MDClone (MDClone Ltd., Beer Sheva, Israel) that uses electronic health record (EHR) data to derive a statistically equivalent data set to the original while ensuring data privacy. Values that are extreme or can be identifiable are censored from the data during the synthesis process.[2]

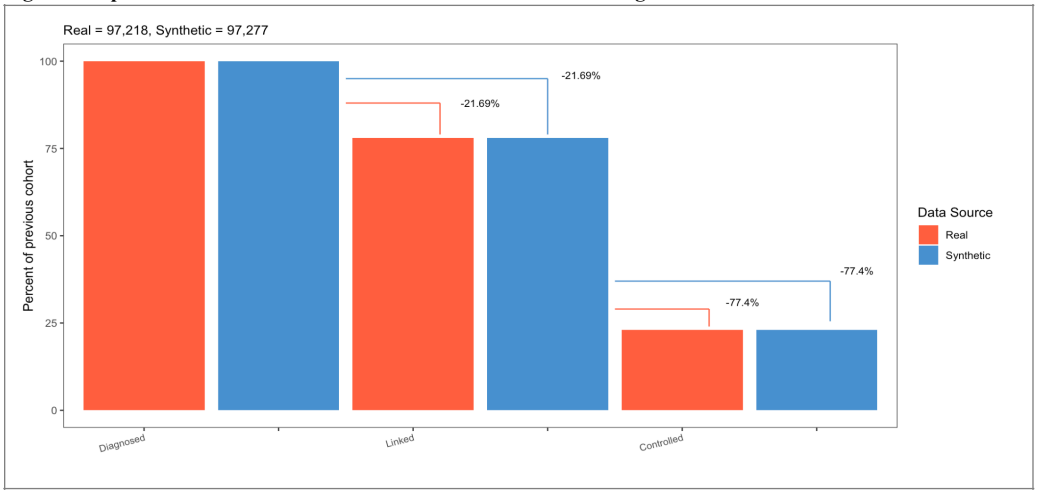
Methods: We adapted the outpatient cascade of care[1] to define an inpatient cascade of care using American Diabetes Association guidelines[3] and consultations with an endocrinologist (Table1). We then constructed the cascade of care using original and synthetic data from the EHR of patients admitted within a day of their diabetes diagnosis to the BJC healthcare system in St. Louis, MO from 2010 to 2019. We defined linkage to care as meeting at least one of the following: insulin administered within 24 hours of admission, HbA1c measured within 24 hours of admission, a consult with a diabetes or endocrinology specialist during or after admission, and/or oral medications prescribed within 3 months of discharge. Patients had HbA1c control if they had at least 1 measure of HbA1c <7%. We compared the proportion of patients who met the criteria for the cascade, as well as the proportion of the cohort lost at each stage, between the original and synthetic data. We then computed 95% confidence intervals to assess for the precision of the estimates. We quantified the distribution of the cohort by sex, race, and age at each stage in the cascade to capture any disparities among those lost along the cascade.

Results:

Table 1: Cascade of care stage data and population demographics (Synthetic and Real data)

Variable	Synthetic (N, %; median ± SD)	Censored (N, %)	Original (N, %; median ± SD)
Diagnosed (w/o complication)	63862 (66%)	395 (<1%)	63993 (66%)
Linked with care HbA1c taken < 24 hours of admission Insulin < 24 hours of admission Oral meds < 3 months of discharge Diabetes or endocrinology consult during	 26166 (27%) 67279 (69%) 21558 (22%) 5966 (6%)	 15 (<1%) 15 (<1%) 34 (<1%) 0 (0%) (<1%)	 26172 (27%) 67276 (69%) 21577 (22%) 5966 (6%)
Controlled HbA1c (<7%)	17204 (18%)	0 (0%)	17203 (18%)
Sex (Female)	46402 (48%)	3436 (4%)	48147 (50%)
Race (White)	68757 (71%)	1108 (1%)	69128 (71%)
Age at diagnosis	65.97 ± 14.16	0 (0%)	65.94 ± 14.16
Length of stay	2.94 ± 6.48	0 (0%)	2.93 ± 6.48
Cardiovascular disease (MI, Stroke, CHF)	31818 (33%)	122 (<1%)	31864 (33%)

Figure 1. Inpatient diabetes cascade of care with % lost from each stage



Conclusion: The results of the synthetic and original data by proportion of patients maintained in the cascade, as well as demographic breakdowns, were sufficiently similar (Table 1), aside from slight variations in sex, which indicated that synthetic data created by this platform will be useful for population-level descriptions and analyses such as quantifying the diabetes cascade of care. This is promising, as synthetic data are easily obtainable, and their use requires no IRB approval. With synthetic data, researchers can think of a question they want to answer, and immediately query the data to investigate their problem.

References

1. Kazemian, P., Shebl, F.M., Mccann, N., Walensky, R.P., Wexler, D.J., 2019. Evaluation of the Cascade of Diabetes Care in the United States, 2005-2016. JAMA Internal Medicine. doi:10.1001/jamainternmed.2019.2396
2. Foraker RE, Yu SC, Gupta A, et al.. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open*. 2020. doi:10.1093/jamiaopen/ooaa060.
3. 2019. Introduction: Standards of Medical Care in Diabetes—2019. Diabetes Care. doi:10.2337/dc19-sint01

Predicting Long-Term Care Demand from Medical and Long-Term Care Insurance Claims in Japan

Jumpei Sato, MS[1], Kazuo Goda, PhD[1], Masaru Kitsuregawa, PhD[1], Tomoki Ishikawa, PhD[2], Naohiro Mitsutake, PhD[2]
[1]The University of Tokyo, Tokyo; [2]Institute for Health Economics and Policy, Tokyo

Introduction

Driven by the rapid aging of the population, Japan introduced public long-term care insurance to reinforce healthcare services for the elderly in 2000. The public long-term care insurance has become a pillar supporting Japan's healthcare system, along with the public medical insurance. The public long-term care insurance groups all individuals aged 65 years and older into eight eligibility levels in accordance with their need for long-term care. The eligibility levels include RLTC5 (having most severe health issues and most intensive care demand) through RS1 (having minor health problems), and Ineligible (being healthy). Individuals of a certain eligibility level are provided with long-term care services according to their approved eligibility level. Precisely predicting individual-level future long-term care demand helps regional healthcare authorities to plan and manage healthcare resources in the region and suggest preventive care to their elderly citizens. In this study, we have explored an effective method for predicting individual-level future long-term care demand from past insurance claims of medical and long-term care services.

Methods

We employed a supervised machine learning approach for predicting future long-term care demand. First, we designed discriminative models that input past insurance claims and output a future eligibility level for each individual. Second, we trained the designed models using the gradient boosting decision tree learning algorithm with the training dataset. Third, we applied the trained models to separate the test dataset to evaluate the effectiveness of the models. In addition to the *multiclass classification*[1] approach, widely used in prediction modeling, we also explored another approach, *class-specific classification*[2]. In the *multiclass classification* approach, the training dataset was directly employed to build a single predictor. In the *class-specific classification* approach, the training dataset was divided into disjoint subsets based on the individuals' current eligibility levels, and a separate predictor was built for each subset. For the model evaluation, we used three evaluation indicators: a weighted average of Precision, Recall, and F-measure (the harmonic mean of Precision and Recall)[3]. For model training and model validation, we utilized the three-year dataset of medical and long-term insurance claims and enrollment records, which were provided by 170 regional public insurance operators covering 42 local government areas in Japan.

Results

We trained the models using insurance claims recorded from April 2016 through March 2017 and evaluated the prediction accuracy of eligibility levels of three / six / twelve months later (i.e., June 2017 / September 2017 / March 2018). The prediction model based on *multiclass classification* achieved practically high accuracy up to twelve months later prediction (Precision: 0.949, 0.923 and 0.872, Recall: 0.950, 0.926, and 0.878, F-measure: 0.949, 0.924, and 0.873 for three, six, and twelve months later prediction, respectively). The model based on *class-specific classification* showed relatively lower accuracy (Precision: 0.866, 0.837 and 0.785, Recall: 0.842, 0.789, and 0.758, F-measure: 0.849, 0.806, and 0.769 for three, six, and twelve months later prediction, respectively).

Conclusion

The results show that the developed prediction models offered a practically high accuracy for the prediction of individual-level future long-term care demand. In the future study, we will extend the training and validation datasets and refine learning algorithms/modeling approaches to further improve the prediction method.

References

1. Aly M. Survey on multiclass classification methods. Neural Netw. 2005;19:1-9.
2. Raitoharju J, Kiranyaz S, Gabbouj M. Training radial basis function neural networks for classification via class-specific clustering. IEEE T Neur Net Lear. 2015;27:2458-71.
3. Behera B, Kumaravelan G. Performance evaluation of deep learning algorithms in biomedical document classification. 11th ICoAC. 2019;220-24.

The COVID-19 DREAM Challenge: enabling continuous benchmarking of models on EHR data

Thomas Schaffter, PhD [1,*] , Timothy Bergquist [2,*] , Yao Yan [3] , Thomas Yu [1] , Yooree Chae [1] , Micheal Mason, PhD [1] , Justin Prosser [4] , Sean Mooney, PhD [2] , Justin Guinney, PhD [1,2]

[1] Sage Bionetwork, Seattle, WA 98121, USA [2] Biomedical Informatics and Medical Education, University of Washington [3] Molecular Engineering & Sciences Institute, University of Washington [4] Institute for Translational Health Sciences, University of Washington

* Are contributing equally

The rapid rise of COVID-19 has challenged healthcare globally. The underlying risks and outcomes of infection are still incompletely characterized even as the world surpasses 25 million infections. Due to the importance and emergent need for better understanding of the condition and the development of patient specific clinical risk scores and early warning tools, we have developed a platform for the COVID-19 DREAM Challenge to support testing analytic and machine learning hypotheses on clinical data without data sharing as a platform to rapidly discover and implement approaches for care. We have previously applied this approach in the successful EHR DREAM Challenge focusing on Patient Mortality Prediction with UW Medicine.

We have the goal of incorporating machine learning and predictive algorithms into clinical care and COVID-19 is an important and highly urgent challenge. In our first iteration, we are facilitating understanding risk factors that lead to a positive test utilizing electronic health recorded data mapped to the OMOP Common Data Model. Data scientists who have built these models are able to build and submit their prediction models to these challenges without access to the data (1). As other sites come online, these models will be distributed to many healthcare institutions across the country who have been on-boarded into the evaluation and benchmarking network (2) where they will be trained and evaluated against the host site's EHR repository (3). The results from each of the sites are returned to the Synapse leaderboard (4) for review by the model submitter (5).

Typically, evaluation of model performance is carried out using cross validation or testing on a hold out set. While these can be useful model validation methods, they do not guarantee that COVID-19 prediction algorithms will generalize to prospective prediction circumstances. Prospective evaluation of multiple competing models, predicting on the same dataset for the same outcome, can remove many of the biases that exist in model validation, and offers a more objective assessment of model performance and generalizability. Here, we report on the performance of the models submitted by 371 participants (and counting) to the different challenge questions that we have identified. We also provide insights into the features used by these models and how well their performance generalizes across datasets from different healthcare institutions.

Acknowledgments: The work presented in this panel reflects the collaboration of many individuals from across the N3C, CTSAs, NCATS, Observational Health Data and Informatics (OHDSI), and the many organizations and companies whose members provided ongoing input, support, and participation. This work has been funded through the National Center for Advancing Translational Sciences, National Institutes of Health, under award number U24 TR002306.

Design of a Clinical Decision Support Tool for Erroneously Documented Vital Signs

Connor Skeehan[1,2], Benjamin H. Slovis, MD, MA[3,4], Melanie McArthur, MSN, RN[1], Jeffrey Riggio MD [3,5]

[1] Information Services and Technology, Thomas Jefferson University, Philadelphia PA
[2] Population Health Intelligence Master's Program (expected 2021), Thomas Jefferson University, Philadelphia PA
[3] Office of the Chief Medical Information Officer, Thomas Jefferson University, Philadelphia PA
[4] Department of Emergency Medicine, Thomas Jefferson University, Philadelphia PA
[5] Department of Internal Medicine, Thomas Jefferson University, Philadelphia PA

Abstract: Electronic Medical Records (EHRs) are associated with decreased vital sign documentation errors.[1] Accurate EHR vital sign data is critical for other clinical decision support systems (CDSS). However, it is still possible for nurses to erroneously document vital signs which can lead to clinical mistakes and poor patient outcomes.[2] The purpose of this study was to leverage the CDSS functionality of our EHR to determine a suitable % change threshold for when a documented vital sign was likely entered in error. Once established we will implement a CDSS to alert clinicians when a suspicious vital sign flowsheet entry has been made and study its effect.

Background: Jefferson Health in Philadelphia PA employs Epic (Epic Systems, Verona WI) as its EHR. The system allows for both automated upload and manually entered vital signs but only alerts providers of potentially erroneously entered weight measurements (based on a percent change). Other vital signs have no alert notifications for extreme changes. The purpose of this study was to determine an appropriate threshold for vital sign decision support tools. Notifying nurses of potentially erroneously entered vital signs may reduce errors and improve quality of care.

Methods: We previously implemented two silent CDSS for Heart Rate (HR) data entry for all inpatient admissions. When a HR entry of at least 75% higher or 50% lower than the previous entry is entered the documentation event is silently recorded. These initial percent thresholds were chosen by clinical informaticists as appropriate starting points for our study. At the time of data entry nurses do have the ability to correct the erroneous HR value by entering either an additional value (thus leaving the error) or correcting the value (overwriting it), but at present are not alerted to the possible error and must recognize it themselves. We currently can distinguish between these two correction methods but can't determine the clinical scenario that led to it. We retrospectively collected all potentially erroneous HR values (HR0) entered for a 30-month period, the values recorded before (HR-1) and after (HR+1), as well as if an entry was corrected or not. The potentially erroneous entries were identified based on the 75% increase or 50% decrease criteria configured in each CDSS. We calculated descriptive statistics for each distribution. We then created two cohorts for each triggering event. The first cohort identified highly likely documentation errors where HR0 was corrected by the nurse at entry or the Δ% from HR0 to HR+1 was documented within one hour and met our previously described threshold criteria (i.e. HR-1 to HR0 had a 75% increase, then HR0 to HR+1 had a 50% decrease or vise-versa within 24 hours). The second cohort included events where the Δ% for HR0 to HR+1 did not meet the threshold for change, implying the data entry was likely accurate. We performed a logistic regression analysis to determine the relationship between Δ% and odds of presumed erroneous entries, thus determining the specific threshold that resulted in greater than 0.5 odds of the CDSS, indicating a likely error.

