# Collaborator Showcase Presentations

## OHDSI Community Call

## Nov. 8, 2022 • 11 am ET

# Upcoming OHDSI Community Calls

| Date | Topic |
|---|---|
| Nov. 15 | Open Network Studies |
| Nov. 22 | 10-Minute Tutorials |
| Nov. 29 | Workgroup Updates |
| Dec. 6 | Fall Publications |
| Dec. 13 | How Did We Do In 2022? |
| Dec. 20 | Holiday-Themed Final Call of 2022 |

# Upcoming OHDSI Community Calls

| Date | Topic |
|---|---|
| Nov. 15 | Open Network Studies |
| Nov. 22 | 10-Minute Tutorials |
| Nov. 29 | Workgroup Updates |
| Dec. 6 | Fall Publications |
| Dec. 13 | How Did We Do In 2022? |
| Dec. 20 | Holiday-Themed Final Call of 2022 |

# Three Stages of The Journey

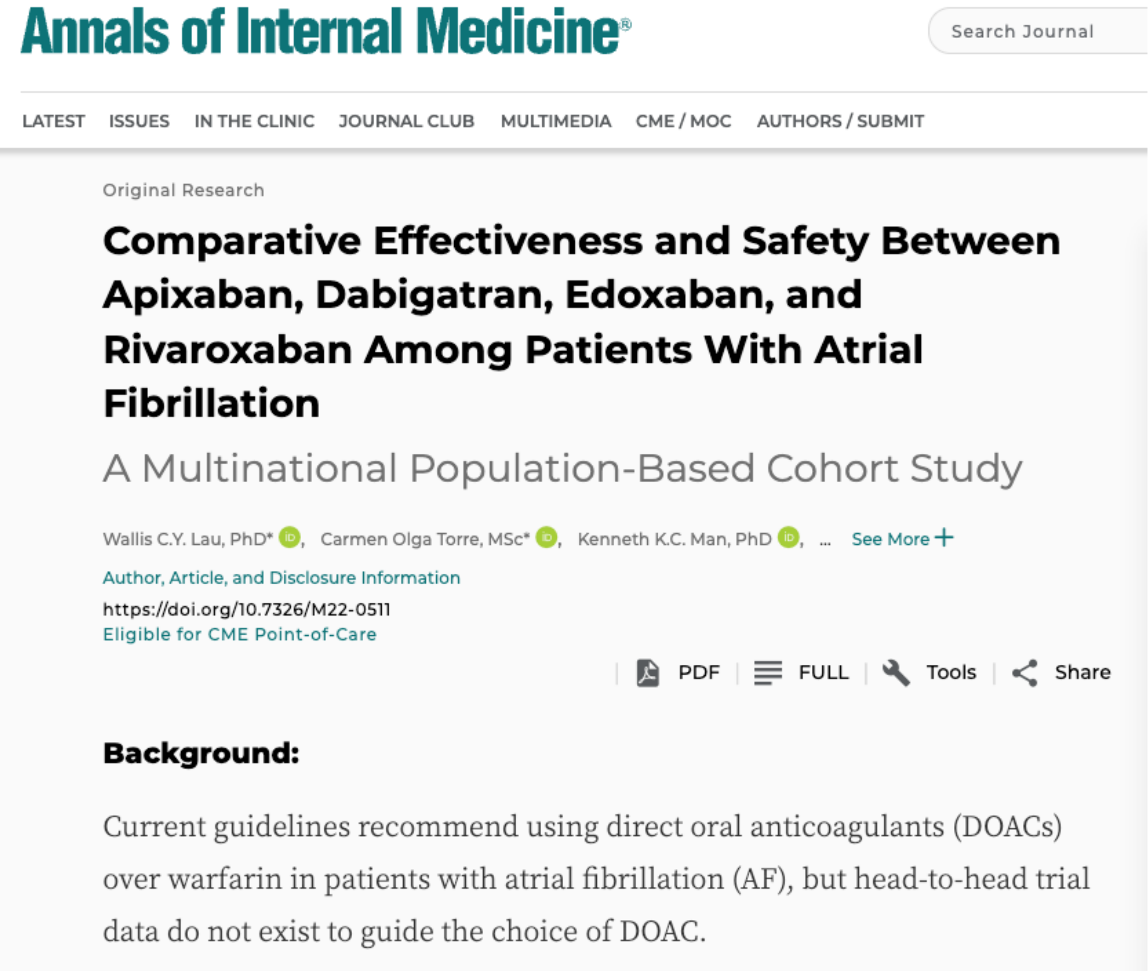
**Where Have We Been?**

**Where Are We Now?**

**Where Are We Going?**

# OHDSI Shoutouts!

Congratulations to the team of **Wallis C.Y. Lau, Carmen Olga Torre, Kenneth K.C. Man, Henry Morgan Stewart, Sarah Seager, Mui Van Zandt, Christian Reich, Jing Li, Jack Brewster, Gregory Y.H. Lip, Aroon D. Hingorani, Li Wei, and Ian C.K. Wong** on the publication of **Comparative Effectiveness and Safety Between Apixaban, Dabigatran, Edoxaban, and Rivaroxaban Among Patients With Atrial Fibrillation** in Annals of Internal Medicine.



Original Research

**Comparative Effectiveness and Safety Between Apixaban, Dabigatran, Edoxaban, and Rivaroxaban Among Patients With Atrial Fibrillation**

A Multinational Population-Based Cohort Study

Wallis C.Y. Lau, PhD*, Carmen Olga Torre, MSc*, Kenneth K.C. Man, PhD, ... See More +

Author, Article, and Disclosure Information

https://doi.org/10.7326/M22-0511

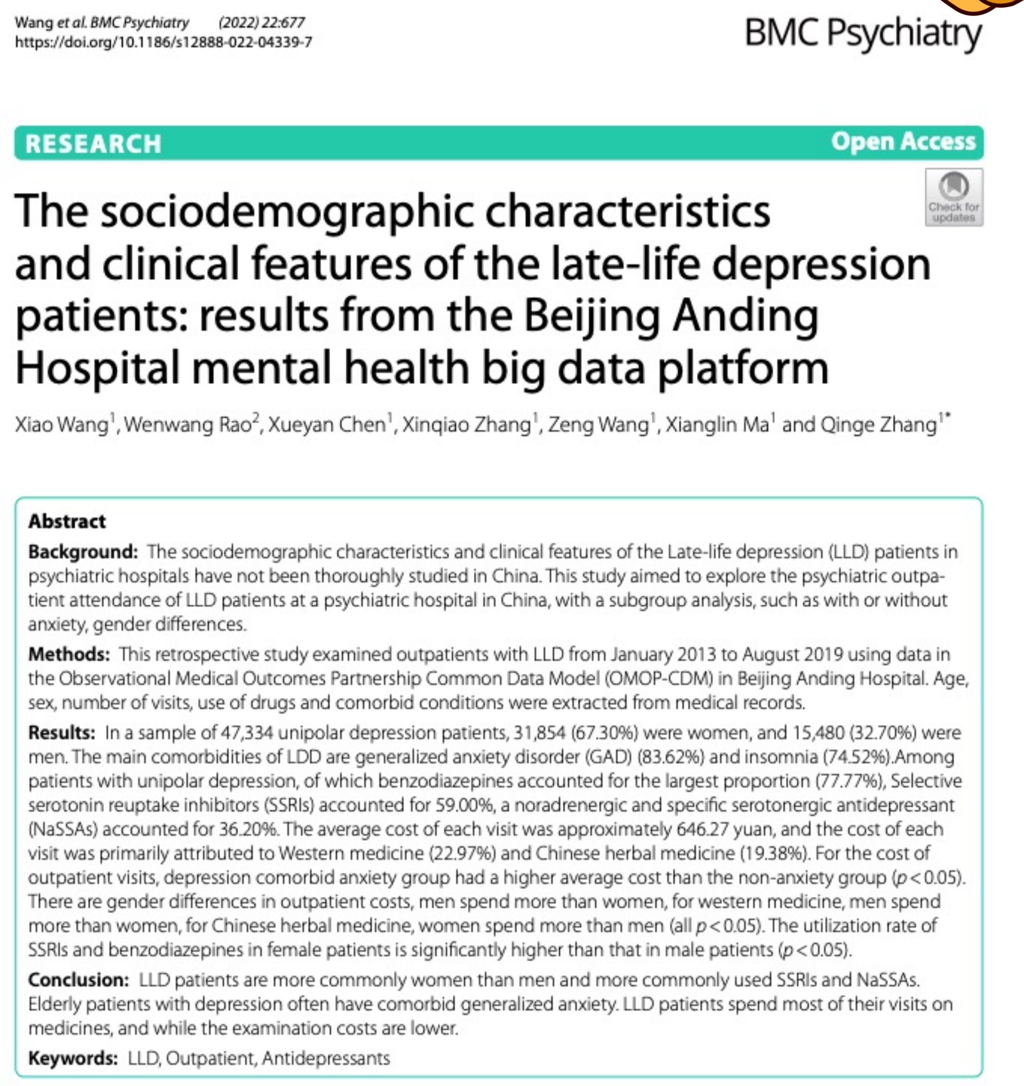
Eligible for CME Point-of-Care

**Background:**

Current guidelines recommend using direct oral anticoagulants (DOACs) over warfarin in patients with atrial fibrillation (AF), but head-to-head trial data do not exist to guide the choice of DOAC.