Results: Between 3/1/2018 and 8/9/2020 our cohort of highly likely errors contained 672 potentially erroneously entered increases in HR and 1,927 potentially erroneously entered decreases in HR in the increase and decrease CDSS, respectively. The median increase in HR was 89.4% (IQR = 80.7%-106.0%) and the mean decrease in HR was 74.1% (SD = 19.0). 9.2% of the increased HR entries were categorized as errors and 49.1% of the decrease entries were categorized as errors. Logistic regression analysis for the increased HR was not significant ($p = 0.083$) but was for the decreased HR (OR = 1.022, $p<0.0001$).

Conclusion: Our logistic regression demonstrated that for every 1% change in the decrease model the odds of an error increased by 1.022. Setting our model at 50% probability of error results in a Δ% of 75.9%, therefore we will set 75% as our new threshold for the decrease CDSS. We will maintain 75% in the increase CDSS. We plan to assess our other vital sign CDSS in a similar manner and establish appropriate error thresholds, and hope to validate our thresholds. We then plan to consult with nursing leadership and implement these alerts to notify nurses of possibly erroneous data entry and examine in effects. Accurate data entry could improve functionality of other CDSS reliant upon vital sign data.

1. Gearing P, Olney CM, Davis K, Lozano D, Smith LB, Friedman B. Enhancing patient safety through electronic medical record documentation of vital signs. J Healthc Inf Manag 2006;**20**(4):40-5
2. Badawy J, Nguyen OK, Clark C, Halm EA, Makam AN. Is everyone really breathing 20 times a minute? Assessing epidemiology and variation in recorded respiratory rate in hospitalised adults. BMJ Qual Saf 2017;**26**(10):832-36 doi: 10.1136/bmjqs-2017-006671[published Online First: Epub Date]|.

Deconvoluting Spatial Transcriptomics Data through Graph-based Artificial Intelligence

Qianqian Song[1], Jing Su[2,3]

[1]Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC, USA; [2]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, USA; [3]Section on Gerontology and Geriatric Medicine, Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA

Introduction. Emerging spatial transcriptomics technologies are able to spatially index transcripts and measure expression profiles, advancing our understanding of precise tissue architectures. However, the resolution of ST data is far lower than single cell level. Transcripts captured at a specific location by a "spot" or a "bead" is usually composed of a mixture of heterogenous cells. For example, Visium, one of the microarray-based spatial transcriptomics techniques developed by 10X Genomics, uses spots of 50 μm diameter, with each spot covering 10-20 cells in average, which varies depending on the tissue histology. Therefore, uncovering the cell compositions within each spot of the spatial transcriptomics data is critical for investigating tissue's molecular and cellular architecture at high resolution. The fast-growing public repository of single-cell RNA sequencing (scRNA-seq) data provides valuable resources for deconvoluting the spatial transcriptomics data. Through integrating existing, well-characterized scRNA-seq data with newly generated spatial transcriptomics data, knowledge learned from previous studies at single-cell level is able to provide insights into spatial data.

Methods and Results. To deconvolute the spatial transcriptomics data, we have proposed a novel, graph-based artificial intelligence approach, Deconvoluting Spatial Transcriptomics data through Graph-based convolutional network (DSTG), for reliable and reproducible identification of cell compositions in the spot-based spatial transcriptomics data. The DSTG approach leverages scRNA-seq data to unveil the cell mixtures in the spatial transcriptomics (ST) data (Figure 1). Our hypothesis is that the captured gene expression on a spot is contributed by a mixture of cells located on that spot. Our strategy is to use the scRNA-seq-derived synthetic spatial transcriptomics data called "pseudo-ST", to predict cell compositions in real-ST data through semi-supervised learning. First, DSTG constructs the synthetic pseudo-ST data from scRNA-seq data as the learning basis of our method. Then DSTG learns a link graph of spot mapping across the pseudo-ST data and real-ST data using shared nearest neighbors. The link graph captures the intrinsic topological similarity between spots and incorporate the pseudo-ST and real-ST data into the same graph for learning. Then, based on the link graph, DSTG is used to learn a latent representation of both local graph structure and gene expression patterns that can explain the various cell compositions at spots. The major advantages of such similarity-based semi-supervised GCN model are: 1) sensitive and efficient, since for each spot, only the features of similar spots (i.e., neighbour nodes) are used; and 2) acquiring generalizable knowledge about the association between gene expression patterns and cell compositions across spots in both pseudo- and real-ST, since the weight parameters in the convolution kernel are shared by all spots. DSTG is validated and applied to real tissue context from mouse cortex, hippocampus, and pancreatic tissues with well-defined structures. Based on the well-characterized scRNA-seq dataset, DSTG is able to learn the precise composition of spatial transcriptomics data.

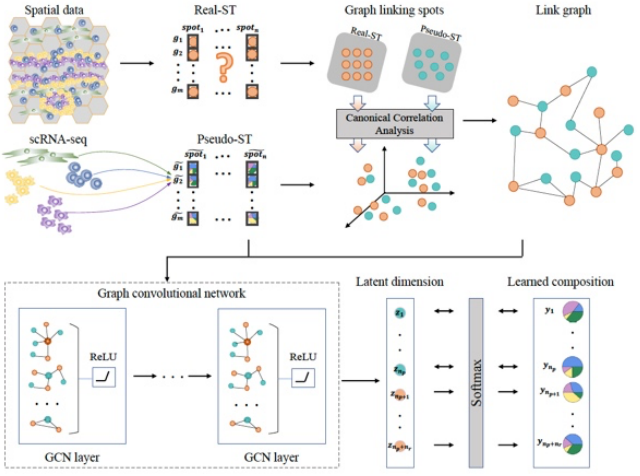


Figure 1. Schematic overview of DSTG

Conclusion. In summary, our study has established a novel computational method that accurately and precisely identify the precise cell-type composition in spatial transcriptomics data. DSTG guides the spatial mapping that will provide deep insights and elucidation of tissue architecture, cell-cell interaction, spatial expression of disease related genes. DSTG is available as an open source software that can be readily used to facilitate the consistency and reproducibility of composition mapping across different studies.

Toward a Neural Semantic Parsing System for EHR Question Answering

Sarvesh Soni, MS, Kirk Roberts, PhD
School of Biomedical Informatics, UTHealth, Houston, Texas

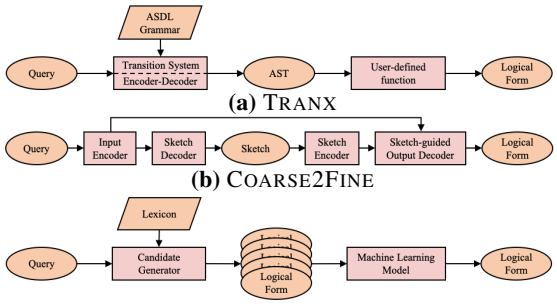
Introduction

Question answering (QA) systems provide a natural way to express an information need in the form of natural language. One of the approaches to interpret and tackle the information need is known as semantic parsing (SP) where the inherently ambiguous input query is mapped to an unambiguous machine-understandable logical form (LF). Most current approaches to clinical SP have been centered around rule-based techniques and traditional machine learning (ML). Recently, many neural SP techniques have emerged that achieve comparable levels of performance as the traditional techniques while overcoming the challenges involved in building traditional systems. In this paper, we systematically assess the efficacy of neural SP when applied to the task of electronic health record (EHR) QA.

Materials and Methods

Data We use two clinical QA datasets for our evaluations, namely, ICU_{DATA} [1] and $\text{FHIR}_{\text{DATA}}$ [2]. Both the datasets consist of patient-specific clinical questions (that can be answered using EHR data) and their corresponding LFs based on λ-calculus. We also report the performance of the evaluated neural models on a general domain dataset, i.e., ATIS [3].

Models We use two neural SP models based on different architectures that are shown to be effective for small datasets, namely, TRANX [4] and COARSE2FINE [5]. To compare the performance in light of a traditional lexicon-based approach, we also report the results of two traditional rule-based and ML approaches for all the evaluated datasets. We collectively refer to these traditional methods as LEXICON-BASED as both of them employ some kind of lexicon to map the natural language phrases from the input utterances to logical predicates present in the λ-calculus (Figure 1).



(a) TRANX

(b) COARSE2FINE

(c) LEXICON-BASED

Figure 1: High-level architectures of the evaluated models.

Table 1: Cross-validation (CV) evaluation.

Model	Corpus		
	ICU_{DATA}	$\text{FHIR}_{\text{DATA}}$	ATIS
TRANX	75.7	91.6	86.2
COARSE2FINE	79.8	94.2	87.7
LEXICON-BASED	95.6	94.2	91.3

Table 2: Cross-dataset evaluation.

Model	Corpus (to test)	
	ICU_{DATA}	$\text{FHIR}_{\text{DATA}}$
TRANX	64.8	66.1
COARSE2FINE	67.4	72.4

Results and Discussion

The performance of the evaluated models are presented in Tables 1 and 2. In the case of CV, both the neural models achieve a lower accuracy on the ICU_{DATA} than that on the $\text{FHIR}_{\text{DATA}}$. This can be related to the higher diversity of logical predicates present in the ICU_{DATA}. The cross-dataset results show that both the models suffered a performance drop when trained on a different dataset and tested on another. These findings highlight the requirement of a generalizable clinical SP dataset. The performance disparity of the neural-based methods between different clinical datasets can be attributed to the diversity present in different corpora. Our error analysis surfaces the most frequent types of errors made by neural models on clinical datasets that, along with our detailed discussion to tackle these types of errors, can serve as a starting point for future research in this domain.

References

[1] Roberts K, Demner-Fushman D. Annotating Logical Forms for EHR Questions. In: LREC; 2016. p. 3772–3778.
[2] Soni S, Gudala M, Wang DZ, et al. Using FHIR to Construct a Corpus of Clinical Questions Annotated with Logical Forms and Answers. In: AMIA Annual Symposium Proceedings. vol. 2019; 2019. p. 1207–1215.
[3] Zettlemoyer L, Collins M. Online Learning of Relaxed CCG Grammars for Parsing to Logical Form. In: EMNLP-CoNLL; 2007. p. 678–687.
[4] Yin P, Neubig G. TRANX: A Transition-Based Neural Abstract Syntax Parser for Semantic Parsing and Code Generation. In: EMNLP: System Demonstrations; 2018. p. 7–12.
[5] Dong L, Lapata M. Coarse-to-Fine Decoding for Neural Semantic Parsing. In: ACL; 2018. p. 731–742.

Query Complexity Analysis on the TriNetX Platform

Ezra Statsky-Frank[1,2], Zuzanna Drebert, PhD[2], Jordan Donovan[2], John E. Doole, Pharm.D., MFA[2], Matvey B. Palchuk, MD, MS, FAMIA[2]
[1]Northeastern University, Boston, MA, [2]TriNetX, Inc., Cambridge, MA

Background

Cohort identification tools can support queries of various complexities, spanning the gamut from basic to more complex ones. Sholle et al[1] found that an overwhelming number of queries on the i2b2 platform could be classified as basic. The goal of this project is to evaluate query complexity of the TriNetX (TNX) live platform.

Methods

Over 200,000 queries from 2019 (to avoid COVID-19 skewing the data) were pulled from a Kibana database, using a script created in R. A query was defined as any search on the platform by users excluding TNX employees, that returned a non-empty result. Seven query complexity metrics were defined, with each corresponding to different features or capabilities of the platform's query builder. Each query was scored on all metrics and all scores normalized on a scale from zero to nine and weighted. These weightings reflect the importance of individual features of the query builder as agreed upon by a group of TNX subject-matter experts. Normalized and weighted scores were added together to arrive at the total complexity score. The mean of the total complexity scores was chosen as the cutoff for determining whether a query is "complex" or not, using graphs generated in R to aid in this decision.

Results

We found that 90% of queries included a diagnosis from ICD-10, and 40% of queries included a temporal restriction. Surprisingly, out of 860,000 unique concepts in the platform's terminology, only 15% appeared in our query corpus. The queries are relatively evenly distributed when ordered by their complexity scores (Figure 1). With the maximum possible score of 63, the mean complexity was observed at 9.78. We segregated the queries by organization type and showed that users from pharmaceutical organizations generated more complex queries with a mean of 11.27 on average and above our complexity cutoff, while queries from healthcare provider organizations were slightly less complex with a mean of 8.98 which falls below our complexity cutoff. This is visible in Figure 1, with the mean for each type as the red lines, and the complexity cutoff as the black lines.

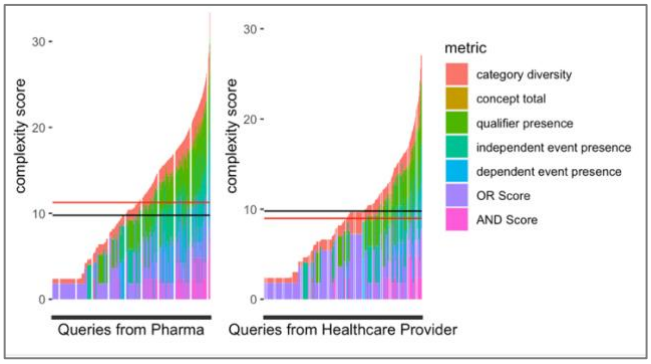


Figure 1. Query Complexity, ordered from lowest to highest score, and segregated by organization type.

Conclusion

We found a smooth query complexity value function, indicating that queries on TNX platform range from simple to complex. Pharma queries appear more complex than healthcare providers on average. These findings can be used by TNX employees to indicate which users may not use the platform to its full ability, and to indicate which platform parts they should focus on improving. These metrics could also be used by programmers at other firms using databases such as OMOP or I2B2 to figure out what their users look for on their platforms, and improve their platforms likewise.

References

1. Sholle ET, Cusick M, Davila MA, Kabariti J, Flores S, Campion TR. Characterizing Basic and Complex Usage of i2b2 at an Academic Medical Center. AMIA Jt Summits Transl Sci Proc. 2020 May 30;2020:589-596.

DEPOT: Revealing Trajectories of Alzheimer's Disease Development for Precision Monitoring and Interventions Using Real-world Evidence

Jing Su, PhD[1,2], Qianqian Song, PhD[3], Mark A. Espeland[2], Jeff D. Williamson[2]
[1]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, USA; [2]Section on Gerontology and Geriatric Medicine, Department of Internal Medicine, [3]Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC, USA

Introduction. It is important to characterize the trajectories, clinical markers, and risk factors through the whole course of Alzheimer's disease (AD). Real-world evidence (RWE) such as electronic medical records (EMRs) provides rich clinical details of real-life scenarios, making it promising *for mining novel clinical features* related with AD progression. However, *the unique characteristics of longitudinal EMR data* (Figure 1), including high sparsity, irregular temporal interval, and informative censoring, challenges traditional approaches.

Data and Method. To fully unleash the power of EMR data in longitudinal study of AD, we develop *the DisEase PrOgression Trajectory* (*DEPOT*) (Figure 2), a novel graph learning algorithm using *beads-on-strings graphs* to represent irregular longitudinal data, hybrid locality-sensitive hashing to cluster heterogenous and sparse clinical encounter data, and Hasse diagram to reconstruct disease progression trajectories. The *hypothesis* of DEPOT is that clinical encounters are sampled from the underlying AD progression trajectories. Therefore, the overall topology of the trajectories can be learned using local similarity among encounters and longitudinal ordering of these encounters. DEPOT learns trajectories as a directed graph with major progression phases as nodes, the transitions between these phases as directed edges, and associated features as clinical markers and risk factors. Once the AD progression trajectory of the whole population is learned, the longitudinal EMR data of an individual patient can be mapped to this overall trajectory for his personalized trajectory. Moreover, the AD risk of this patient can be determined by his current location on the graph, and the clinical interventions can be recommended by identifying the clinically modifiable risk factors.

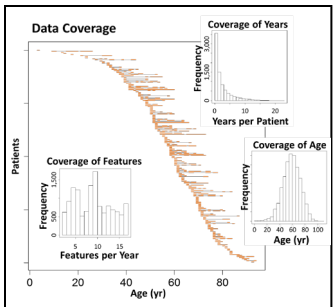


Figure 1. Data Coverage.

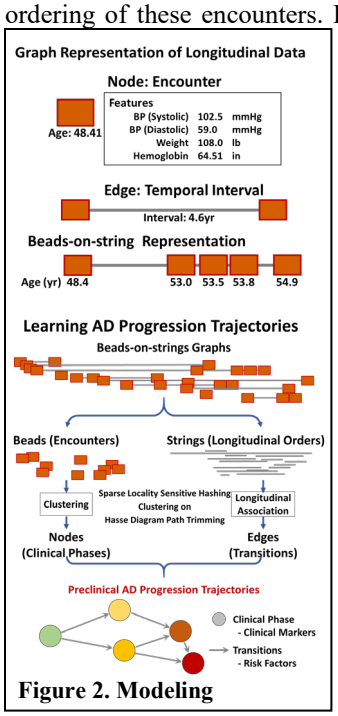


Figure 2. Modeling

Results and Discussion. The Wake Forest Baptist Health's Diabetic AD Risk dataset (all with type 2 diabetes; age of encounters: 58.9 ± 13.8yr; cohort size: 33,125; AD-related case: 3,171; features: 17 essential laboratory, vital, and demographical concepts defined by the Carolinas Collaborative EMR network) is used in this study. The constructed overall trajectory of AD progression includes 15 clinical phases and 27 transitions identified from longitudinal orders (Figure 3). Interestingly, the declining kidney function is closely associated with the increased risk of developing AD-related diseases, which is consistent with recent clinical discoveries (1). We also construct the personalized trajectory for an individual patient. Further annotations using EMR data such as ICD codes and procedure codes show consistent results with clinical markers. Patient-specific trajectory enables personalized clinical cares for monitoring and managing AD.

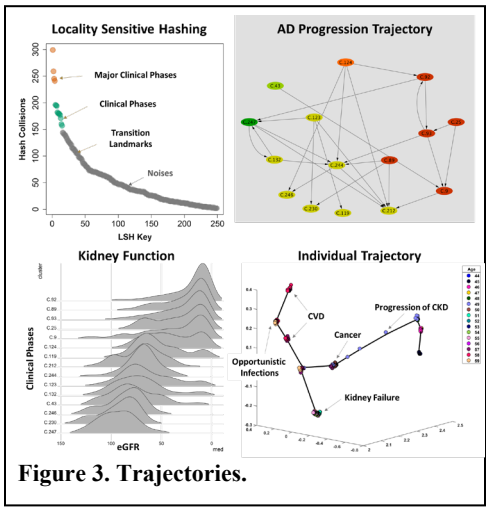


Figure 3. Trajectories.

Conclusion. We have developed a novel graph-based learning model (DEPOT) to address the challenges in utilizing longitudinal EMR data, which successfully constructs the AD progression trajectories and paves new ways for individual and precision care of AD.

References

1. Etgen T. Kidney disease as a determinant of cognitive decline and dementia. Alzheimers Res Ther. 2015;7(1):29.

The PREVENT Tool: An Implementation of the Yelp Fusion Application Programming Interface (API)

Esra Taner, MS[1], Randi Foraker, PhD, MA, FAHA[1], Simon Lara[2], Marcelo Lopetegui, MD[2], Maura Kepper, PhD[1,3]

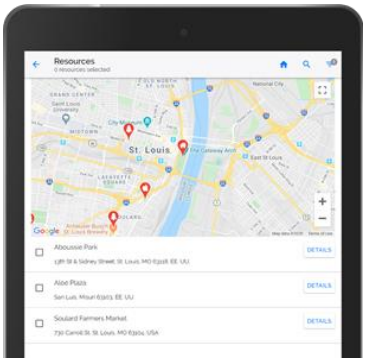
[1]Washington University in St. Louis, School of Medicine Institute for Informatics, St. Louis, MO, USA, [2]Universidad del Desarrollo, Concepción, Chile, [3]Washington University in St. Louis, Brown School, Prevention Research Center, St. Louis, MO, USA

Introduction

Despite the rapid emergence of health information technologies (HIT), they have rarely been utilized to promote physical activity and healthy food intake and address several of the social determinants of health (SDOH) to prevent chronic disease[1,2,3]. Our PREVENT tool is one of the first HITs that strives to address this gap by integrating and hosting an interactive map of community resources to help providers link patients to resources at the point of care to help them be active and eat healthy foods.

Methods

The PREVENT tool hosts resources within the following ten categories: Fitness/Sports Center, Parks and Playgrounds, Community/Recreation Centers, Grocery Stores, Farmer's Markets, CSA's, Food Pantries, Food Organizations, Community Gardens, and Weight Management Programs. The Yelp Fusion API was chosen to populate resources in these categories since it hosts content from 50 million businesses, and provides extensive documentation, which promotes a user-friendly interface for obtaining and extracting information. This API hosts information within "All Categories" and "Category Details", respectively. For example, the Yelp parent category "Active Life" includes diverse subcategories such as "ATV Rentals/Tours", "Skiing", and "Parks." Subcategories were selected and grouped into the new PREVENT tool categories to help users select resources that will help them be physically active and eat healthy food. If subcategories were irrelevant or inaccessible in the Greater St. Louis region they were excluded.



Results

The PREVENT tool includes an interactive map that displays YELP resources within the newly created categories (Figure 1). Resource categories are searchable, color coded and have unique icons for a user-friendly interface. The tool is intended to be used by healthcare teams during a clinical care. This abstract presents only a portion of the tool—PREVENT displays patient's cardiovascular health score, provides personalized, evidence-based physical activity and nutrition prescriptions, and includes the interactive map of community resources to help patients locate resources that will help them be active and eat healthy foods. After the visit, PREVENT sends patients an email with the community resources, prescriptions, and educational resources to promote sustained behavioral modification, along with automatic follow-up.

Conclusion

The inclusion of a community resource map with individualized health recommendations has the potential to assist patients in achieving sustained behavioral change. The impact of this tool may be limited if a patient lives within a resource-poor region and/or a patient lacks the means to utilize resources (e.g., reliable transportation).

Figure 1. PREVENT Tool Map

References

1. Tung E, Peek M. Linking community resources in diabetes care: a role for technology?. Curr Diab Rep. 2015;15(7):45
2. Hood K, Hilliard M, Piatt G, Ievers-Landis C. Effective strategies for encouraging behavior change in people with diabetes. Diabetes Management. 2015;5(6):499-510.
3. Payne P, Lussier Y, Foraker R, Embi P. Rethinking the role and impact of health information technology: informatics as an interventional discipline. BMC Medical Informatics and Decision Making. 2016;16(1).

Leveraging Free-text Clinical Notes for Accurate Diagnostic Decision Support Using Pre-trained Language Representation Models

Rui Tang, MS[1], Haishen Yao, MS[1], Zhaowei Zhu, MS[1], Xingzhi Sun, PhD[1]
[1]Ping An Healthcare Technology, Beijing, China

Introduction

Diagnostic Decision Support Systems (DDSS) provide physicians with assistance in the clinical diagnostic decision-making task. DDSSs have been widely used to reduce medical errors, i.e. misdiagnosis (incorrect diagnosis) and missed diagnosis (lack of diagnosis). In this work, we aimed to leverage free-text clinical notes for accurate diagnostic decision support using advanced pre-trained language representation techniques. We utilized a pre-trained language representation model, which is BERT, to obtain the representation of free-text clinical notes and built a diagnostic model to provide diagnostic decision support. The free-text clinical notes in this work refer to chief complaint (CC) text, which record a patient's major health problem or concern and its time course, e.g. chest pain for past four hours.

Methods

We constructed the dataset with 434,168 visits from a real-world dataset collected from the hospitals in a Chinese province, filtering out visits which did not contain valid CC text and did not correspond to a diagnosis code that follows the ICD-10. We extracted symptoms from CC text using an entity extraction tool. There are 1,593 unique diagnosis codes and 1,549 unique symptoms occurred in CC text. Each visit in the dataset contains CC text (free-text), symptoms extracted from CC text (structured data), concatenated symptom text (free-text), and a diagnosis code.

We built three diagnostic models trained on structured data and free-text data. We applied BERT to build two diagnostic models using free-text data. Specifically, the model *BERT-text* was trained on CC text and the model *BERT-sym* was trained on concatenated symptom text. The baseline model *NN-sym* was trained by a neural network using structured symptoms. We randomly divided the dataset in training, validation, and testing set in 8:1:1 ratio. We used the training set to develop the models. The hyperparameters of models were adjusted based on validation set. We calculated top-k accuracy (accuracy@k) on testing set to assess the performance of diagnostic models. We assigned k to 1, 3, 5, and 10, and computed accuracy@k respectively.

Results

Table 1. Model performance using structured or free-text data.

Model	**Accuracy@1**	**Accuracy@3**	**Accuracy@5**	**Accuracy@10**
NN-sym	32.77%	50.56%	57.87%	70.21%
BERT-sym	35.15%	54.86%	64.48%	75.44%
BERT-text	**59.68%**	**79.34%**	**84.96%**	**90.43%**

Table 1 showed the improvements for *BERT-sym* and *BERT-text* compared to the baseline model *NN-sym*. *BERT-text* achieved the best performance with accuracy@1 as 59.68%, which was significantly higher than *BERT-sym* and *NN-sym* by more than 20%. *BERT-sym* outperformed the baseline *NN-sym*, although their input information was same. It demonstrated that leveraging free-text clinical notes and using pre-trained language models can boost the performance of diagnostic models.

Conclusion

In this work, we utilized pre-trained language models to obtain the representation of free-text clinical notes and built diagnostic models that can assist physicians in making diagnostic decisions. Our model can diagnose 1,593 types of diseases (defined in ICD-10) and achieve the good performance with the accuracy@10 more than 90%. The model learnt from the free-text clinical notes outperformed the one learnt from the structured data extracted from free-text clinical notes. This demonstrated the effectiveness of pre-trained language models on patient data representation and diagnostic model learning.

Identifying Bad Actors in a Direct-to-Patient COVID-19 Registry

Kevin P. Timms, PhD, Emma Brinkley, Kalyani Hawaldar, MS, Charles Tirrell, Leon Rozenblit, JD, PhD,
IQVIA, Durham, NC

Introduction

Direct-to-patient (DTP) data registries offer a means to rapidly obtain self-reported symptoms, disease-specific information, and treatment effectiveness and safety. Recognizing that self-reports via open web interfaces can attract submissions from bad actors (bots or low-effort or malicious humans) who can submit inaccurate information thereby corrupting data quality, we explored ways to identify these actors in a large DTP COVID-19 registry[1]. This exploratory work sheds light on the relative amount of bad actor submissions to this dataset, and more generally on the possible suitability of various approaches for such data quality analyses in similar registries.

Methods

We approached the problem through iterative data examination to identify rules for detecting bad actors. To explore possible signals of bad intent, we examined two text fields (email address and a field on clinical trial enrollment), user authentication data, clinical data, and results of an "identity resolution" service[2]. Finally, we developed our broadest set of potential DQ "flags" and applied Machine Learning (ML) to derive predictive models using those flags.

Results

First, we found systematic patterns in email addresses. Two examples include strings matching the Python regular expressions *[a-zA-Z]\d{2,3}[a-zA-Z].*@[hotmail|outlook]* and *^yy\d.*@hotmail* (517 and 17 cases were flagged, respectively). A similar analysis on the "name of clinical trial" field showed 318 field submissions had the value "treatments", 225 had "1", and 12 had joke or offensive submissions. The frequency of these patterns, some of which were unusually unique or irrelevant, suggest systematic submission by bots or bad actors. User authentication (sign up process) information also shed light on suspicious DQ: 300 IP addresses attempted authentication more than three times, one from Vietnam attempting 654 times, and we geolocated 49 sets of fully *completed* surveys from non-US IPs. Furthermore, email bounce back information elucidated user registration from 1,440 non-existent emails. Next, a set of rules was developed to flag improbable or unexpected clinical data points. For example, 879 users reported being pregnant males and 196 under 4 ft tall. Additionally, possible duplicates or bad actor submissions were identified based on identical demographic datapoints and submission timing. Lastly, we provided 6 personally identifiable information (PII) database fields (e.g., name, address) to a third-party identity resolution vendor. Using proprietary algorithms and databases they returned a 0-100 score (s_i) for each submitter indicating likelihood they are not a real person[2]; e.g., phone number 000-000-0000 leads to a lower score. 8.070 users were labeled a score below 50. Finally, we estimated multiple regression models using Python's sklearn in which various sets of binary flags alluded to above — i.e., those based on authentication or clinical data — were features and s_i was the outcome (in lieu of a verified 'truth' label). Examining feature importance across models, we found that flags based on IP geolocation, email bounce backs, and clinical trial 'names' were regularly important across models. Even though these regression models were not of high predictive ability, the consistency of the importance of these flags suggest they are reflecting some bad actor phenomenon also captured in the identity resolution confidence score.