# OHDSI Shoutouts!

Congratulations to the team of **Xiao Wang, Wenwang Rao, Xueyan Chen, Xinqiao Zhang, Zeng Wang, Xianglin Ma, Qinge Zhan** on the publication of **The sociodemographic characteristics and clinical features of the late-life depression patients: results from the Beijing Anding Hospital mental health big data platform** in BMC Psychiatry.



Wang *et al. BMC Psychiatry* (2022) 22:677
https://doi.org/10.1186/s12888-022-04339-7

BMC Psychiatry

RESEARCH Open Access

## The sociodemographic characteristics and clinical features of the late-life depression patients: results from the Beijing Anding Hospital mental health big data platform

Xiao Wang[1], Wenwang Rao[2], Xueyan Chen[1], Xinqiao Zhang[1], Zeng Wang[1], Xianglin Ma[1] and Qinge Zhang[1*]

**Abstract**

**Background:** The sociodemographic characteristics and clinical features of the Late-life depression (LLD) patients in psychiatric hospitals have not been thoroughly studied in China. This study aimed to explore the psychiatric outpatient attendance of LLD patients at a psychiatric hospital in China, with a subgroup analysis, such as with or without anxiety, gender differences.

**Methods:** This retrospective study examined outpatients with LLD from January 2013 to August 2019 using data in the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) in Beijing Anding Hospital. Age, sex, number of visits, use of drugs and comorbid conditions were extracted from medical records.

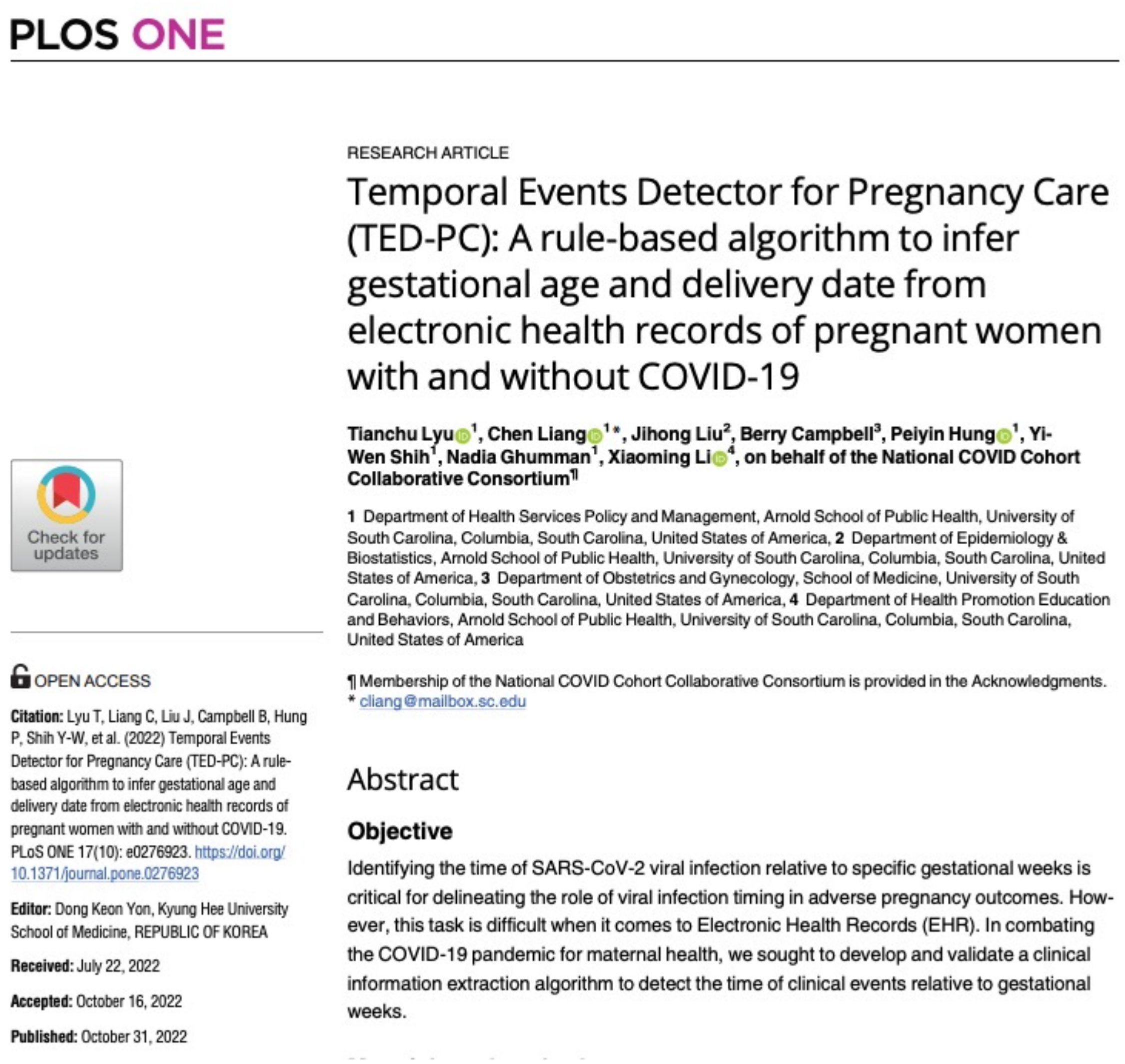
**Results:** In a sample of 47,334 unipolar depression patients, 31,854 (67.30%) were women, and 15,480 (32.70%) were men. The main comorbidities of LDD are generalized anxiety disorder (GAD) (83.62%) and insomnia (74.52%).Among patients with unipolar depression, of which benzodiazepines accounted for the largest proportion (77.77%), Selective serotonin reuptake inhibitors (SSRIs) accounted for 59.00%, a noradrenergic and specific serotonergic antidepressant (NaSSAs) accounted for 36.20%. The average cost of each visit was approximately 646.27 yuan, and the cost of each visit was primarily attributed to Western medicine (22.97%) and Chinese herbal medicine (19.38%). For the cost of outpatient visits, depression comorbid anxiety group had a higher average cost than the non-anxiety group ($p < 0.05$). There are gender differences in outpatient costs, men spend more than women, for western medicine, men spend more than women, for Chinese herbal medicine, women spend more than men (all $p < 0.05$). The utilization rate of SSRIs and benzodiazepines in female patients is significantly higher than that in male patients ($p < 0.05$).

**Conclusion:** LLD patients are more commonly women than men and more commonly used SSRIs and NaSSAs. Elderly patients with depression often have comorbid generalized anxiety. LLD patients spend most of their visits on medicines, and while the examination costs are lower.

**Keywords:** LLD, Outpatient, Antidepressants

# OHDSI Shoutouts!

Congratulations to the team of **Tianchu Lyu, Chen Liang, Jihong Liu, Berry Campbell, Peiyin Hung, Yi-Wen Shih, Nadia Ghumman, Xiaoming Li, and members of the National COVID Cohort Collaborative Consortium** on the publication of **Temporal Events Detector for Pregnancy Care (TED-PC): A rule-based algorithm to infer gestational age and delivery date from electronic health records of pregnant women with and without COVID-19** in PLOS ONE.



PLOS ONE

RESEARCH ARTICLE

## Temporal Events Detector for Pregnancy Care (TED-PC): A rule-based algorithm to infer gestational age and delivery date from electronic health records of pregnant women with and without COVID-19

**Tianchu Lyu[1], Chen Liang[1]*, Jihong Liu[2], Berry Campbell[3], Peiyin Hung[1], Yi-Wen Shih[1], Nadia Ghumman[1], Xiaoming Li[4], on behalf of the National COVID Cohort Collaborative Consortium¶**

**1** Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America, **2** Department of Epidemiology & Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America, **3** Department of Obstetrics and Gynecology, School of Medicine, University of South Carolina, Columbia, South Carolina, United States of America, **4** Department of Health Promotion Education and Behaviors, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, United States of America

¶ Membership of the National COVID Cohort Collaborative Consortium is provided in the Acknowledgments.
* cliang@mailbox.sc.edu

Check for updates

OPEN ACCESS

**Citation:** Lyu T, Liang C, Liu J, Campbell B, Hung P, Shih Y-W, et al. (2022) Temporal Events Detector for Pregnancy Care (TED-PC): A rule-based algorithm to infer gestational age and delivery date from electronic health records of pregnant women with and without COVID-19. PLoS ONE 17(10): e0276923. https://doi.org/10.1371/journal.pone.0276923

**Editor:** Dong Keon Yon, Kyung Hee University School of Medicine, REPUBLIC OF KOREA

**Received:** July 22, 2022

**Accepted:** October 16, 2022

**Published:** October 31, 2022

## Abstract

### Objective

Identifying the time of SARS-CoV-2 viral infection relative to specific gestational weeks is critical for delineating the role of viral infection timing in adverse pregnancy outcomes. However, this task is difficult when it comes to Electronic Health Records (EHR). In combating the COVID-19 pandemic for maternal health, we sought to develop and validate a clinical information extraction algorithm to detect the time of clinical events relative to gestational weeks.

# OHDSI Shoutouts!

# OHDSI Shoutouts!

Dr. Thamir Alshammari
@T_M_Alshammari

Well we'll deserved George, it is always great to work and learn from you.
@OHDSI
@ColumbiaDBMI

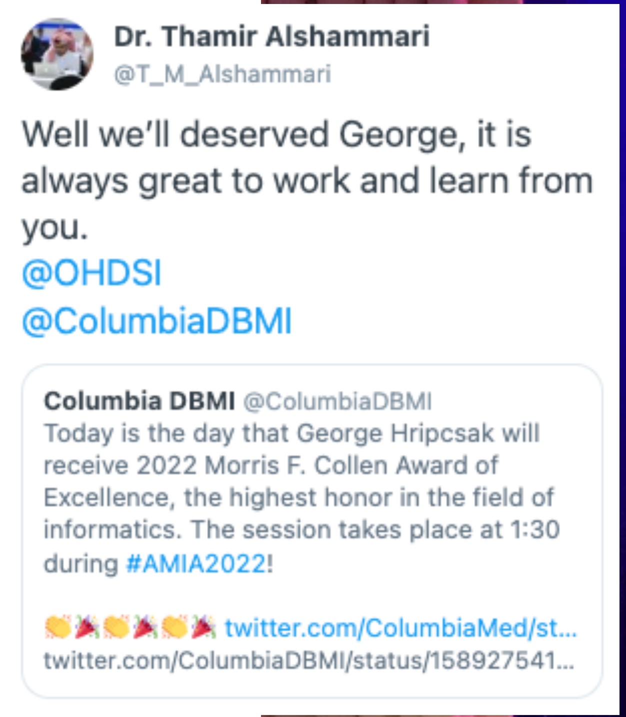
Columbia DBMI @ColumbiaDBMI
Today is the day that George Hripcsak will receive 2022 Morris F. Collen Award of Excellence, the highest honor in the field of informatics. The session takes place at 1:30 during #AMIA2022!

👏🎉👏🎉👏🎉 twitter.com/ColumbiaMed/st...
twitter.com/ColumbiaDBMI/status/158927541...

Kristin Kostka @kricketchirps 15h
The most humble leader I know! So awesome to see our fearless @OHDSI leader recognized. I know I wouldn't be where I am today without George's mentorship! He's a true star in the observational health data science and informatics field. 💫

Columbia DBMI @ColumbiaDBMI
Congrats to DBMI chair George Hripcsak on receiving the 2022 Morris F. Collen Award of Excellence yesterday! #AMIA2022 @Columbia @ColumbiaPS @DataSciColumbia @AMIAinformatics

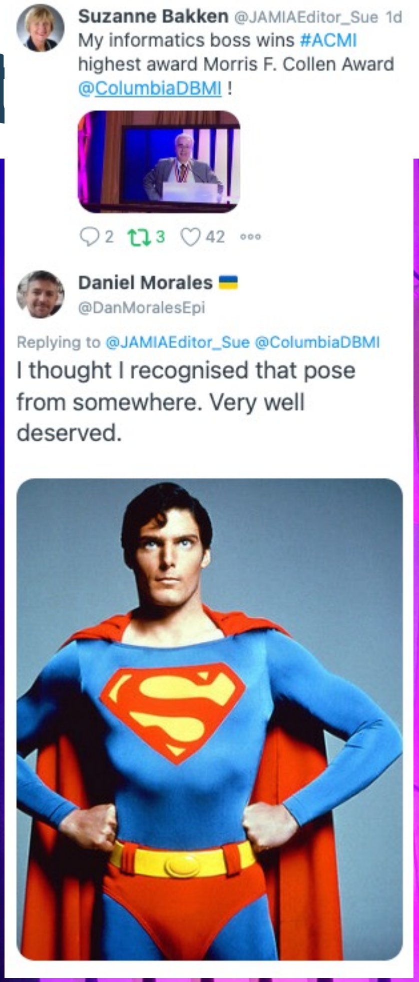
# OHDSI Shout



**Dr. Thamir Alshammari**
@T_M_Alshammari

Well we'll deserved George, it is always great to work and learn from you.
@OHDSI
@ColumbiaDBMI

> **Columbia DBMI** @ColumbiaDBMI
> Today is the day that George Hripcsak will receive 2022 Morris F. Collen Award of Excellence, the highest honor in the field of informatics. The session takes place at 1:30 during #AMIA2022!
>
> 👏🎉👏🎉👏🎉 twitter.com/ColumbiaMed/st...
> twitter.com/ColumbiaDBMI/status/158927541...

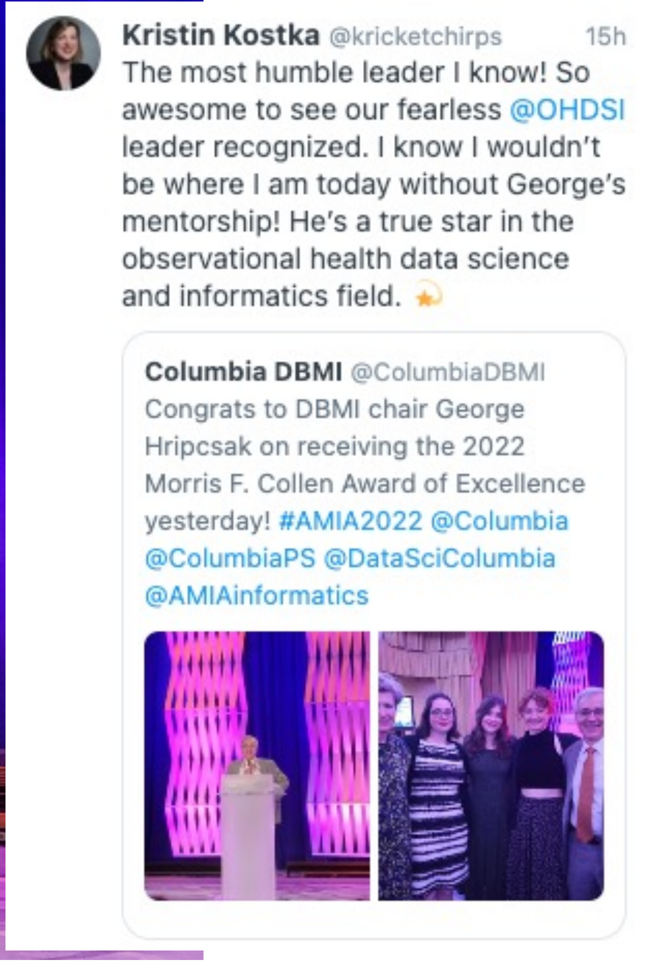


**Suzanne Bakken** @JAMIAEditor_Sue 1d
My informatics boss wins #ACMI highest award Morris F. Collen Award @ColumbiaDBMI !

2 3 42

**Daniel Morales** 🇺🇦
@DanMoralesEpi

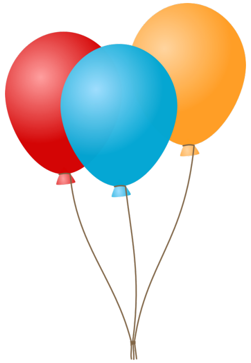
Replying to @JAMIAEditor_Sue @ColumbiaDBMI
I thought I recognised that pose from somewhere. Very well deserved.



**Kristin Kostka** @kricketchirps 15h
The most humble leader I know! So awesome to see our fearless @OHDSI leader recognized. I know I wouldn't be where I am today without George's mentorship! He's a true star in the observational health data science and informatics field. 💫

> **Columbia DBMI** @ColumbiaDBMI
> Congrats to DBMI chair George Hripcsak on receiving the 2022 Morris F. Collen Award of Excellence yesterday! #AMIA2022 @Columbia @ColumbiaPS @DataSciColumbia @AMIAinformatics

# OHDSI Shoutouts!



**Any shoutouts from the community? Please share and help promote and celebrate OHDSI work!**

Have a study published? Please send to sachson@ohdsi.org so we can share during this call and on our social channels.
Let’s work together to promote the collaborative work happening in OHDSI!