Conclusion

These analyses identified patterns of data submissions that are confidently from some form of bad actors. The results also highlight how regression-based methods to identify bad actors require accurate 'truth' labels (which were not available here). In lieu of such labels, the repeated and unexpected patterns in free text fields found here suggests future bad actor identification by researchers can likely benefit significantly from first analyzing free text submissions.

References

1. IQVIA COVID Active Research Experience (CARE) project [Internet]. Durham (NC): IQVIA, Inc.; 2020 [cited 2020 Aug 21]. Available from: https://www.helpstopcovid19.com/
2. Infutor [Internet]. Oakbrook Terrace (IL): Infutor Data Solutions, Inc.; c2020. Lead scoring and verification; c2020 [cited 2020 Aug 21]. Available from: https://infutor.com/consumer-identities/scoring-verification/

Consult Tracking Management Solution and Scheduler Workflow Metrics in the U.S. Department of Veterans Affairs

Alyssa Tsai, B.S.[1] , Evelynn J. Gerardo, B.S.[1], Linda S. Droshine, B.S., M.A.[1], Michael D. Watson, B.A.[1], David V. LaBorde, M.D., M.B.A.[1]
[1]Document Storage Systems, Juno Beach, FL, USA

Abstract

Specialty care scheduling delays can adversely impact patient access, the timeliness of care, and even patient outcomes. In this report, we share the findings from a retrospective analysis of two scheduling process metrics before and after the implementation of a facility wide, consult management solution used by schedulers / medical support assistants in the U.S. Department of Veterans Affairs.

Introduction

Successful completion of referrals for specialty medical or surgical care, often called consults, are critical to the delivery of highly reliable care. One of the first actions required when a consult is requested is the scheduling of an appointment with the specialty provider clinic. Delays in scheduling these visits can adversely impact access to care, patient outcomes, patient satisfaction, and patient safety; as such it is a point of patient vulnerability[1]. Moreover, failure to complete consults can create financial liabilities under population health, value based care and even fee-for-service financial models[1]. Tools that can help track and streamline scheduler / medical support assistant (MSA) workflows can potentially be of benefit and improve the time to consult completion. Herein we report an analysis of two scheduling workflow process metrics before and after the implementation of a consult tracking management software system at a medium size U.S. Department of Veterans Affairs (VA) medical center (VA complexity level 2).

Methods

A consult management tracking software system (*Consult Tracking Manager, DSS, Inc., Juno Beach, FL*) was procured by the facility and fully integrated into the system of record. The system went live at the facility March 13, 2018. As part of routine health care operations before and after the implementation of the consult tracking software, the facility tracked the following process metrics: i) average days from first active to first scheduled (AD-FFATFS); and ii) average days from first forwarded from to first scheduled (AD-FFFTFS). A one-tailed paired t-test was used to compare the AD-FFATFS and AD-FFFTFS metrics for an 18-month period prior to go-live to the same metrics for an 18 month period after go-live. In the retrospective analysis, the null hypothesis was that there would be no difference in the average AD-FFATFS and AD-FFFTFS metrics after system go live; the alternative hypothesis was that the difference in both would be greater than zero. Type I error was permitted to be no greater than 0.05.

Results

The baseline monthly AD-FFATFS in the 18 month pre-go-live period was 7.4 days ± 1.0 (SD) and the baseline monthly AD-FFFTFS was 10.0 days ± 2.4 (SD) as compared to the average AD-FFATFS post-implementation which was 4.6 days ± 0.6 (SD) and the average AD-FFFTFS which was 7.8 days ± 2.9 (SD); the differences observed were statistically significant for AD-FFATFS ($p = 6.68 \times 10^{-10}$) and AD-FFFTFS ($p = 0.01$) and the null hypothesis was refuted for both process metrics.

Table 1. t-Test: Paired Two Sample for Means - Average Days From First Active to First Scheduled (AD-FFATFS)

	Pre-Go-Live	*Post-Go-Live*
Mean	7.4	4.6
Variance	0.9	0.3
Observations	18	18
Pearson Correlation	0.241	
Hypothesized Mean Difference	0	
df	17	
t Stat	11.78	
P(T<=t) one-tail	6.68E-10	

Table 2. t-Test: Paired Two Sample for Means - Average Days From First Forwarded From to First Scheduled (AD-FFFTFS)

	Pre-Go-Live	*Post-Go-Live*
Mean	10.0	7.8
Variance	5.6	8.5
Observations	18	18
Pearson Correlation	0.198	
Hypothesized Mean Difference	0	
df	17	
t Stat	2.76	
P(T<=t) one-tail	0.01	

Discussion

In this single facility retrospective analysis, the AD-FFATFS and AD-FFFTFS metrics improved in the time period that temporally followed deployment of a software intervention targeting consult management workflow improvement. Further work could include testing reproducibility at other facilities and a similar prospective analysis.

References

1. Patel MP, Schettini P, O'Leary CP, Bosworth HB, Anderson JB, Shah KP. Closing the referral loop: an analysis of primary care referrals to specialists in a large health system. J Gen Intern Med. 2018;33(5):715-721.

Development and Evaluation of a Natural Language Processing (NLP) Pipeline for Extracting Endoscopic Measurements

Mohammed Ullah[1], Sean T. Pompea, MFA[1], Thomas R. Campion, Jr., PhD[1,2,3,4], Evan T. Sholle, MS[1]

[1]Information Technologies and Services Department, Weill Cornell Medicine, New York, NY; [2]Clinical and Translational Science Center, Weill Cornell Medicine, New York, NY; [3]Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY; [4]Department of Pediatrics, Weill Cornell Medicine, New York, NY

Introduction

Crohn's Disease (CD) and ulcerative colitis are two particularly debilitating aspects of inflammatory bowel disease (IBD), a condition that primarily affects the gastrointestinal tract. However, CD does not present similarly in all patients. Numerous structured scoring systems can evaluate CD severity, including the Rutgeerts Score, which evaluates the potential for recurrence of CD after resection, the Simple Endoscopic Score for Crohn's Disease (SES-CD), which assesses the presence, size, and frequency of ulcers and stenosis in the ileocolonic segment of the gastrointestinal tract, and the Mayo Endoscopic Score, which serves as a key indicator of ulcerative colitis. Though these measurements are critical for both clinical care and research, they are often stored as unstructured data points and are not amenable to automated extraction, necessitating the use of natural language processing (NLP), a suite of computational techniques heretofore underutilized in this domain[1]. We developed and evaluated an NLP pipeline for the extraction of Rutgeerts, SES-CD, and Mayo scores.

Methods

We obtained notes for patients treated for IBD from our institution's EHR system. Rutgeerts and SES-CD scores were only extracted from endoscopy reports, while Mayo scores were isolated from either endoscopy reports or physician progress notes. Based on an initial review of notes, we iteratively developed two rule-based algorithms. To extract Rutgeerts scores, a regular expression in Python was created as follows: "(rutgeerts|rutgeert|rutgers)\s*(score\s*|was\s*|is\s*|of\s*){0,4}(i?[0-4])". To address variability in reports of SES-CD and Mayo scores, we implemented a finite state machine to extract these values.

NLP Performance Evaluation

The performance of the NLP pipeline was then evaluated on a set of notes distinct from the corpus through which the regular expressions were developed. Each case evaluated was categorized as either true positive (NLP-derived and manual scores matched), false positive (NLP extracted a score but manual review did not OR NLP extracted an incorrect score), false negative (manual review extracted a score but NLP did not), or true negative. Using these results, we constructed a confusion matrix and calculated precision, recall, and F-score.

Results

This NLP pipeline exhibited success in extracting Rutgeerts, SES-CD, and Mayo scores with f-scores of 1.00, 1.00, and 0.99 from samples of 51, 320, and 200 notes, respectively. Rutgeerts and SES-CD scores were extracted with a perfect level of recall and precision, while Mayo scores were extracted with a recall of 0.98 and precision of 0.99. Code is freely available at http://www.github.com/wcmc-research-informatics.

Discussion

Despite variability in endoscopy reports in patient records, this NLP pipeline demonstrated an effective and consistent extraction. We encourage investigators to use our publicly available pipeline to assist in evaluating the portability of this method in extracting Rutgeerts, SES-CD, and Mayo scores.

References

1. Imler TD, Morea J, Kahi C, Imperiale TF. Natural Language Processing Accurately Categorizes Findings From Colonoscopy and Pathology Reports. Clinical Gastroenterology and Hepatology. 2013;11(6):689–94.

MEDTYPE: Improving Medical Entity Linking with Semantic Type Prediction

Shikhar Vashishth, PhD[1], Rishabh Joshi, BS[1], Denis Newman-Griffis, PhD[2], Ritam Dutt, MS[1], Carolyn Rose, PhD[1]
[1]Carnegie Mellon University, Pittsburgh, PA; [2]University of Pittsburgh, Pittsburgh, PA

Introduction

Identifying the standardized concepts referred to in an unstructured text is a critical component of biomedical natural language processing, enabling harmonization across different documents for search and semantic analysis. Medical entity linking, also referred to as medical concept normalization, is the task of harmonizing different surface forms for the same concepts: for example, identifying that documents mentioning *Amyotrophic lateral sclerosis* and *Lou Gehrig's Disease* are referring to the same disease. A key step in this process is candidate generation, the identification of candidate medical concepts a text could be referring to. This process is prone to overgeneration, producing too many candidates that make normalization difficult and lead to erroneous predictions. In this work, we propose an intermediate step between candidate generation and final normalization decisions that alleviates the overgeneration. For this, we present MEDTYPE, a fully modular system that prunes out irrelevant candidate concepts based on the predicted semantic type of an entity mention. MEDTYPE utilizes a Transformer-based encoder for modeling the context of a given mention. To address the dearth of annotated training data for medical entity linking, we also present two novel large-scale datasets: WIKIMED, biomedical subset of Wikipedia corpus for entity linking, and PUBMEDDS, a distantly-supervised dataset of medical entity mentions, which help improve the performance on the task.

Method

We experimented with incorporating MedType into standard medical entity linking tools and evaluated its impact on entity linking performance in four datasets: NCBI, Bio CDR, ShARe, MedMentions for medical entity linking, as well as our novel WIKIMED dataset. The chosen datasets span across different domains such as biomedical research articles, Electronic Health Records (EHR), and general domain text. Thus, they allow us to evaluate the generality of MEDTYPE across diverse domains. MEDTYPE is evaluated for pruning candidate concepts generated from five entity linking models: MetaMap, cTAKES, MetaMapLite, QuickUMLS, and ScispaCy. We run four sets of experiments for demonstrating the effectiveness of our approach. Experiment 1 evaluates how incorporating MEDTYPE in existing entity linking systems helps improve medical entity linking. Experiment 2 demonstrates that the proposed datasets WIKIMED and PUBMEDDS are effective for semantic type prediction. Experiment 3 analyzes the gains obtained using the proposed datasets. Experiment 4 quantifies how type-based filtering helps prune irrelevant candidates.

Results

Experiment 1. The results show that the type-based filtering of candidates using MEDTYPE enhances entity linking systems across all 25 settings evaluated. MEDTYPE obtains statistically significant improvement of up to 4.2% F1-score over default entity linking without using semantic type prediction. Experiment 2 shows a substantial gain in performance on using the proposed datasets. Overall, we get an average absolute increase of 9.8, 12.7, and 13.4 AUC on type prediction from WIKIMED, PUBMEDDS, and the combined corpora respectively. Experiment 3. The results show that for semantic types such as *Sign or Symptoms*, utilizing WIKIMED and PUBMEDDS gives an absolute increase in F1-score of 24, 28 on Bio CDR and 12, 14 on ShARe respectively. Experiment 4. The type-based filtering removes all incorrect candidates in 36.5% cases while in 25.6% cases it helps to reduce the candidate set size. This clearly demonstrates the practical benefits of our proposed approach.

Conclusion

We found that filtering out irrelevant candidate concepts based on the predicted semantic type improves entity linking performance for a variety of popular medical entity extraction toolkits across several benchmark datasets. We further present two novel large-scale datasets: WIKIMED and PUBMEDDS. Pre-training on these datasets substantively improves MEDTYPE performance, and we share these datasets with the community as a resource for medical entity linking research at `http://github.com/svjan5/medtype`.

Ascertaining Chronic Condition Status from Electronic Health Records with Patient Problem Lists versus Encounter Diagnosis Records

Robert Voss, MS[1], Teresa Schmidt, PhD[1], Jean O'Malley, MPH[1], John Heintzman, MD MPH[1,2], Nathalie Huguet, PhD[2], Miguel Marino, PhD[2], David A. Dorr, MD MS[2], Nicole Weiskopf, PhD[2], Jennifer Lucas, PhD MPH[2], Steele Valenzuela, MS[2], Kate Peak, MPH[2], Nate Warren, MPH[1], Ana Quiñones, PhD[2]
[1]Ochin, Inc., Portland, OR USA [2]Oregon Health & Science University, Portland, OR USA.

Introduction: To guide research on chronic condition multimorbidity and its impact on patients and healthcare delivery systems, the US Department of Health and Human Services (HHS) has identified conditions most impactful to patient multimorbidity burden[1]. Electronic health records (EHR) offer rich and timely source of clinical data which may be more appropriate for many studies[3], though data quality for patient identification is an area of ongoing investigation[4]. We used the Centers for Medicare and Medicaid Services Chronic Conditions Warehouse (CCW) algorithms[2] to identify chronic conditions in EHR encounter and problem list records. Problem lists are not consistently used, and diagnoses are not always recorded in EHR encounters[4]. Comparing ascertainment between these sources will inform EHR users who lack access to or cannot utilize claims data.

Methods: We used EHR records from Ochin Inc. and Health Choice Network patients in the ADVANCE (Accelerating Data Value across a National Community Health Center Network) clinical research network 2012-2016. We used ICD diagnosis codes from 39 CCW groupings to identify patients' chronic conditions, comparing prevalence by medical chart problem lists versus encounter diagnoses that met CCW time and count requirements.