# Three Stages of The Journey

**Where Have We Been?**

**Where Are We Now?**

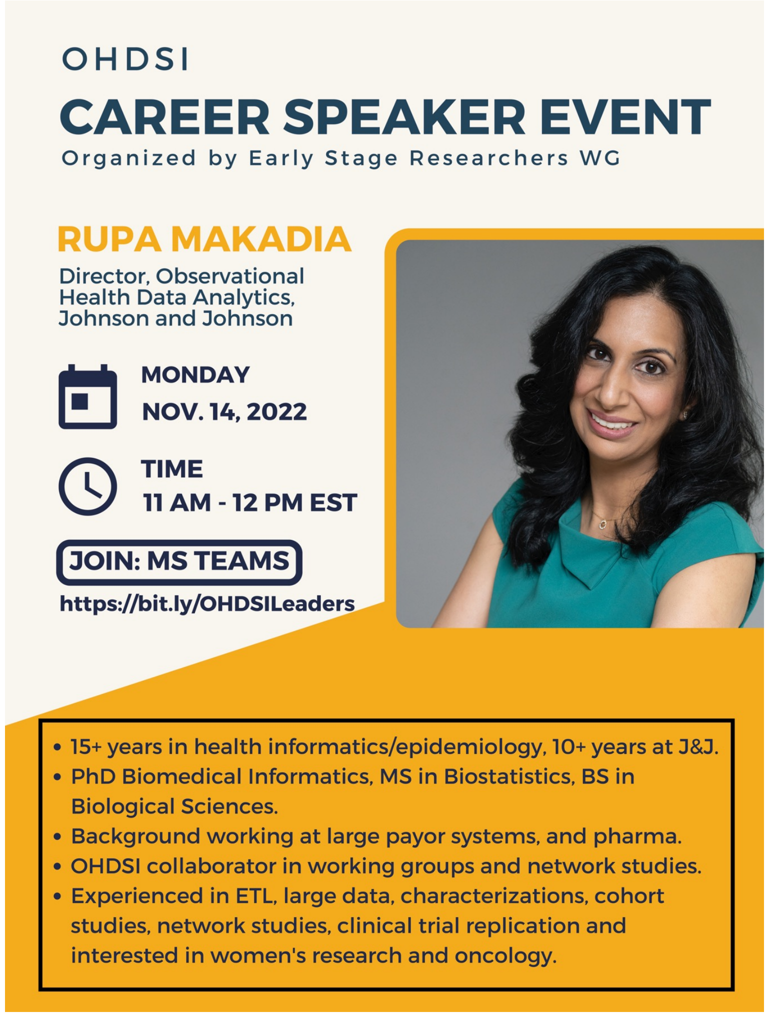
**Where Are We Going?**

# Upcoming Workgroup Calls

| Date | Time (ET) | Meeting |
|---|---|---|
| Tuesday | 12 pm | Common Data Model Vocabulary Subgroup |
| Tuesday | 6 pm | Eyecare & Vision Research |
| Wednesday | 9 am | Patient-Level Prediction |
| Wednesday | 10 am | FHIR and OMOP Digital Quality Measurements Subgroup (ZOOM) |
| Wednesday | 11 am | Open-Source Community |
| Wednesday | 2 pm | Natural Language Processing |
| Thursday | 12 pm | FHIR and OMOP Oncology Subgroup |
| Thursday | 7 pm | Dentistry |
| Friday | 9 am | GIS – Geographic Information System Development |
| Friday | 9 am | Phenotype Development & Evaluation |
| Friday | 10:15 am | Clinical Trials |
| Friday | 10 pm | China Chapter |
| Monday | 10 am | Africa Chapter |
| Monday | 11 am | Early-Stage Researchers |

ohdsi.org/upcoming-working-group-calls/

# Early-Stage Researchers Career Speaker Series



OHDSI

## CAREER SPEAKER EVENT

Organized by Early Stage Researchers WG

**RUPA MAKADIA**

Director, Observational Health Data Analytics, Johnson and Johnson

MONDAY
NOV. 14, 2022

TIME
11 AM - 12 PM EST

JOIN: MS TEAMS
https://bit.ly/OHDSILeaders

- 15+ years in health informatics/epidemiology, 10+ years at J&J.
- PhD Biomedical Informatics, MS in Biostatistics, BS in Biological Sciences.
- Background working at large payor systems, and pharma.
- OHDSI collaborator in working groups and network studies.
- Experienced in ETL, large data, characterizations, cohort studies, network studies, clinical trial replication and interested in women's research and oncology.

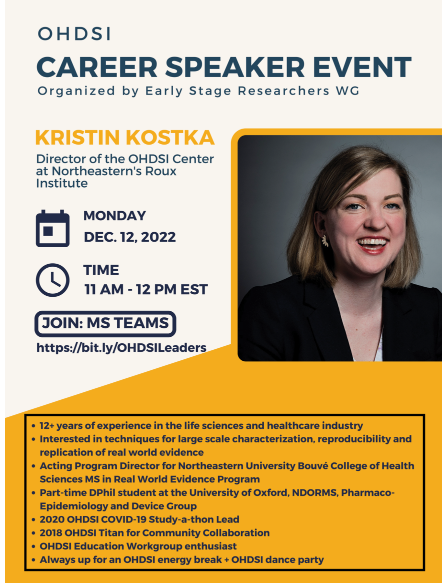
**August: Asieh Golozar** ▶

**September: Jenna Reps** ▶

**October: Jenny Lane** ▶

**November: Rupa Makadia (Nov. 14, 11 am ET)**

**December: Kristin Kostka (Dec. 12, 11 am ET)**



OHDSI

## CAREER SPEAKER EVENT

Organized by Early Stage Researchers WG

**KRISTIN KOSTKA**

Director of the OHDSI Center at Northeastern's Roux Institute

MONDAY
DEC. 12, 2022

TIME
11 AM - 12 PM EST

JOIN: MS TEAMS
https://bit.ly/OHDSILeaders

- 12+ years of experience in the life sciences and healthcare industry
- Interested in techniques for large scale characterization, reproducibility and replication of real world evidence
- Acting Program Director for Northeastern University Bouvé College of Health Sciences MS in Real World Evidence Program
- Part-time DPhil student at the University of Oxford, NDORMS, Pharmaco-Epidemiology and Device Group
- 2020 OHDSI COVID-19 Study-a-thon Lead
- 2018 OHDSI Titan for Community Collaboration
- OHDSI Education Workgroup enthusiast
- Always up for an OHDSI energy break + OHDSI dance party

**bit.ly/OHDSILeaders**

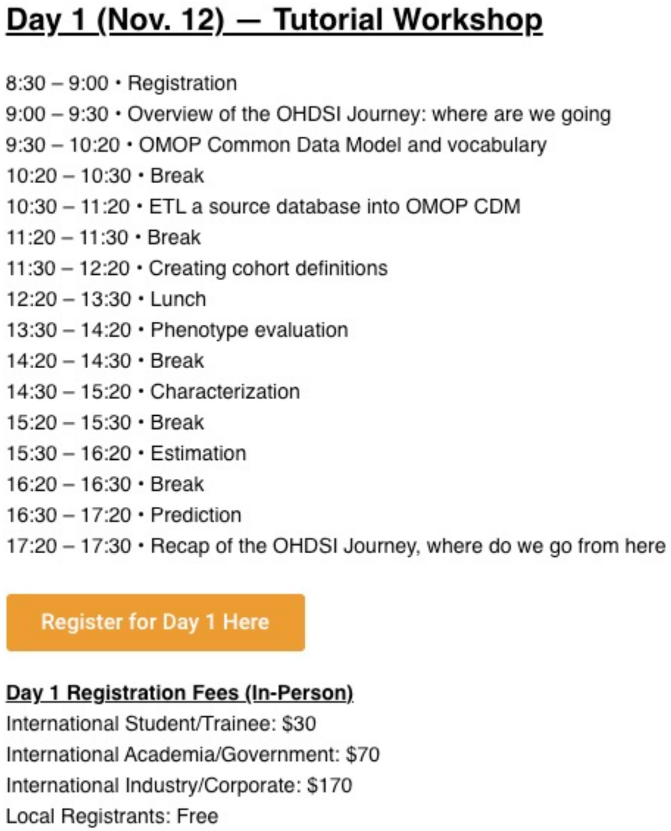
# Open-Source Community WG Meeting

Please join the next Open-Source Community WG meeting, which will include a presentation from **Laurie Arp** around supporting open-source sustainability planning.

Laurie Arp's professional interests focus on the intersection of collections, technology, and people. As the Director of DuraSpace Community Supported Software, Laurie directs community supported open-source programs housed at LYRASIS including ArchivesSpace, CollectionSpace, DSpace, Fedora, and VIVO.

**Wednesday, 11 am ET**

# 2022 OHDSI APAC Symposium



**Day 1 (Nov. 12) — Tutorial Workshop**

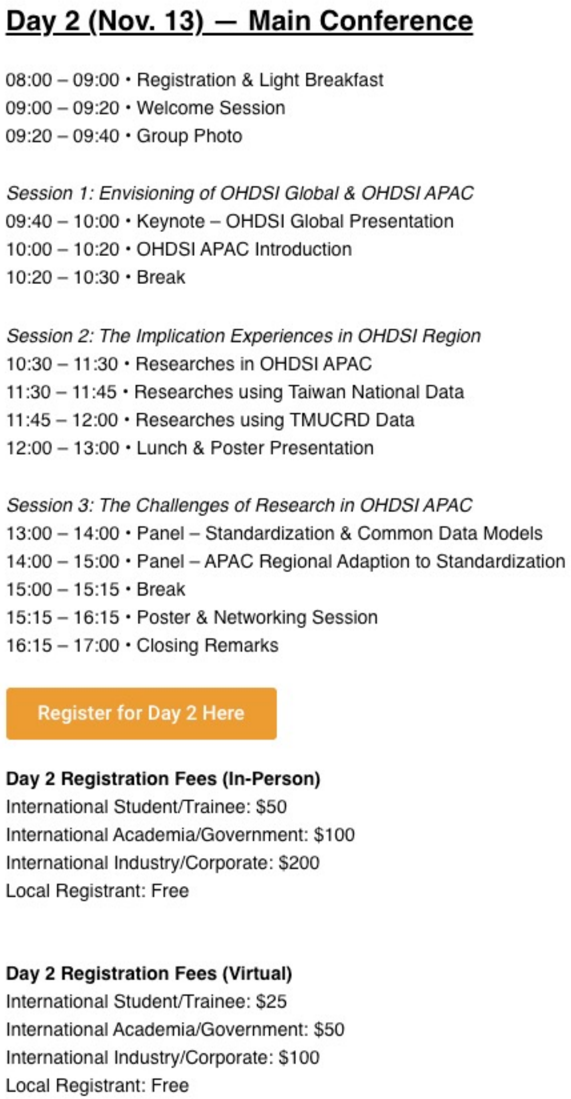
8:30 – 9:00 • Registration
9:00 – 9:30 • Overview of the OHDSI Journey: where are we going
9:30 – 10:20 • OMOP Common Data Model and vocabulary
10:20 – 10:30 • Break
10:30 – 11:20 • ETL a source database into OMOP CDM
11:20 – 11:30 • Break
11:30 – 12:20 • Creating cohort definitions
12:20 – 13:30 • Lunch
13:30 – 14:20 • Phenotype evaluation
14:20 – 14:30 • Break
14:30 – 15:20 • Characterization
15:20 – 15:30 • Break
15:30 – 16:20 • Estimation
16:20 – 16:30 • Break
16:30 – 17:20 • Prediction
17:20 – 17:30 • Recap of the OHDSI Journey, where do we go from here

Register for Day 1 Here

**Day 1 Registration Fees (In-Person)**
International Student/Trainee: $30
International Academia/Government: $70
International Industry/Corporate: $170
Local Registrants: Free

2022 APAC OHDSI Symposium

Nov. 12 - 13 • Taipei Medical University



**Day 2 (Nov. 13) — Main Conference**

08:00 – 09:00 • Registration & Light Breakfast
09:00 – 09:20 • Welcome Session
09:20 – 09:40 • Group Photo

*Session 1: Envisioning of OHDSI Global & OHDSI APAC*
09:40 – 10:00 • Keynote – OHDSI Global Presentation
10:00 – 10:20 • OHDSI APAC Introduction
10:20 – 10:30 • Break

*Session 2: The Implication Experiences in OHDSI Region*
10:30 – 11:30 • Researches in OHDSI APAC
11:30 – 11:45 • Researches using Taiwan National Data
11:45 – 12:00 • Researches using TMUCRD Data
12:00 – 13:00 • Lunch & Poster Presentation

*Session 3: The Challenges of Research in OHDSI APAC*
13:00 – 14:00 • Panel – Standardization & Common Data Models
14:00 – 15:00 • Panel – APAC Regional Adaption to Standardization
15:00 – 15:15 • Break
15:15 – 16:15 • Poster & Networking Session
16:15 – 17:00 • Closing Remarks

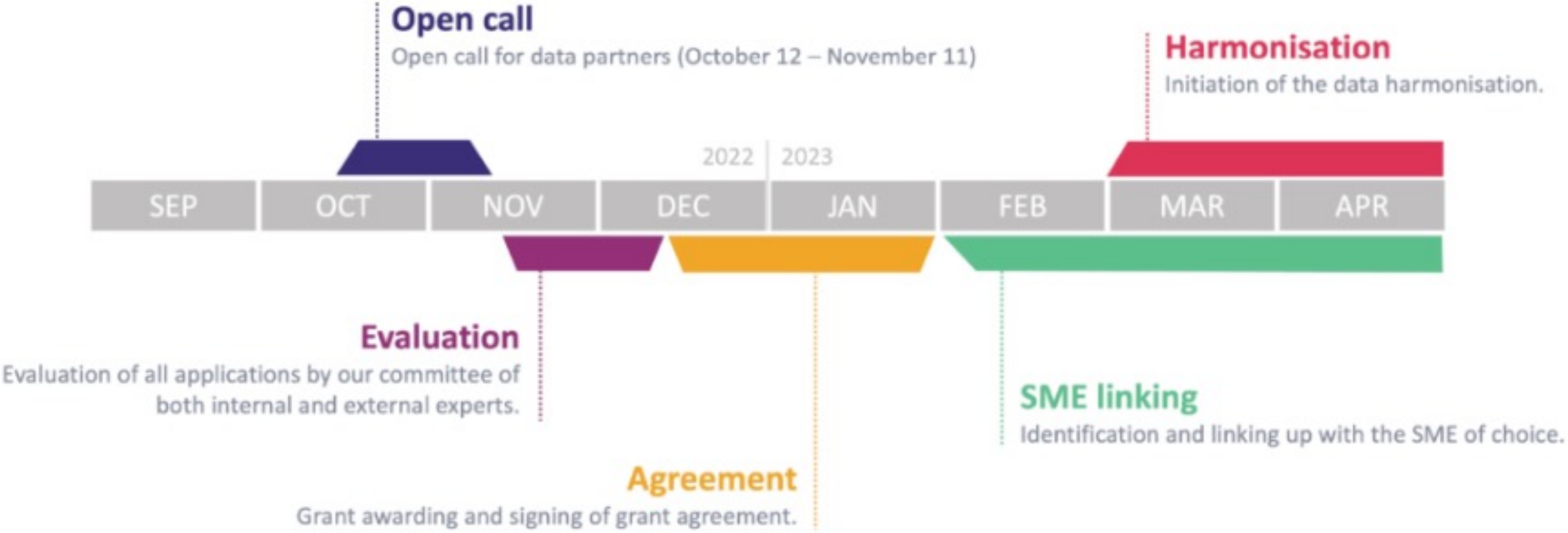
Register for Day 2 Here

**Day 2 Registration Fees (In-Person)**
International Student/Trainee: $50
International Academia/Government: $100
International Industry/Corporate: $200
Local Registrant: Free

**Day 2 Registration Fees (Virtual)**
International Student/Trainee: $25
International Academia/Government: $50
International Industry/Corporate: $100
Local Registrant: Free

ohdsi.org/2022apacsymposium

# 2022 OHDSI APAC Symposium



**Open call**

Open call for data partners (October 12 – November 11)

**Harmonisation**

Initiation of the data harmonisation.

2022 | 2023

SEP OCT NOV DEC JAN FEB MAR APR

**Evaluation**

Evaluation of all applications by our committee of both internal and external experts.

**Agreement**

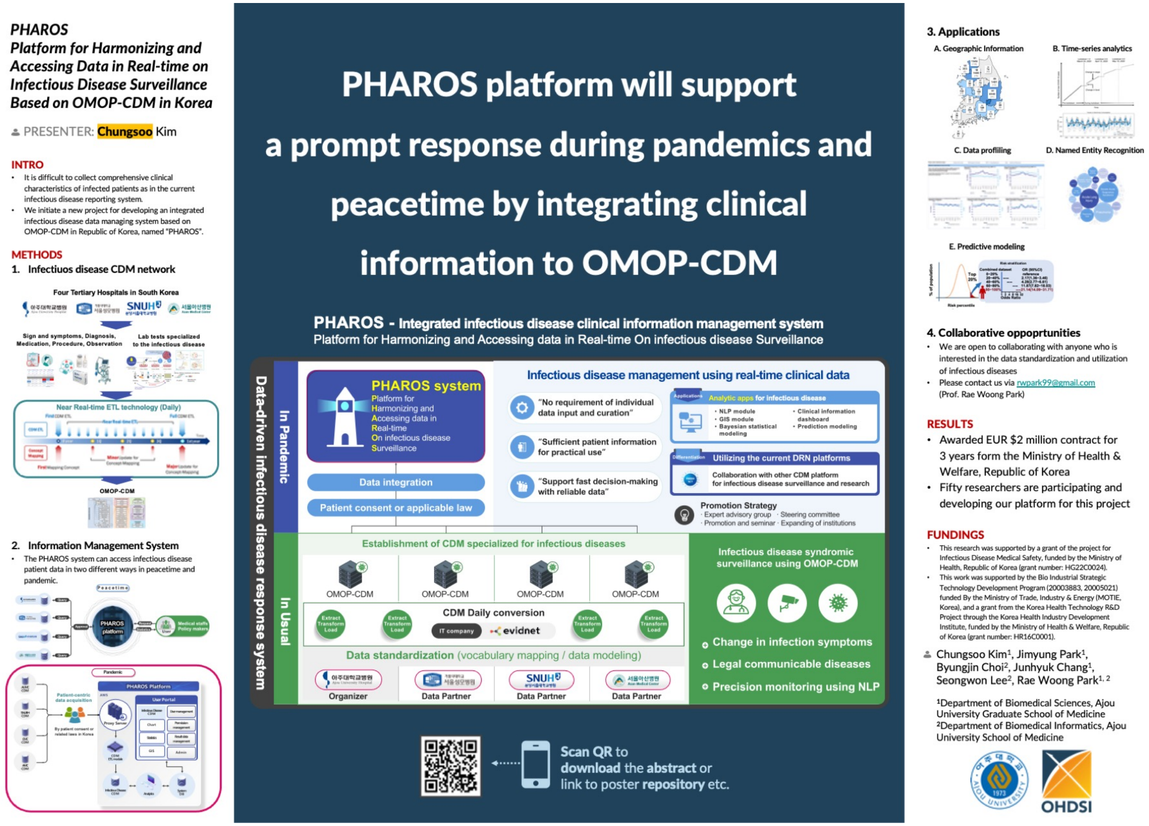
Grant awarding and signing of grant agreement.