Results: We assessed 1,191,110 patients aged 45+ with at least two office visits at one of 540 community health centers in 24 US states. We found 2,298,119 chronic conditions; 58.2% of patients had at least one. Encounter diagnosis and problem list ascertainment overlapped considerably: 59.4% of conditions were identified by both sources by study end, while 12.3% were identified only in encounter diagnoses and 28.3% were only in problem lists. Onset dates by these sources were similar, with 69.9% of dually identified conditions appearing within 6 months of one another (78.1% within 12 months). Overlap varied by condition (Figure 1).

Discussion: Application of CCW algorithms to EHR encounter data, without reference to the patients' problem list, may undercapture chronic condition prevalence among community health center patients in the United States.

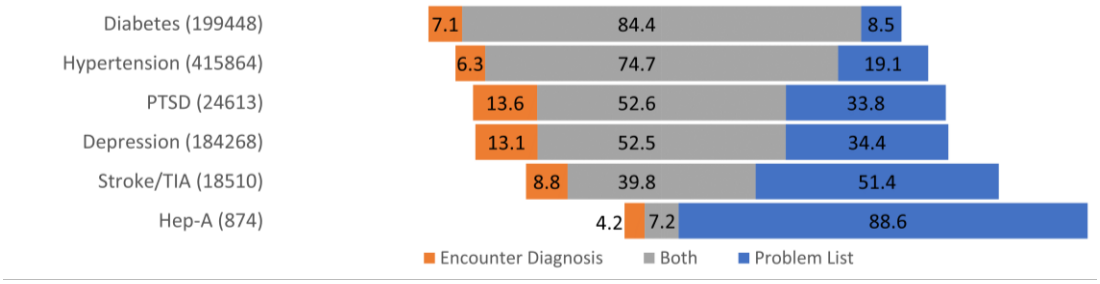


Figure 1: Chronic condition ascertainment in problem list and encounter diagnosis, EHR records. (A subset of full results. Patient counts in parenthesis, patient proportion (%) in bars).

References:

1- US Department of Health & Human Services. About the Multiple Chronic Conditions Initiative. 2016. https://www.hhs.gov/ash/about-ash/multiple-chronic-conditions/about-mcc/index.html. Accessed Aug. 18, 2020.
2- Centers for Medicare and Medicaid Services. CCW Chronic Condition Algorithms.pdf. Chronic Conditions Data Warehouse. https://www2.ccwdata.org/web/guest/condition-categories. Accessed Aug. 18, 2020.
3-Shephard E, Stapley S, Hamilton W. The use of electronic databases in primary care research. Fam Pract. 2011; (28):352–354.
4-Singer A, Yakubovich A, Kroeker AL, Dufault B, Duarte R, Katz A. Data quality of electronic medical records in Manitoba: do problem lists accurately reflect chronic disease billing diagnoses? J Am Med Inform Assoc. 2016;23:1107–1112.

Network analysis to characterize chronological relationships in comorbidity trajectories: breast cancer as a case model

Clark D. Wang[1,2], William Zhang[1,2], Jonathan X. Wang, MS[1], Julian C. Hong, MD, MS[1]
[1]UCSF, San Francisco, CA; [2]UC Berkeley, Berkeley, CA

Introduction

The health trajectory of accumulating acute and late comorbidities has significant consequences for patients with cancer, including short term complications and long term survivorship. Network analysis, particularly directed graphs, offer an opportunity to use electronic health record (EHR) data to characterize these patterns of comorbidity in a population, facilitating the identification of management priorities. Applications of directed graphs in characterizing the trajectory of patients with cancer have been limited. The objective of this study was to develop a directed graph to model population-wide chronological relationships between comorbidities and then analyze this graph focusing on breast cancer as a case model.

Methods

Utilizing a single institution clinical data warehouse, we identified patients with at least 5 years of EHR history. Historical patient data from a prior EHR was excluded. First instances of each three-digit ICD-10-CM diagnosis were identified, excluding Z (factors influencing health status and contact with health services) and R (symptoms, signs and abnormal clinical and laboratory findings) diagnoses. Diagnoses affecting fewer than 5,000 patients were excluded. An adjacency matrix of relative risk (RR) between all pairs of diagnoses was generated. NetworkX 2.5[1] (Python 3.7.7) was used to create a directed graph of comorbidities, using a threshold of RR > 1. As a case model, we then focused our study on the subgraph induced on breast cancer (C50) and nodes one outgoing edge away, or subsequent diagnoses. Hyperlink-Induced Topic Search (HITS) algorithm was performed on the subgraph to identify top hubs (preceding) and authorities (subsequent). A Personalized PageRank (PPR) algorithm was performed on the larger graph from the perspective of breast cancer, identifying relevant subsequent diagnoses.

Results

The study population included 340,292 patients. A directed graph was produced with 292 diagnoses and 43,333 edges (connections between diagnoses) after filtering as described. A subgraph, limited to breast cancer and subsequent diagnoses, included 154 diagnoses and 16,175 edges. In this subgraph, top authorities were pulmonary edema (J81), adverse effect of primarily systemic and hematological agents (T45), respiratory failure (J96), and polyneuropathies (G62). Additionally, the diagnoses with the highest PPR scores from the original graph, from the perspective of breast cancer, are adverse effect of primarily systemic and hematological agents (T45), pulmonary edema (J81), and polyneuropathies (G62).

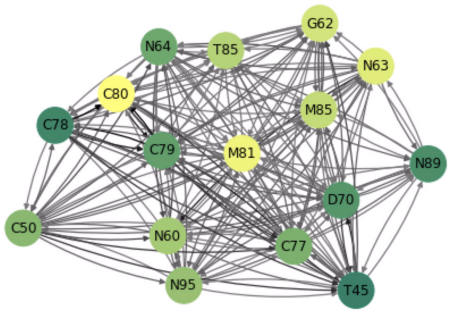


Figure 1: Subgraph of C50, and its top 15 subsequent diagnoses by edge weight (RR). Darker diagnoses indicated a heavier edge from C50, and darker edges have higher RR.

Conclusion and Further Study

Directed graphs can model chronological relationships between comorbidities. We preliminarily evaluated diagnoses following breast cancer, generating hypotheses for common key diagnoses based on different approaches. Investigation is ongoing to characterize the complex interactions of comorbidities to target priorities to inform acute and long-term management.

References

1. Hagberg A, Schult D, Swart P. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008). 2008;1:11-15.

Leveraging Broadcast Text-Messages to Deliver Real-time Clinical Guidance to Hospital Employees During the Covid-19 Pandemic

Cheyenne Williams, BS[1], Aditi Rao, PhD, RN[1], Justin B. Ziemba MD, MSEd[1], Jennifer S. Myers, MD[1], and Neha Patel, MD, MS[1].

[1]University of Pennsylvania, Philadelphia, PA, USA

Background

Responses to the Covid-19 pandemic introduced drastic policy and practice changes necessary for protecting the safety of patients and staff. Hospital employees are a critical population for receiving consistent and timely communication about these changes[1]. Previously, urgent staff communications were limited to large email listservs that historically have poor readership[2], and text-messaging has been limited to communcation within care teams. Recognizing the need for an alternative communication platform, we developed a novel use of broadcast text-messaging to supplement traditional mass email.

Methods

Since 2013 Penn Medicine has utilized a HIPAA-compliant messaging app for communication about inpatient care[3]. A multidisciplinary team leveraged this app to deliver real-time, mass communication or "broadcast" texts to on-service users at the Hospital of the University of Pennsylvania. Users included, but were not limited to, physicians, nurses, respiratory therapists, case managers, social workers, and pharmacists. Users received the broadcast messages on any device with the app installed, including hospital and personal smartphones. Hospital leaders developed message content, focusing primarily on safe infection control practices, proper personal protective equipment use, and Covid-19 census updates. Messages were sent 3 times per week between 8-9am. Message recipients were also able to directly reply to the broadcasts with feedback or questions. The utility of this innovation was evaluated through analysis of "read" receipts which are automatically collected from all users. Effectiveness of this method was assessed by rates of occupational exposures to Covid-19 and by two cross-sectional attitudinal surveys administered to all text- broadcast recipients.

Results

On average, messages were sent to 1,997 users (range = 1,799-2,030; 13% of total app users) per broadcast. Analysis of "read" receipts revealed that on average, 1,198 (60%) texts were consistently read within 24 hours of delivery, 679 (34%) were read in 2 hours, and 326 (16%) were read in 10 minutes. Readership peaked and fell in the first week of messaging but remained consistent throughout the remainder of the intervention (Figure 1). A survey administered after 1 month (response rate 10%) revealed that, 163 (79%) users found broadcasts "valuable," 152 (73%) users would recommend these broadcast texts to their colleagues, and 114 (55%) users preferred texts to email. A second survey at 3 months (response rate 7%) revealed that 109 (80%) users continued to find broadcasts "valuable." Broadcast utilization correlated with a decrease in average daily occupational exposure events (21 pre-messaging vs. 1 post-messaging).

Conclusions

Broadcast text-messages sent to employee smartphones are an effective strategy for urgent communications. Our outcomes measurements are limited by low survey response rates and by occupational exposures being counfounded by factors like improvements in employee comfort and the regional caseload. Additionally, readership rates may be lowered by users who are not working within 24 hours of when messages were sent. Hospitals may leverage text-messaging during times of routine operations and crisis management to improve system-wide communcations.

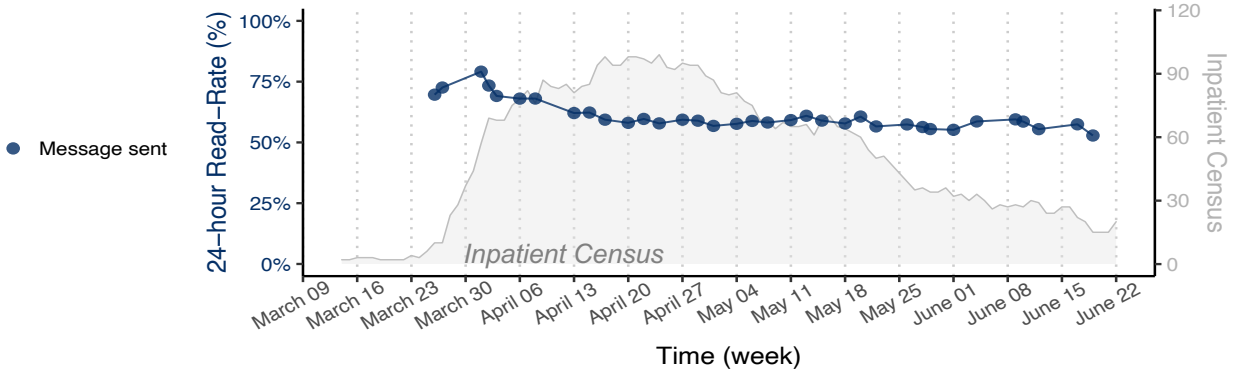


Figure 1: Aggregate 24-hour readership over course of broadcast messaging

References

1. Adams JG, Walls RM. Supporting the health care workforce during the COVID-19 global epidemic. *JAMA*. 2020;323(15):1439-1440
2. Paul IM, Levi BH. Metastasis of e-mail at an academic medical center. JAMA Pediatr. 2014;168(3):290-291.
3. Patel N, Siegler JE, Stromberg N, Ravitz N, Hanson CW. Perfect Storm of Inpatient Communication Needs and an Innovative Solution Utilizing Smartphones and Secured Messaging. Appl Clin Inform. 2016 Aug 10;7(3):777–89.

Automatic Gender Detection on Twitter for Health-related Cohort Studies

Yuan-Chi Yang, PhD[1], Mohammed Ali Al-Garadi, PhD[1], Abeed, Sarker, PhD[1,2]
[1]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA
[2]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA;

Introduction

Biomedical research involving social media (SM) data is moving from population-level to targeted, cohort-level data analysis. Though crucial for biomedical studies, SM user's demographic information (e.g., gender) is often not explicitly known from SM profile. Here we present an automatic and scalable gender classification system for SM and we illustrate how gender information can be incorporated into a SM-based Toxicovigilance study.

Methods

We collected two publicly-available datasets: (i) the gender-labeled users (Dataset-1), and (ii) the users who have self-reported prescription medication (PM) misuse (Dataset-2, the Toxicovigilance dataset with no ground truth labels). We first used Dataset-1 to training/evaluating the gender detection pipeline. We employed user's data including profile and tweets for classification and experimented with machine-learning algorithms including support vector machines and deep learning architectures and released packages including M3 pipeline.[1] We further developed meta-classifiers that strategically use the predicted scores from these classifiers. Applying the best-performing pipeline to Dataset-2, we further tested the system. First, we extracted the gender information for the users who revealed their gender publicly on their linked Facebook profiles. Using this gender information as the golden label, we grouped these users as a test set to evaluate our pipeline. Second, we inferred the gender distributions for three commonly abused PM categories (tranquilizers, stimulants, and pain relievers containing opioids) and compared them with data from Substance Abuse and Mental Health Services Administration (SAMHSA)[2] and CDC Wonder Database.[3]

Results

We collected 67,181 (35,812 female and 31,369 male) for Dataset-1 and 175,063 users for Dataset-2, respectively. In Dataset-2, we identified the gender of 413 users (155 female and 258 male) using their linked Facebook profiles, who make up the test set for Dataset-2. The best performing classifier was a meta-classifier based on SVM on tweets and M3 on profiles (Dataset-1 accuracy: 94.4% [95%-CI: 94.0%-94.8%]; Dataset-2: 94.4% [95%-CI: 92.0%-96.6%]). For stimulants and tranquilizers, the inferred female proportions using the meta-classifier are close to the data from the SAMHSA [tranquilizers: 0.50 *vs*. 0.50 stimulants: 0.50 *vs*. 0.45]. The inferred female proportion for pain reliever is close to the Opioid overdose Emergency Department visit from CDC Wonder Database [0.38 *vs.* 0.37].[3]

Discussion and Conclusion

Our gender detection pipeline shows promising results, with accuracy over 94% even on unseen data (Dataset-2). Further improvements on classification performances may be achieved by using more annotated data and using more sophisticated architectures. The similarity between the inferred gender distributions and those from the surveys shows that not only a SM Toxicovigilance study could be corroborated by surveys but also such a system can further helps us understand the PM misusers. We leave identifying transgender users for future work.