**SME linking**

Identification and linking up with the SME of choice.

**Ehden.eu**

# #OHDSISocialShowcase This Week



*PHAROS*
***Platform for Harmonizing and Accessing Data in Real-time on Infectious Disease Surveillance Based on OMOP-CDM in Korea***

PRESENTER: Chungsoo Kim

**INTRO**

- It is difficult to collect comprehensive clinical characteristics of infected patients as in the current infectious disease reporting system.
- We initiate a new project for developing an integrated infectious disease data managing system based on OMOP-CDM in Republic of Korea, named "PHAROS".

**METHODS**

1. Infectiuos disease CDM network

Four Tertiary Hospitals in South Korea
Sign and symptoms, Diagnosis, Medication, Procedure, Observation
Lab tests specialized to the infectious disease
Near Real-time ETL technology (Daily)
OMOP-CDM

2. Information Management System

- The PHAROS system can access infectious disease patient data in two different ways in peacetime and pandemic.

## PHAROS platform will support a prompt response during pandemics and peacetime by integrating clinical information to OMOP-CDM

**PHAROS - Integrated infectious disease clinical information management system**
Platform for Harmonizing and Accessing data in Real-time On infectious disease Surveillance

Data-driven infectious disease response system
In Pandemic
In Usual
PHAROS system
Platform for Harmonizing and Accessing data in Real-time On infectious disease Surveillance
Data integration
Patient consent or applicable law
Infectious disease management using real-time clinical data
"No requirement of individual data input and curation"
"Sufficient patient information for practical use"
"Support fast decision-making with reliable data"
Applications
Analytic apps for infectious disease
· NLP module
· GIS module
· Bayesian statistical modeling
· Clinical information dashboard
· Prediction modeling
Differentiation
Utilizing the current DRN platforms
Collaboration with other CDM platform for infectious disease surveillance and research
Promotion Strategy
· Expert advisory group · Steering committee
· Promotion and seminar · Expanding of institutions
Establishment of CDM specialized for infectious diseases
OMOP-CDM
OMOP-CDM
OMOP-CDM
OMOP-CDM
Extract Transform Load
CDM Daily conversion
IT company
evidnet
Data standardization (vocabulary mapping / data modeling)
Organizer
Data Partner
Data Partner
Data Partner
Infectious disease syndromic surveillance using OMOP-CDM
Change in infection symptoms
Legal communicable diseases
Precision monitoring using NLP

Scan QR to **download** the **abstract** or link to poster **repository** etc.

**3. Applications**

**4. Collaborative oppoprtunities**

- We are open to collaborating with anyone who is interested in the data standardization and utilization of infectious diseases
- Please contact us via rwpark99@gmail.com (Prof. Rae Woong Park)

**RESULTS**

- Awarded EUR $2 million contract for 3 years form the Ministry of Health & Welfare, Republic of Korea
- Fifty researchers are participating and developing our platform for this project

**FUNDINGS**

- This research was supported by a grant of the project for Infectious Disease Medical Safety, funded by the Ministry of Health, Republic of Korea (grant number: HG22C0024).
- This work was supported by the Bio Industrial Strategic Technology Development Program (20003883, 20005021) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea), and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR16C0001).

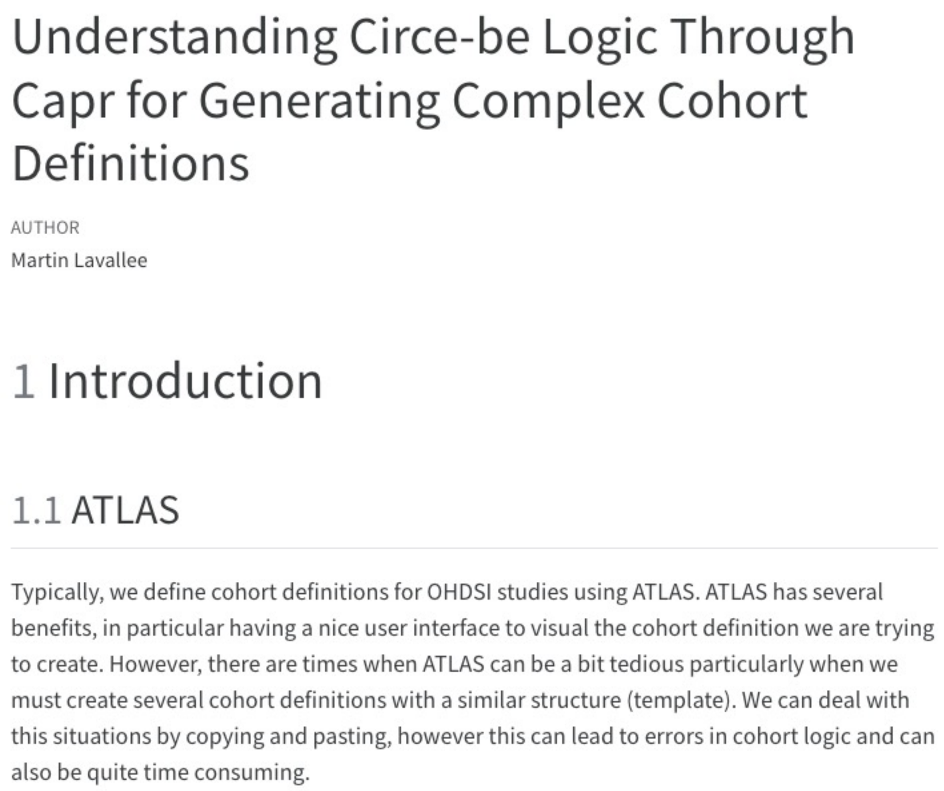
Chungsoo Kim[1], Jimyung Park[1], Byungjin Choi[2], Junhyuk Chang[1], Seongwon Lee[2], Rae Woong Park[1, 2]

[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine
[2]Department of Biomedical Informatics, Ajou University School of Medicine

**MONDAY**

**PHAROS, Platform for Harmonizing and Accessing Data in Real-time on Infectious Disease Surveillance Based on OMOP-CDM in Korea** (Chungsoo Kim, Jimyung Park, Byungjin Choi, Seongwon Lee, Rae Woong Park)

# #OHDSISocialShowcase This Week



## Understanding Circe-be Logic Through Capr for Generating Complex Cohort Definitions

AUTHOR
Martin Lavallee

## 1 Introduction

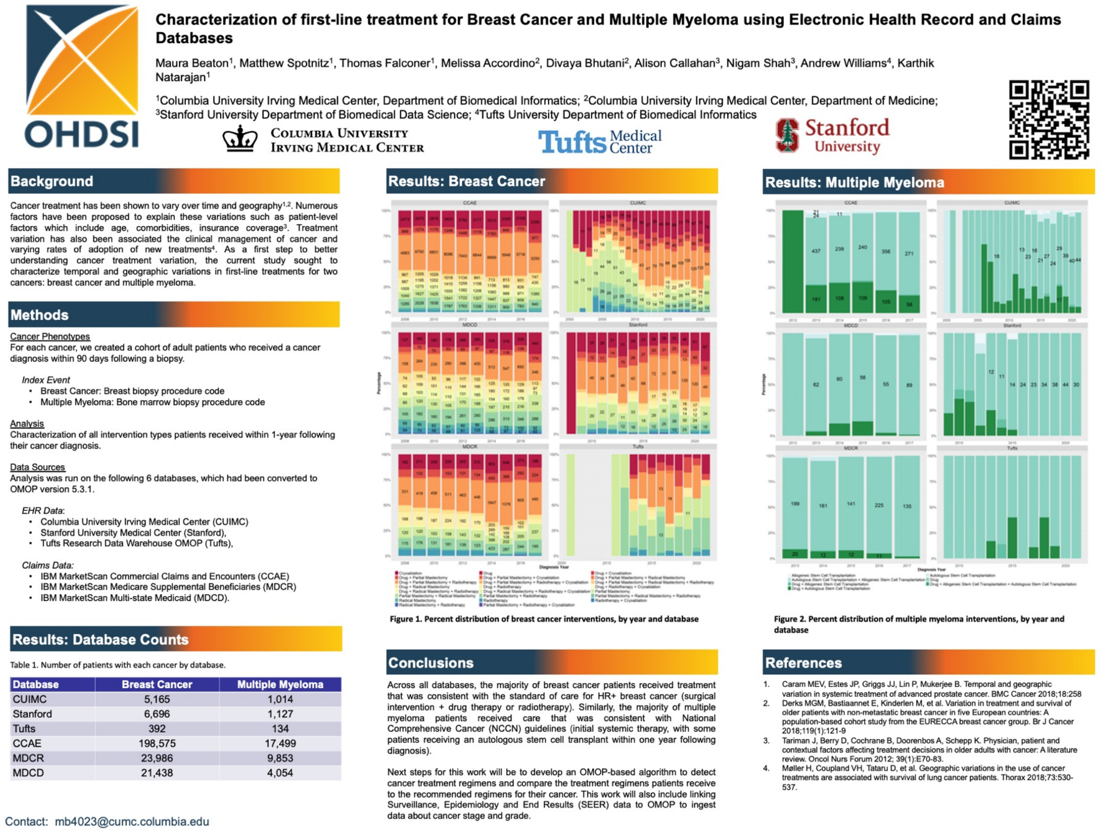
### 1.1 ATLAS

Typically, we define cohort definitions for OHDSI studies using ATLAS. ATLAS has several benefits, in particular having a nice user interface to visual the cohort definition we are trying to create. However, there are times when ATLAS can be a bit tedious particularly when we must create several cohort definitions with a similar structure (template). We can deal with this situations by copying and pasting, however this can lead to errors in cohort logic and can also be quite time consuming.

**TUESDAY**

**Understanding circe-be logic through Capr for generating complex cohort definitions (Martin Lavallee, Adam Black, Asieh Golozar)**

# #OHDSISocialShowcase This Week



## Characterization of first-line treatment for Breast Cancer and Multiple Myeloma using Electronic Health Record and Claims Databases

Maura Beaton[1], Matthew Spotnitz[1], Thomas Falconer[1], Melissa Accordino[2], Divaya Bhutani[2], Alison Callahan[3], Nigam Shah[3], Andrew Williams[4], Karthik Natarajan[1]

[1]Columbia University Irving Medical Center, Department of Biomedical Informatics; [2]Columbia University Irving Medical Center, Department of Medicine; [3]Stanford University Department of Biomedical Data Science; [4]Tufts University Department of Biomedical Informatics

### Background

Cancer treatment has been shown to vary over time and geography[1,2]. Numerous factors have been proposed to explain these variations such as patient-level factors which include age, comorbidities, insurance coverage[3]. Treatment variation has also been associated the clinical management of cancer and varying rates of adoption of new treatments[4]. As a first step to better understanding cancer treatment variation, the current study sought to characterize temporal and geographic variations in first-line treatments for two cancers: breast cancer and multiple myeloma.

### Methods

Cancer Phenotypes
For each cancer, we created a cohort of adult patients who received a cancer diagnosis within 90 days following a biopsy.

*Index Event*
- Breast Cancer: Breast biopsy procedure code
- Multiple Myeloma: Bone marrow biopsy procedure code

Analysis
Characterization of all intervention types patients received within 1-year following their cancer diagnosis.

Data Sources
Analysis was run on the following 6 databases, which had been converted to OMOP version 5.3.1.

*EHR Data*:
- Columbia University Irving Medical Center (CUIMC)
- Stanford University Medical Center (Stanford),
- Tufts Research Data Warehouse OMOP (Tufts),

*Claims Data:*
- IBM MarketScan Commercial Claims and Encounters (CCAE)
- IBM MarketScan Medicare Supplemental Beneficiaries (MDCR)
- IBM MarketScan Multi-state Medicaid (MDCD).

### Results: Database Counts

Table 1. Number of patients with each cancer by database.

| Database | Breast Cancer | Multiple Myeloma |
|---|---|---|
| CUIMC | 5,165 | 1,014 |
| Stanford | 6,696 | 1,127 |
| Tufts | 392 | 134 |
| CCAE | 198,575 | 17,499 |
| MDCR | 23,986 | 9,853 |
| MDCD | 21,438 | 4,054 |

Contact: mb4023@cumc.columbia.edu

### Results: Breast Cancer

Figure 1. Percent distribution of breast cancer interventions, by year and database

### Results: Multiple Myeloma

Figure 2. Percent distribution of multiple myeloma interventions, by year and database

### Conclusions

Across all databases, the majority of breast cancer patients received treatment that was consistent with the standard of care for HR+ breast cancer (surgical intervention + drug therapy or radiotherapy). Similarly, the majority of multiple myeloma patients received care that was consistent with National Comprehensive Cancer (NCCN) guidelines (initial systemic therapy, with some patients receiving an autologous stem cell transplant within one year following diagnosis).

Next steps for this work will be to develop an OMOP-based algorithm to detect cancer treatment regimens and compare the treatment regimens patients receive to the recommended regimens for their cancer. This work will also include linking Surveillance, Epidemiology and End Results (SEER) data to OMOP to ingest data about cancer stage and grade.

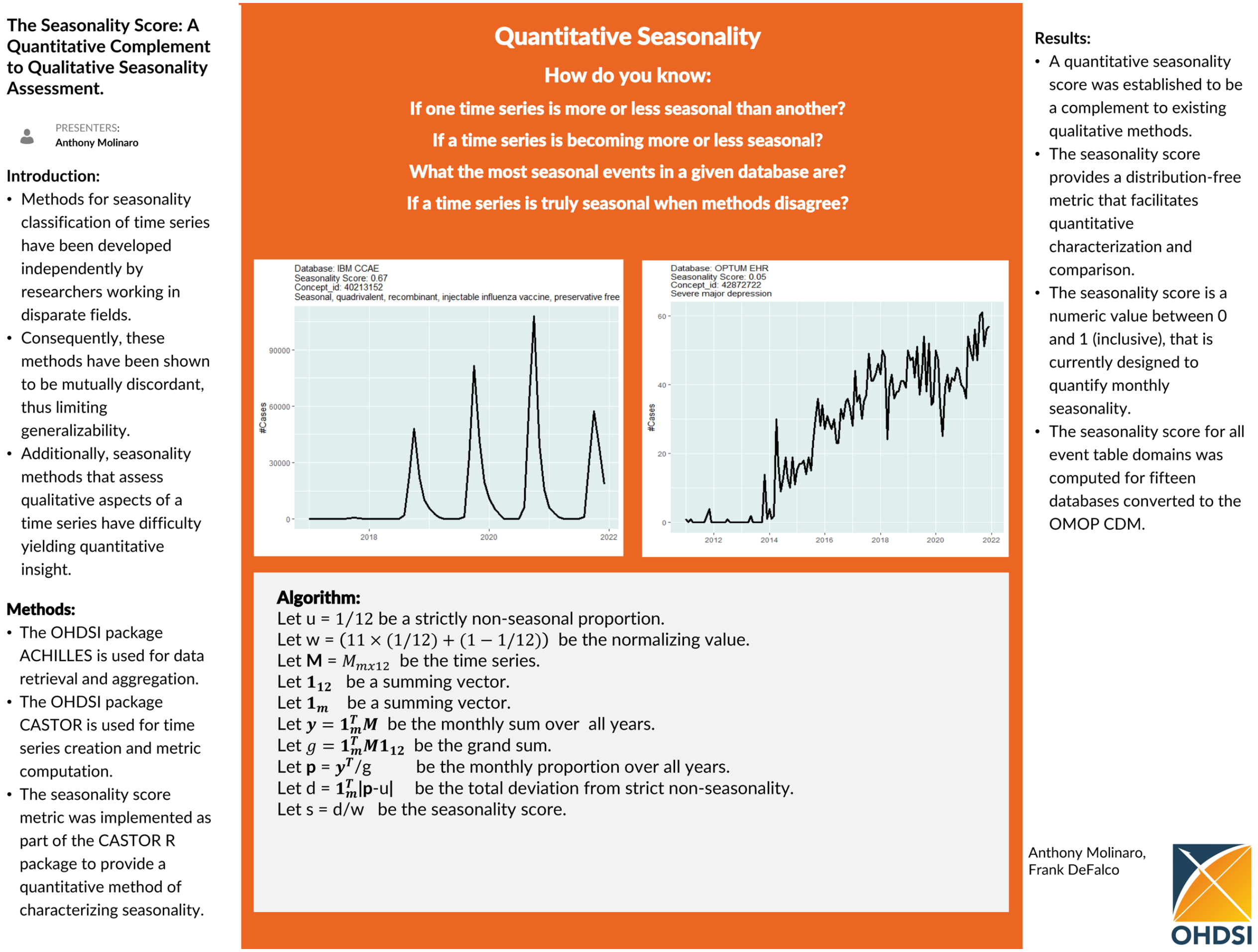
### References

1. Caram MEV, Estes JP, Griggs JJ, Lin P, Mukerjee B. Temporal and geographic variation in systemic treatment of advanced prostate cancer. BMC Cancer 2018;18:258
2. Derks MGM, Bastiaannet E, Kinderlen M, et al. Variation in treatment and survival of older patients with non-metastatic breast cancer in five European countries: A population-based cohort study from the EURECCA breast cancer group. Br J Cancer 2018;119(1):121-9
3. Tariman J, Berry D, Cochrane B, Doorenbos A, Schepp K. Physician, patient and contextual factors affecting treatment decisions in older adults with cancer: A literature review. Oncol Nurs Forum 2012; 39(1):E70-83.
4. Møller H, Coupland VH, Tataru D, et al. Geographic variations in the use of cancer treatments are associated with survival of lung cancer patients. Thorax 2018;73:530-537.