Ethically speaking, though we only use the publicly available data and adhere to Twitter API's use terms, we agree that Twitter users' perception toward user profiling may vary and limited our work to population research.[4]

References

1. Wang Z, Hale S, Adelani DI, et al. Demographic inference and representative population estimates from multilingual social media data. Paper presented at: The World Wide Web Conference2019.
2. Wilson N KM, Seth P, Smith H IV, Davis NL. . Drug and Opioid-Involved Overdose Deaths — United States, 2017–2018. *MMWR Morb Mortal Wkly Rep.* 2020;69:290-297.
3. Prevention CfDCa. Annual Surveillance Report of Drug-Related Risks and Outcomes — United States Surveillance Special Report. In: Centers for Disease Control and Prevention USDoHaHS, ed2019.
4. Williams ML, Burnap P, Sloan L. Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology.* 2017;51(6):1149-1168.

Identifying Barriers and Facilitators to Integrating Patient-Generated Health Data into EHR

Jiancheng Ye

Feinberg School of Medicine, Northwestern University, Chicago, USA

Introduction

In recent years, health-related data are increasingly collected by technologies such as portable devices with embedded sensors, remote monitoring devices, wearable devices, and smartphone apps. These data can be collected continuously outside of the clinical settings and be shared with health care providers (HCPs). Health-related data that is generated, recorded, or gathered by patients outside of the clinical setting without the assistance of HCPs is termed patient-generated health data (PGHD). Coupled with deployed EHRs, patient portals, and secure messaging, these new types of data enable patients to actively engage in the health care process, further improving the connection with their HCPs. In this way, the breadth, depth, and continuity of traditional health-related data are expanded, thus contributing to improved treatment adherence, health outcomes, healthcare quality, and patient safety.

Methods

This study examines how interactions with EHR-integrated PGHD may result in physician burnout and to identify the potential contributing factors using a mixed-methods approach.

Results

In this paper, we focus on two main elements of PGHD: Patient-reported outcomes (PROs) and mobile health (e.g. mobile apps, wearable devices, or portable devices, etc.).

PRO

PROs are assessments of patients' health conditions reported directly from patients in the form of questionnaires. PROs have been increasingly recognized as necessary elements of clinical information. The most commonly adopted PRO measurements include general health perceptions, quality of life, functioning, etc. Evidence has shown that health-related outcomes reported by patients have higher accuracy than clinical reports, and that patient reporting can improve patient-provider communication, patient satisfaction, and symptom management. Widespread adoptions of PROs in performance evaluation cater to the growing interest in integrating PROs into EHR systems and patient portals.

Mobile health

The growing market of smartphones, mobile apps, and wearables or portable devices that could be connected with smartphones have been increasingly harnessed to support health monitoring and management. Because healthcare systems have been interdependent on EHR capacities due to the widespread adoption and legislation of meaningful use, the integration of data generated by various devices into EHR becomes a novel and critical capacity of hospital information systems.

We identified three factors related to PGHD that can contribute to physician burnout: Technostress, Workflow-related issues, and Time pressure.

Conclusion

Incorporation of PGHD into EHR could lead to information overload through several known causes, including increasing time spent on learning new technologies; increasing amounts of available information, especially uncommon health-related information presentations; and increasing new collaborative work. Physicians may feel overwhelmed by the increased information overload, which could cause several negative influences, notably stress and burnout, resulting in increased medical errors and decreasing the quality of health care delivery and medical decision making. Furthermore, the mechanism through which PGHD is integrated into EHR would divert physicians' attention, thus increasing the chances of alarm fatigue, where alarms are ignored by physicians, potentially happening during serious situations. Integrating PGHD with EHR can aid in patient-provider communication, clinical decision-making, and improved outcomes. However, the ever-evolving technology and a large amount of data may also cause physician burnout, impairing job satisfaction, healthcare delivery and services, health care quality, and patient safety. This article outlines a number of challenges related to PGHD-EHR integration that may cause physician burnout. Addressing these issues could relieve physician burnout and better support PGHD-EHR integration. Moreover, incorporating artificial intelligence and algorithm-based approaches would improve clinical decision making and health care delivery. Advancing the algorithms to innovate the clinical support systems will also reduce physician burden and foster clinical decision-making.

Identify Cancer Patients at Risk for Heart Failure using Electronic Health Record and Genetic Data

Zehao Yu[1], Xi Yang[1], Yiqing Chen[2], Ruogu Fang[4], William R Hogan[1], Yan Gong[2,3], Yonghui Wu[1]
[1]Department of Health Outcome & Biomedical Informatics, College of Medicine, University of Florida; [2]Department of Pharmacotherapy and Translational Research and Center for Pharmacogenomics and Precision Medicine, College of Pharmacy, University of Florida; [3]University of Florida Health Cancer Center; [4]J. Crayton Pruitt Family Department of Biomedical Engineering, Herbert Wertheim College of Engineering, University of Florida, Gainesville, Florida, USA

Abstract

This study examined how genetic data can be used with Electronic Health Record (EHR) data to identify cancer patients at risk for heart failure. We explored four machine learning models for heart failure prediction using EHR and genetic data. The best area under the curve score improved from 77.32% to 77.48% after combined genetic data.

Introduction

This study examined how genetic data can be used with Electronic Health Record (EHR) data to identify cancer patients at risk for heart failure. We explored four machine learning models for heart failure prediction using EHR and genetic data.

Methods

Using the UK Biobank data, we identified cases (cancer patients who had heart failure after cancer diagnoses) using ICD-9 and ICD-10 codes and matched each of them to up to 10 controls (cancer patients without heart failure) by age, gender, race, and cancer type. We compared machine learning methods including Logistics Regression, SVM, Random Forest, and Gradient Boost (GB) for heart failure prediction using patient's EHR and genetic data[1]. To incorporate the genetic data, we included several heart failure-related single-nucleotide polymorphisms (SNPs) from previous studies as well as newly identified SNPs through genome-wide association analysis. Based on the best prediction model using only EHR data, we further added heart failure-related SNPs to examine how patients' genetic data could improve the prediction of heart failure among cancer patients. Following the standard machine learning procedure, we developed machine learning models using the training set and evaluated their performance using the test set.

Results

In the UK Biobank data, we identified 1,874 cases and matched them to 18,131 controls, resulting in a cohort of 20,005 patients. We divided the cohort into a training set of 16,004 patients and a test set of 4,001 patients using stratified sampling. When only using the EHR data, the GB model achieved the best area under the curve (AUC) score of 77.32% among the 4 machine learning models. After combining 17 heart failure-related SNPs[2] with the EHR, the AUC score of the GB model was improved to 77.48%.

Conclusion

This study demonstrates that genetic data can be used to further improve heart failure prediction on top of EHR data among cancer patients.

References

1. Yang X, Gong Y, Waheed N, March K, Bian J, Hogan WR, et al. Identifying Cancer Patients at Risk for Heart Failure Using Machine Learning Methods. arXiv:191000582 [cs, q-bio, stat] [Internet]. 2019 Oct 1 [cited 2020 Mar 31]; Available from: http://arxiv.org/abs/1910.00582

2. Thomas Mark R., Lip Gregory Y.H. Novel Risk Markers and Risk Assessments for Cardiovascular Disease. Circulation Research. 2017 Jan 6;120(1):133–49.

Integration between physician and nurse terminology in the care of patients with heart failure

Abdul Zakkar, MD[1], Haleh Vatani, MS[1], Barbara Di Eugenio, PhD[1], Carolyn Dickens, PhD, RN[1], Pamela Martyn-Nemeth, PhD RN[1], Karen Dunn Lopez, PhD RN[2], Amer K Ardati, MD[1], Andrew D Boyd, MD[1]

[1]University of Illinois at Chicago, Chicago, Illinois; [2]University of Iowa, Iowa City, Iowa

Introduction

Physicians and nurses are in constant communication in the hospital setting, yet their use of language in their care documentation often differs[1]. Previous studies have identified the need for including information on nursing care to truly reflect the total care provided by healthcare professionals, and nursing data has been implicated in readmission risk[1,2] The objective of this study is to examine the integration between physician and nurse terminology in the care of patients with heart failure, and to assess the inclusion of nursing care in their discharge notes.

Methods

As part of our larger research study (R01 CA225446-01), nursing terms were obtained from a publication focusing on heart failure patients including nursing diagnosis terms (NANDA-I), nursing interventions terms (NIC), and nursing outcomes terms (NOC)[3] for a total of 272 terms. We then obtained medical terms from 33 physician discharge summaries of heart failure patients by using the clinical Text Analysis Knowledge Extraction System (cTAKES)[4]. cTAKES mapped the health-related terms physicians used with the Unified Medical Language System (UMLS) ontology to extract the Concept Unique Identifiers (CUIs). UMLS is a comprehensive list of clinical concepts and terms from various controlled vocabularies that provides a mapping structure among vocabularies and enables linking synonymous terms from various terminology systems. We mapped the nursing terms to the UMLS CUIs as well. These two sets UMLS CUIs were compared to find similarities. This project was conducted under IRB #2019-0353

Results

17,020 health concepts were extracted from 33 physician discharge summaries of patients with heart failure using cTAKES. From those concepts, 2076 unique UMLS CUIs were extracted with an accuracy of 90.5%. 238 unique UMLS CUIs were then also extracted from 243 terms form published heart failure nursing documentation. We found that 6.7% (16/238) of the unique UMLS terms used by nurses treating patients with heart failure are also shared with the physician discharge notes. Of note, shared terms such as “presence” and “mobility” did not align contextually between medical and nursing terms. Of the 16 shared terms, 7 are NANDA-I, 2 are NOC, and 7 are NIC. 29 of the 33 physician notes contained at least one nursing term.

Conclusion

Our analysis builds on the findings of our team’s earlier research[1,]. Nurse and physician documentation for patients with heart failure reveals their different domains of care. Even when overlapping, the meanings can be different, which reveals the limitations of semantic mapping and computational analysis Given the small overlap in care terms, it is important that patients must receive information on care provided to them both by nurses and physicians. Nursing documentation data should be included in predictive algorithms to obtain a fuller picture of the patient care and response to treatment.

References

1. Boyd AD, Lopez KD, Lugaresi C, Macieira T, Sousa V, Acharya S, Balasubramanian A, Roussi K, Keenan GM, Lussier YA, Burton M. Physician nurse care: A new use of UMLS to measure professional contribution: Are we talking about the same patient: a new graph matching algorithm?. *International journal of medical informatics*. 2018 May 1;113:63-71.
2. Sanson G, Welton J, Vellone E, Cocchieri A, Maurici M, Zega M, Alvaro R, D’Agostino F. Enhancing the performance of predictive models for Hospital mortality by adding nursing data. *International journal of medical informatics*. 2019 May 1;125:79-85.
3. Park, Hyejin. “Identifying Core NANDA-I Nursing Diagnoses, NIC Interventions, NOC Outcomes, and NNN Linkages for Heart Failure.” *International Journal of Nursing Knowledge*, vol. 25, no. 1, Feb. 2014, pp. 30–38., doi:10.1111/2047-3095.12010.
4. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*. 2010 Sep 1;17(5):507-13.

Blockchain for DSCSA Compliance in Hospital-at-Home Programs

Stephanie Zawada, MS[1]
[1]Mayo Clinic College of Medicine and Science, Rochester, MN

Abstract

Regulatory relief implemented during the COVID-19 Public Health Emergency expanded access to remote health care, including telehealth and Hospital-at-Home. Remote care delivery involving prescriptions necessitates new models for pharmaceutical supply chain tracking. Current research suggests that blockchain can facilitate secure drug tracking with mobile devices. This project outlines and recommends methods to assess the feasibility of blockchain use cases for Drug Supply Chain Security Act (DSCSA) compliance in Hospital-at-Home programs.

Introduction

In November 2020, the Centers for Medicare & Medicaid Services (CMS) expanded its *Hospital Without Walls* initiative by launching the *Acute Hospital Care At Home* program, offering new waivers and regulatory relief for hospitals to deliver comprehensive care to remote patients for the duration of the COVID-19 Public Health Emergency (PHE). To ensure patient safety and compliance with the DSCSA, new strategies for tracking the delivery of drug and vaccine products to a patient's home are needed. Blockchain, a distributed, peer-to-peer public ledger, offers one method of real-time tracking for the pharmaceutical supply chain[1]. Feasible blockchain use cases for DSCSA compliance in Hospital-at-Home (HaH) programs are proposed in this project[2].

Blockchain Use Cases for DSCSA Compliance in Hospital-at-Home Programs

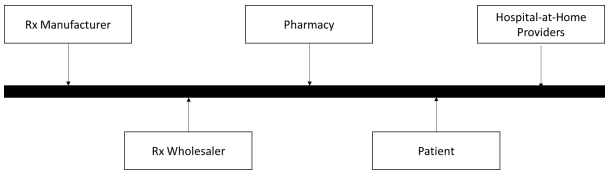


Figure 1. Prescription Delivery Model for HaH

The prescription delivery model (Figure 1) places increased responsibility on HaH programs to ensure patients receive safe medications. With blockchain, the manufacturer can create the first block and verification hash for an individual drug unit. Then, the wholesaler and pharmacy add the second and third blocks, respectively, to the ledger, sequentially verifying the unit using its unique hash. The patient receives the medication from the pharmacy delivery, verifies the unit using a mobile app barcode scanner, and adds the fourth block. Lastly, the HaH provider reviews the record – the chain of blocks – to ensure the integrity of the medication delivery process.

Table 1. Blockchain Use Cases and Technical Considerations for HaH DSCSA Compliance

Use Case	Technical Considerations
HaH Provider to Patient	HaH software platform must be on a DSCSA-compliant blockchain network
Patient to HaH Provider	HaH software platform must be on a DSCSA-compliant blockchain network
	HaH mobile app must have a barcode-scanning camera feature for patients and delivery personnel
	Drug-delivery sensors must use Bluetooth data exchange over the DSCSA-compliant blockchain network

Based on this model, two use cases and their associated technical considerations for blockchain specific to HaH DSCSA compliance were identified (Table 1).

Conclusion

Although more research is needed to identify how blockchain can best support DSCSA compliance, it is unknown whether HaH programs will continue to receive reimbursement support after the PHE. Thus, it is unlikely that rapid deployment of blockchain for pharmaceutical supply chain tracking will occur during the PHE.

References

1. Chien W, et al. The Last Mile: DSCSA Solution through Blockchain Technology. Blockchain and Healthcare Today. 2020;3.
2. The Drug Supply Chain Security Act and Blockchain. Center for Supply Chain Studies. 2018.

OSPred – A Digital Health Aid for Rapid Analysis of Early Endpoints (PFS, ORR) and Overall Survival (OS) Correlates in Non-Small Cell Lung Cancer (NSCLC) Clinical Trials

Youyi Zhang Ph.D.[1], Jiabu Ye Ph.D.[2], Feng Liu Ph.D.[2], Sreenath Nampally MBA[1], Imran Khan, MS[1], Jim Weatherall Ph.D.[3], Cristina Duran CIMA[4], Renee Bailey Iacona Ph.D.[2], Faisal Khan Ph.D.[1], Pralay Mukhopadhyay Ph.D.[2], and Khader Shameer Ph.D.[1*]

[1]Artificial Intelligence & Analytics, Data Science & Artificial Intelligence, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, MD, USA; [2]Biometrics Oncology, Oncology R&D, AstraZeneca, Gaithersburg, MD, USA; [3]Data Science & Artificial Intelligence, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK; [4]Digital R&D, AstraZeneca, Cambridge, UK

**Corresponding Author:* Shamee.Khader@astrazeneca.com

Introduction

In clinical trials that assess novel therapeutic modalities to target NSCLC, early endpoints are evaluated to determine the safety and identify evidence of biological drug activity with endpoints such as Progression-Free Survival (PFS) and Overall Survival (OS). Curating trial-level endpoint data and developing a model that evaluates the correlation between early and late endpoints contribute in detecting novel and early surrogate endpoints, which potentially leads to faster and cost-effective oncology trials. We built an interactive dashboard to visualize the correlation trends across early-to-late endpoints in clinical trials may further empower the clinical teams to rapidly test and validate any hypothesis testing during interim analysis.

Methods

A robust endpoint data set of 81 NSCLC studies, 156 observations on 35 drugs was compiled from multiple public data sources to investigate the correlation between early and late endpoints. Meta-regression model that incorporates both fixed effects and random effects was implemented to measure the correlation between early and late endpoints borrowing strength across different Mechanism of Action (MoA) cohorts. Specifically, the earlier and late endpoints that we considered in our analysis are PFS at 6 months (PFS6m), PFS at 4 months (PFS4m) and Hazard Ratio (HR) of OS. An interactive analysis tool, OSPRED, was developed using R Shiny [1], with required R packages "ggplot2", "metafor", "boot", "dplyr" and "mvtnorm", to visualize the correlation results using historical study and to predict late endpoint with user-defined early endpoint input. Technical details have been fully described in our associated methodology abstract [2]. The user input in our OSPRED interactive platform involves the targeted Mechanism of Action (MoA), pre-calculated PFS odds ratios (OR) and the hazard ratio (HR) at different time points as numeric values. A dropdown list on the left input panel in **Figure 1** allows users to choose a MoA from PDL1, EGFR, VEGFR and DNA [2]. A set of radio buttons allow the user to select PFS odds ratio type (Median PFS, PFS 4 months and PFS 6 months). Key statistics such as Spearman Rank Correlation (ranging [-1,1]), R Square (amount of variance of the outcome accounted by the regressors), I-squared (residual heterogeneity/amount of unaccounted variability in the regression), p-value, predictive value with 95% confidence intervals are displayed. The bottom panel contains two charts, left chart shows the regression plot for HR OS over user-defined early endpoint for historical data and the predicted point of HR OS based on user input (highlighted in green "X"). The chart on the right illustrates our predicted distribution of HR OS with the point predicted value and confidence intervals.

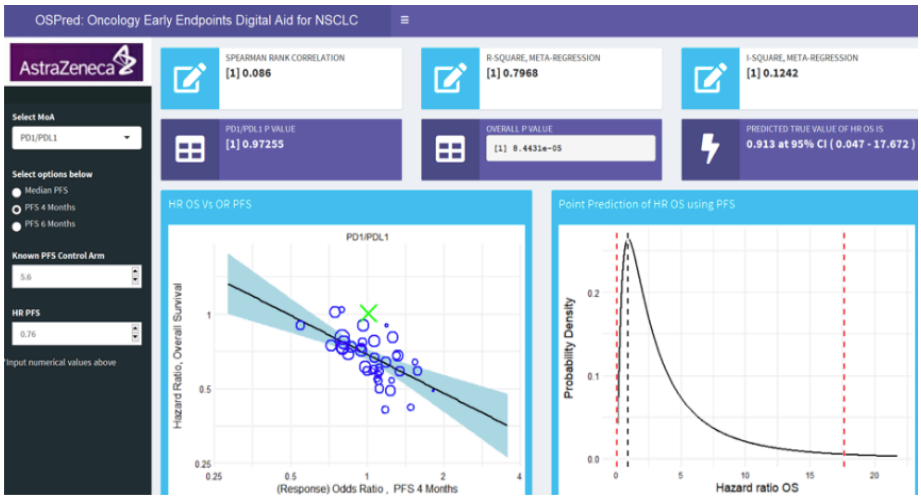


Figure1: OSPred dashboard for interactive analyses of early-to-late endpoints in NSCLC trials

Results

OSPred offers interactive visualization of clinical trial endpoint correlations with reference to a large pool of past NSCLC studies and provides early indications of potential efficacy of the targeted investigational drug with user-defined inputs. Our tool has been applied to several internal studies with the results being referred to when doing interim analysis. Its focused capability has the potential to digitally transform and accelerate decision making with data-driven insights in drug development process.

References

1. https://cran.r-project.org/web/packages/shiny/shiny.pdf
2. https://jitc.bmj.com/content/8/Suppl_3/A398.1

Demonstration of a Self-service De-identified COVID-19 Data Lake

Rajan Chandras MS MSc[1], Stephen B. Johnson PhD [2], Claudia Pulgarin MA MS[2], Megan D. Winner MD MS FACS[3], Chinyere J. Okpara MS[3], Eduardo Iturrate MD[1, 4]
[1]Medical Center Information Technology, NYU Langone Health, New York NY, [2]Department of Population Health, NYU Langone Health, New York NY, [3]NYU Winthrop Hospital, NYU Langone Health, Mineola NY, [4]Department of Medicine, NYU Langone Health, New York NY

Abstract

Clinical informaticists, researchers and information management teams at NYU Langone Health collaborated to quickly create a self-service, de-identified COVID-19 Data Lake in response to a rapidly growing demand from users seeking to understand virus behavior, its impact on clinical processes and outcomes, and care and treatment options. We will describe the steps taken to create the data lake, the challenges encountered, and the outcome and usage, and will demonstrate various features of the data lake.

Description

The COVID-19 pandemic has spurred interest in research on the pathogen in the context of social determinants, clinical practices, treatment options and outcomes. This has created an urgent requirement for relevant COVID data that is de-identified and supports self-service analysis as well as data sharing[1]. To support research and collaboration, the data repository should allow for the integration of diverse data sets, data access controls, data organization and scalable growth[2].

At NYU Langone Health, we created a COVID-19 Data Lake using the Hadoop big data platform that meets all these requirements. Our goal was to establish an institutional resource to democratize data, promote advanced analytics, and accelerate COVID research and innovation. The data lake provides a secure, flexible, scalable, integrated repository for self-service research data management and advanced analytics. The resource enables users to re-identify patients (with appropriate regulatory approval) for enrollment in trials or other longitudinal studies, and provides the means to extend the data with other internal or external data sets in an agile and iterative manner.

As of the date of submission of this proposal, over 150 NYU researchers, informaticists and trainees have applied for use of the data, with about half accessing the database at least once. Reasons cited for use of the repository range from self-learning and exploratory research (hope to identify vulnerable populations) and focused research (explore impact of heart failure on COVID-19 outcomes), to preparation for interventional studies and machine learning and predictive analytics (patterns of COVID-19 disease manifestation/progression using machine learning). The COVID Data Lake also spurred the design and delivery of an introductory clinical informatics training class for medical students and researchers that will enhance the curriculum and provide a foundation for emerging clinical informaticists.

The session will include a presentation using slides on the project and the solution architecture as well as challenges encountered and key findings, followed by a demonstration of the data lake.

Deployment Status

The COVID-19 data lake system has been fully deployed, and continues to be improved iteratively.

Acknowledgments

Satyaki Adusumally MS, Nforce Technologies, Sachin S Ghalme MCA, Clairvoyant, LLC

References

1. Randi E Foraker, Albert M Lai, Thomas G Kannampallil, Keith F Woeltje, Anne M Trolard, Philip R O Payne. Transmission dynamics: Data sharing in the COVID-19 era. Learn Health Syst .2020 Jun 28;e10235
2. Kim Suntae. Functional Requirements for Research Data Repositories. International Journal of Knowledge Content Development & Technology. 2018. Mar, 8(1): 25-36

DataKnots: A Framework for Building Domain Specific Query Languages

Clark C. Evans; Kyrylo Simonov, PhD

DataKnots[1] is an extensible data processing framework. It lets collaborative research teams build their own domain specific query languages (DSQLs) that reflect their conceptual models and vocabularies. DataKnots is based upon an algebraic framework, Query Combinators[2], which formalizes how query components can be defined and combined in a consistent manner. While DataKnots includes standard query operations (group, filter, sum, etc.), they are not given special treatment over domain specific operations. One can add new query components by encapsulating existing queries, lifting Julia language subroutines to queries, and authoring novel transformations using a pipeline construction interface. DataKnots is a platform for creating an ecosystem of mutually interoperable DSQLs, so that collaborative research groups could easily customize their query systems to fit the kinds of data sources and analysis methods they use.

To show that ergonomic, high-performance DSQLs could be rapidly constructed, we built a DSQL inspired by the Clinical Quality Language (CQL) and used it to implement CMS124v7 “Cervical Cancer Screening”. Specifically, we constructed a processing pipeline that converts JSON encoded Fast Healthcare Interoperability Resources (FHIR) to its clinical quality measure (CQM) score. The pipeline loads JSON to an in-memory FHIR representation, converts to an intermediate Quality Data Model (QDM) form, then implements CMS124v7 logic to calculate the CQM score. This layered approach separates concerns, giving us a place to put encoding logic specific to FHIR that doesn’t belong in a CQM, as well as a place to isolate electronic health record vendor differences.

DataKnots4FHIR[3] took 27 working days to implement. New measures can now be added with incremental effort. We benchmarked it using synthetic patient bundles from Synthea. Computation of CMS124v7 over 1,000 patients averages 76ms per patient with a single core on a i7-4770 desktop computer, while the reference implementation[5] averaged 607ms per patient. Our bottleneck is memory usage, with a high-water mark of 11mb per patient. JSON parsing is expensive (17ms/patient). These benchmarks motivate future work on a custom JSON parser, suitable to our vectorized representation, that extracts FHIR lazily based upon the exact fields needed for a given computation.

CQL is a highly-targeted DSQL, created specifically for implementing clinical quality measures and decision logic. Using DataKnots, we were able to rapidly construct a DSQL that matches CQL with comparable functionality and ergonomics. Moreover, we were able to represent not only a CQM calculation, but the entire processing pipeline, including conversion from JSON to FHIR and conversion of FHIR to QDM. For this application, DataKnots was extended with relevant data types, including concepts and value-sets from clinical vocabularies. Informed by clinical quality measure guidelines, we also defined a datetime interval with operators, such as **`and_previous`** and **`during`**. Critically, these domain specific operators are treated no differently from built-in operations such as **`filter`**. Because its formalized approach permits new query operators to be seamlessly integrated, DataKnots can be used to build distinct DSQLs, each having a conceptual model and vocabulary responsive to its research domain and audience.

```
define "PapTest Within 5 Years"
 ("Pap Test with Results" PapTestOver30YearsOld
  with ["Patient Characteristic Birthdate"] BirthDate
    such that Global."CalendarAgeInYearsAt"(
           BirthDate.birthDatetime,
           start of PapTestOver30YearsOld.relevantPeriod) ≥ 30
    and PapTestOver30YearsOld.relevantPeriod 5 years or
                less before end of "Measurement Period")
```

CMS124v7 fragment using CQL with QDM

https://ecqi.healthit.gov/sites/default/files/ecqm/measures/CMS124v7.html

```
@define PapTestWithin5Years =
  let birthDate ⇒ PatientCharacteristicBirthdate.
                    BirthDateTime,
      previous5years ⇒ interval(MeasurePeriod.end).
                         and_previous(5years)
    PapTestWithResults.
    filter(years_between(relevantPeriod.start, birthDate) ≥ 30
          && relevantPeriod.during(previous5years))
 end
```

an equivalent using DataKnots – cms124.jl

https://github.com/rbt-lang/DataKnots4FHIR.jl/blob/master/doc/src/cms124v7.jl

DataKnots is MIT/Apache licensed, is well documented, and has extensive regression tests. Our approach has multiple applications: we have additionally prototyped a DSQL for Observational Health Data Sciences and Informatics (OHDSI) cohort construction[4]. We are actively searching for a pilot project.

1. Evans CC, Simonov K, DataKnots Query System for Julia (https://github.com/rbt-lang/DataKnots.jl/)
2. Evans CC, Simonov K, Query Combinators (https://arxiv.org/abs/1702.08409)
3. Evans CC, DataKnots4FHIR : Query Adapters for FHIR (https://github.com/rbt-lang/DataKnots4FHIR.jl/)
4. Evans CC, Simonov K, DSQLs for Medical Research (https://www.biorxiv.org/content/10.1101/737619v2)
5. Database Consulting Group, CQL Measure Processing Component (https://github.com/DBCG/cqf-ruler)

The Pediatric Cancer Data Commons

A demonstration of a novel implementation and extension of the Gen3 infrastructure for cohort discovery and data sharing

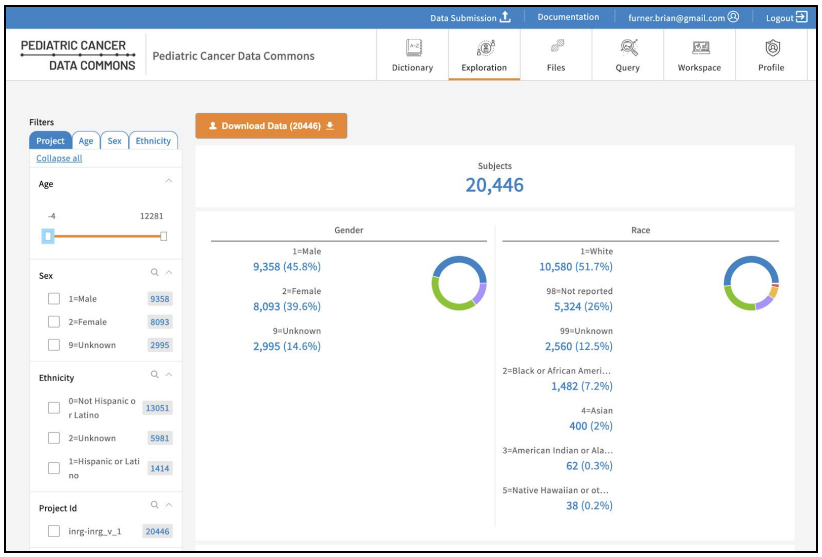
Luca Graglia, MS, MBA, Shazia Sathar, MS, Monica Palese, MPH,
Brian Furner, MS, Samuel Volchenboum, MD, PhD
University of Chicago, Chicago, IL

Description

Purpose: Data commons have become an essential way to collect and share research data, as they can obviate the need for large data downloads and expensive and redundant tool creation, while facilitating interoperability through standardization. The Gen3 platform, created for the implementation of the Genomic Data Commons, is a freely-available infrastructure that supports data model integration and faceted search of large sets of data. With the growth of the National Cancer Institute's Cancer Research Data Commons (CRDC), platforms built on Gen3 may be more easily integrated and interoperable.

Problems it addresses: Access to data for pediatric cancer research is hampered by the paucity of cases and the lack of interoperability between all types of data - clinical, genomic, imaging, and others. Further, there are few available platforms for cohort discovery, data fulfillment, and visualization and analytics. Allowing researchers to connect disparate data sources creates opportunities to make new discoveries. The development and deployment of the Gen3 infrastructure and associated new microservices and features solves these challenges and lowers the bar for researchers around the world. By making all code freely available via an open source license, innovations can be adopted and integrated by other research groups.

Features that make it innovative: Development and deployment of a Gen3 instance requires a high level of technical skill and expertise, and if extensions to basic functionality are required, a multi-talented team of programmers, front-end developers, and experts in data standards, security, and regulation. The Pediatric Cancer Data Commons (PCDC) team at the University of Chicago is leveraging the Gen3 infrastructure for deployment of data commons for multiple childhood malignancies. This is the first example of a centralized platform for data search and sharing for pediatric cancer that leverages international consensus data standards. Further, users have access to analytic tools and visualizations that facilitate hypothesis generation. Additionally, since universal identifiers are leveraged by different nodes in the CRDC ecosystem and other external data sources, it is possible to create extended data sets which include multiple data types.



Current State

Mature, internationally-balloted data dictionaries have been created for all the major pediatric cancer types. Several data commons, including the neuroblastoma and pediatric rhabdomyosarcoma are currently deployed in a Gen3 staging environment with go-live expected by the end of the calendar year. Current funding is supporting the development of these Gen3 commons with extensions of functionality to include discrete authorization, harmonized data ingestion, faceted search, data visualization, and research request submission and fulfillment. The demonstration will include cohort discovery, data model illustration, authentication and authorization, project request functionality, and analytics and visualization.

Gaining Enriched Insight into Patient Populations by Appending Claims Data to National Health Survey Responses and a Market Segmentation System

Brandi Hodor, BS[1], Elisabeth Scheufele, MD, MS[1], George Popa, Jr., MS, MHSA[1], William Kassler, MD, MPH[1]
[1]IBM Watson Health, Cambridge, MA

Description of the Demo

Optimizing the health of a population requires further insight into the lives of patients outside of what is available from clinical and administrative data. Socioeconomic status, access to health care, social needs such as food scarcity, and neighborhood issues such as safety are non-clinical determinants of health, and affect a patient's ability to take care of themselves, and their loved ones, and to manage their medical concerns. Knowledge regarding the social and environmental aspects of patients' lives are not typically available at the point of care, or to policy-makers developing population level strategies. PULSE/PRIZM, an appended dataset of responses from a national health and health behavior survey with an extensive market segmentation system, can provide these types of insights not typically available to the medical provider or public health entity.

The PULSE™ Survey is an annual US national health care and health behavior survey that conducts interviews on numerous topics, including health care access and utilization, social determinants of health, access to technology, health seeking behavior and personal perspectives on healthcare. The annual survey comprises approximately 100 questions that are revised yearly, covering at least 80,000 respondents each year. The survey data is combined with PRIZM®, a market segmentation system, built on multiple public and private data sources including the US Census, data on educational level, socio-economic details, and local purchasing tendencies. Market segmentation systems have allowed the commercial market to focus marketing and sales at the hyperlocal, market segment level. By taking this same approach, this dataset has significant potential in applying these insights into identifying attitudes, lifestyles, and behaviors around barriers and uncovering gaps to medical care or improving health. Appending PRIZM data to PULSE responses at the block group level provides an enriched dataset where PULSE responses are localized to market segment level and insights on health and health behaviors are linked to socioeconomic and educational data to provide a more complete picture of the population investigated.

Further population detail is achieved when linking the health processes and outcomes found in medical claims data to the PULSE/PRIZM dataset which can provide hyper-localized sociodemographic and behavioral details to patients in the claims data. The health survey responses and the market segmentation insights augment the population represented in the claims data, and helps provide peer level location-based insight so entities of interest can use the dataset to identify opportunities or deficiencies in medical or health care and then identify potential means by which to bridge gaps.

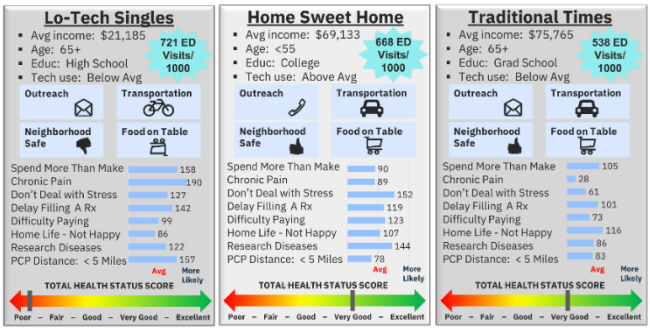


Figure 1: ED Utilization Use Case using PULSE/PRIZM and claims data

This demonstration will describe these datasets in detail and apply them to real-world use cases. The presentation will describe the PULSE health and health behavior survey, the PRIZM market segmentation model, and the dataset that is created when these are appended at the block group level. One use case involves the preventive care practice of mammograms; using the appended dataset to identify a location where mammograms are underperformed and to identify the potential reasons for gaps and potential means to bridge them. Another use case (Figure 1), involving the claims appended dataset, investigates emergency department utilization, identified from the claims data. The hyperlocalized dataset will be used to uncover possible reasons of overutilization in a patient population, then identify opportunities by which to develop a strategy to decrease inappropriate ED overutilization.

A statement of the degree to which the system or service has been deployed, as of the date of submitting the proposal: The PULSE/PRIZM dataset has been active and applied to health-related strategies for the last 30 years.

Synchronized Coordination of The ACT Network to Rapidly Identify COVID-19 Patients in an Evolving Global Crisis

Anupama Maram, MS[1], Griffin M. Weber, MD, PhD[1], Philip Trevvett[1]
[1]Harvard Medical School, Boston, MA

Introduction

The COVID-19 pandemic caused by SARS-CoV-2 remains a significant and impactful issue for global health, economics and society. A plethora of information and data has been generated and disseminated since its emergence in December 2019, and it is mission critical for researchers to keep up with this data from across the world at a time of uncertainty and constantly evolving guidelines and clinical practice. Unlike prior pandemics, national informatics infrastructure has played an indispensable role in the response to the novel coronavirus COVID-19 pandemic. Leveraging open source informatics tools and The ACT Network has allowed researchers to rapidly respond to evolving requirements and accelerate timelines in order to share data, locate patient cohorts, and study disease prevalence. The Accrual to Clinical Trials (ACT) network is a nationwide federation of Clinical and Translational Science Award (CTSA) institutions that share aggregate patient counts from electronic health record data. The network consists of local installations of Informatics for Integrating Biology at the Bedside (i2b2) EHR data repositories that are linked by the Shared Health Research Information Network (SHRINE) platform. The SHRINE platform includes a web-based query tool that allows researchers to construct complex Boolean queries to obtain real time aggregate count of patients at participating hospitals who meet a given set of inclusion and exclusion criteria. Because of the national scope of the ACT network, researchers have access to patient sets with regional diversity helping with clinical trial cohort discovery and study feasibility. To date, the network connects over 45 CTSA sites and contains data on more than 125 million patients[1]. The SHRINE user interface (UI) was recently updated with a more intuitive, user-friendly UI with modern usability standards of design, look-and-feel, and accessibility. The coordinated release of frequent data refreshes, new medical coding criteria, and features have enabled researchers to access just in time updates in order to address this national crisis.

Discussion

The expansion of the user community driven by continued growth of the ACT network to over half of the CTSA consortium, and the availability of a large number of patient records, necessitates a new UI and user experience for SHRINE/ACT to be as intuitive as possible for novice users while conveying complex query construction and eliminating the need for extensive training. By chance, the evolving pandemic coincided with the release of the new interface. To better serve the research community, the ACT Operations team coordinated releases among the i2b2, SHRINE, and ACT development teams. New ontology development efforts to incorporate specialized COVID-19 ontology specific for COVID-19 phenotype were deployed on a frequent basis to address the changing terminology including diagnosis and LOINC concepts. To help researchers identify cohorts with changing clinical definitions, the team modified the UI to locate emerging ICD-10, CPT, HCPCS, LOINC codes, incorporate existing codes, and display newly created derived terms of particular interest in COVID-19 research such as illness severity, mechanical ventilation in a dedicated space. As COVID-19 clinical data flows into local sites, limited data sets (aggregate counts) are accessed in the network, aided by frequent data refreshes to ensure up to date information. This coordination efforts extends across the sites of the network, allowing researchers of the ACT Network to query the total numbers of patients at each participating site meeting the inclusion or exclusion criteria for demographics (age, gender, race, etc.), diagnoses (ICD9/10 codes), lab results, and most frequently prescribed medications in just under a few minutes.

Conclusion

A wealth of data has already been generated on COVID-19 since early January 2020. Nevertheless, key questions remain regarding understanding at-risk populations, disease transmission, and progression. Leveraging the ACT Network with newly added medical concepts, a flexible and responsive UI, and rapid development of features has equipped researchers to swiftly identify patient populations specific to the COVID-19 pandemic.

References

1. The ACT Network. [Internet]. 2020 Aug 27. Available from http://www.actnetwork.us/national/faqs-46EU-1377S2.html

Funded by the NIH National Center for Advancing Translational Sciences grant numbers UL1TR002541, UL1TR000005.

Deploying i2b2 as study-specific application for clinical data analysis

Kavishwar B. Wagholikar MBBS PhD[1,3,5], Layne Ainsworth MBA[5], Nilesh Keriwala [4], Akshay Zagade BE[4], Rupendra Chulyadyo [5], Sachin Wakle BE[4], Swati Shevade [4], Ashutosh Wakhure[4], Sadanand Walte [4], Yuri Ostrovsky PhD [4], Kira Chaney MPH [3], David Zelle MS [3], Angela Miller MA[3], Michael Oates[5], Samuel J. Aronson ALM, MA[5], Benjamin M. Scirica MPH MD[1,2], Shawn N. Murphy MD PhD[1,3]

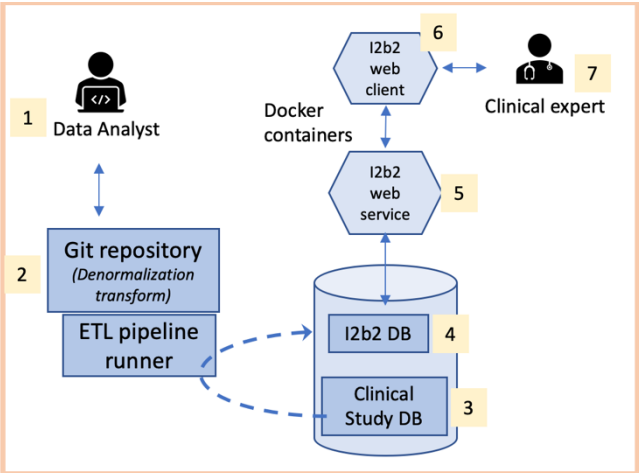
[1]Harvard Medical School, Boston, MA;
[2]Brigham and Women's Hospital, Boston, MA; [3]Massachusetts General Hospital, Boston, MA ;
[4] Persistent Systems, Pune, India; [5] Mass General Brigham, Boston, MA

Background

Reporting on clinical studies typically requires effort from a data analyst to decipher the study-database's schema to develop SQL queries for generating counts of patients that meet particular clinical-criteria. However considerable domain knowledge is often required to implement the clinical-criteria using structured query language (SQL), and hence the data-analyst is often paired with a clinical-expert who can explain the criteria in terms of the elements available in the database. This collaboration often requires weeks or months to yield a set of SQL statements that produce the desired analysis. However, the SQL statements are only amenable to modification or troubleshooting by the data-analyst— the clinical expert cannot modify the SQL queries due to unfamiliarity with the SQL syntax, which renders it difficult to independently validate the results or to perform additional analysis.

Methodology and Design

We have developed an alternative approach wherein the data-analyst de-normalizes the study-database into an i2b2 like format rather than synthesizing the data elements into aggregate counts. The resulting de-normalized form is then imported into i2b2 and the i2b2-webclient is used by the clinical expert to auto-generate the SQL queries for the analysis.



This approach inherently entails that the project database be denormalized and loaded into i2b2, using the extensions that we have developed to,
i) create and load custom ontologies and
ii) import the study-data into i2b2 on a daily basis.

Easy install of i2b2 is facilitated by using docker containers. [1] Figure 1 shows the system architecture that we will demonstrate in this presentation. The goal of the system is to augment the validity and reproducibility of the study analysis by autogenerating the SQL and enabling the clinical staff to directly query the data. Our system has been deployed in the production setting for a clinical study. One limitation of our system currently is the lack of graphical visualization of the results.

Figure 1. The data-analyst creates data-transforms to de-normalize the study data into i2b2 format, that is queried by clinician using the web-client. Docker containers are used to deploy the Extract-Transform and Load (ETL) pipeline and the i2b2 web-client and web-services.

References

1. Wagholikar KB, Dessai P, Sanz J, Mendis ME, Bell DS, Murphy SN. Implementation of informatics for integrating biology and the bedside (i2b2) platform as Docker containers. BMC Med Inform Decis Mak 2018;18:66.