**WEDNESDAY**

**Characterization of first-line treatment for Breast Cancer and Multiple Myeloma using Electronic HealthRecord and Claims Databases (Maura Beaton, Matthew Spotnitz, Thomas Falconer, Melissa Accordino,DivayaBhutani, Alison Callahan, Nigam Shah, Jake Gillberg, Andrew Williams, Karthik Natarajan)**

# #OHDSISocialShowcase This Week



**The Seasonality Score: A Quantitative Complement to Qualitative Seasonality Assessment.**

PRESENTERS:
Anthony Molinaro

**Introduction:**

- Methods for seasonality classification of time series have been developed independently by researchers working in disparate fields.
- Consequently, these methods have been shown to be mutually discordant, thus limiting generalizability.
- Additionally, seasonality methods that assess qualitative aspects of a time series have difficulty yielding quantitative insight.

**Methods:**

- The OHDSI package ACHILLES is used for data retrieval and aggregation.
- The OHDSI package CASTOR is used for time series creation and metric computation.
- The seasonality score metric was implemented as part of the CASTOR R package to provide a quantitative method of characterizing seasonality.

## Quantitative Seasonality

**How do you know:**

**If one time series is more or less seasonal than another?**

**If a time series is becoming more or less seasonal?**

**What the most seasonal events in a given database are?**

**If a time series is truly seasonal when methods disagree?**

**Algorithm:**

Let u = 1/12 be a strictly non-seasonal proportion.

Let w = $(11 \times (1/12) + (1 - 1/12))$ be the normalizing value.

Let **M** = $M_{mx12}$ be the time series.

Let $\mathbf{1}_{12}$ be a summing vector.

Let $\mathbf{1}_{m}$ be a summing vector.

Let $\boldsymbol{y} = \mathbf{1}_m^T \boldsymbol{M}$ be the monthly sum over all years.

Let $g = \mathbf{1}_m^T \boldsymbol{M} \mathbf{1}_{12}$ be the grand sum.

Let **p** = $\boldsymbol{y}^T$/g be the monthly proportion over all years.

Let d = $\mathbf{1}_m^T$|**p**-u| be the total deviation from strict non-seasonality.

Let s = d/w be the seasonality score.

**Results:**

- A quantitative seasonality score was established to be a complement to existing qualitative methods.
- The seasonality score provides a distribution-free metric that facilitates quantitative characterization and comparison.
- The seasonality score is a numeric value between 0 and 1 (inclusive), that is currently designed to quantify monthly seasonality.
- The seasonality score for all event table domains was computed for fifteen databases converted to the OMOP CDM.

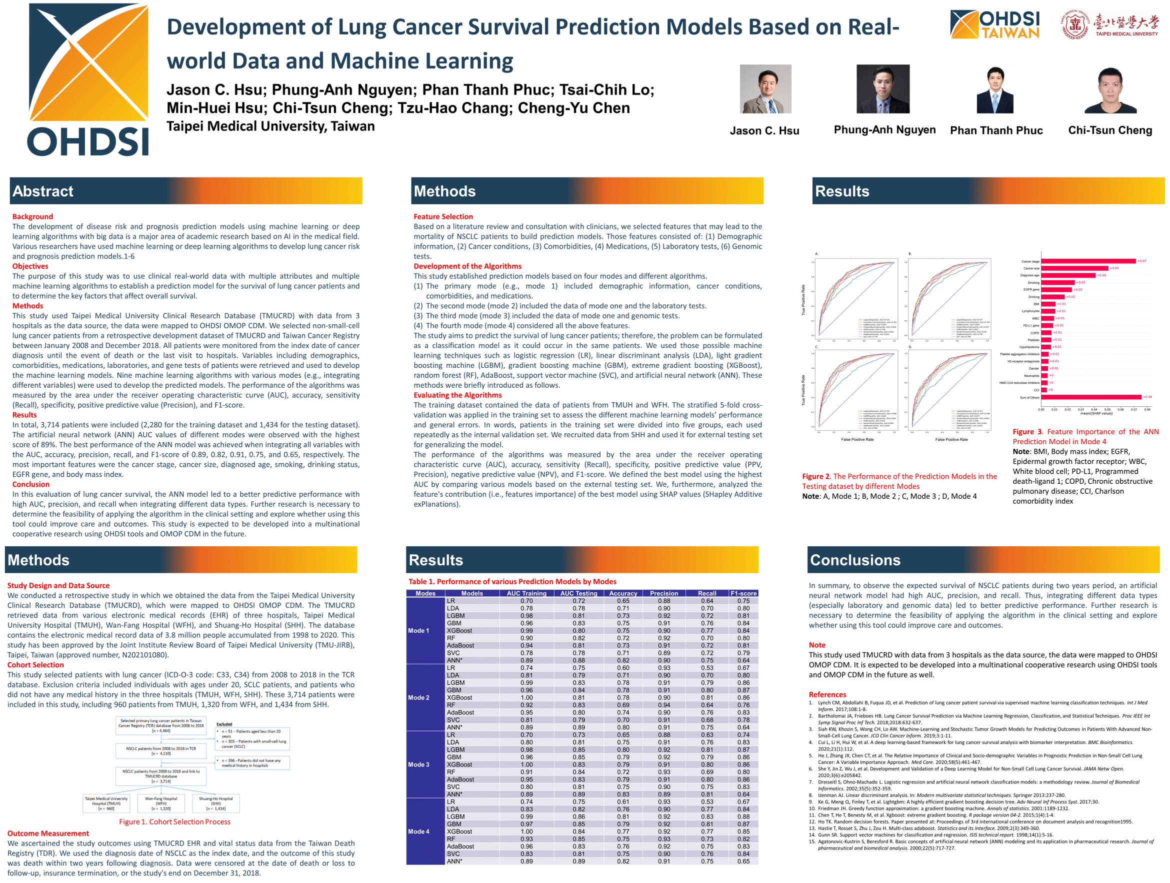
Anthony Molinaro,
Frank DeFalco

OHDSI

THURSDAY

## The Seasonality Score: A Quantitative Complement to Qualitative Seasonality Assessment (Anthony Molinaro, Frank DeFalco)

# #OHDSISocialShowcase This Week



## Development of Lung Cancer Survival Prediction Models Based on Real-world Data and Machine Learning

**Jason C. Hsu; Phung-Anh Nguyen; Phan Thanh Phuc; Tsai-Chih Lo; Min-Huei Hsu; Chi-Tsun Cheng; Tzu-Hao Chang; Cheng-Yu Chen**
**Taipei Medical University, Taiwan**

Jason C. Hsu
Phung-Anh Nguyen
Phan Thanh Phuc
Chi-Tsun Cheng

### Abstract

**Background**
The development of disease risk and prognosis prediction models using machine learning or deep learning algorithms with big data is a major area of academic research based on AI in the medical field. Various researchers have used machine learning or deep learning algorithms to develop lung cancer risk and prognosis prediction models.1-6

**Objectives**
The purpose of this study was to use clinical real-world data with multiple attributes and multiple machine learning algorithms to establish a prediction model for the survival of lung cancer patients and to determine the key factors that affect overall survival.

**Methods**
This study used Taipei Medical University Clinical Research Database (TMUCRD) with data from 3 hospitals as the data source, the data were mapped to OHDSI OMOP CDM. We selected non-small-cell lung cancer patients from a retrospective development dataset of TMUCRD and Taiwan Cancer Registry between January 2008 and December 2018. All patients were monitored from the index date of cancer diagnosis until the event of death or the last visit to hospitals. Variables including demographics, comorbidities, medications, laboratories, and gene tests of patients were retrieved and used to develop the machine learning models. Nine machine learning algorithms with various modes (e.g., integrating different variables) were used to develop the predicted models. The performance of the algorithms was measured by the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity (Recall), specificity, positive predictive value (Precision), and F1-score.

**Results**
In total, 3,714 patients were included (2,280 for the training dataset and 1,434 for the testing dataset). The artificial neural network (ANN) AUC values of different modes were observed with the highest score of 89%. The best performance of the ANN model was achieved when integrating all variables with the AUC, accuracy, precision, recall, and F1-score of 0.89, 0.82, 0.91, 0.75, and 0.65, respectively. The most important features were the cancer stage, cancer size, diagnosed age, smoking, drinking status, EGFR gene, and body mass index.

**Conclusion**
In this evaluation of lung cancer survival, the ANN model led to a better predictive performance with high AUC, precision, and recall when integrating different data types. Further research is necessary to determine the feasibility of applying the algorithm in the clinical setting and explore whether using this tool could improve care and outcomes. This study is expected to be developed into a multinational cooperative research using OHDSI tools and OMOP CDM in the future.

### Methods

**Study Design and Data Source**
We conducted a retrospective study in which we obtained the data from the Taipei Medical University Clinical Research Database (TMUCRD), which were mapped to OHDSI OMOP CDM. The TMUCRD retrieved data from various electronic medical records (EHR) of three hospitals, Taipei Medical University Hospital (TMUH), Wan-Fang Hospital (WFH), and Shuang-Ho Hospital (SHH). The database contains the electronic medical record data of 3.8 million people accumulated from 1998 to 2020. This study has been approved by the Joint Institute Review Board of Taipei Medical University (TMU-JIRB), Taipei, Taiwan (approved number, N202101080).

**Cohort Selection**
This study selected patients with lung cancer (ICD-O-3 code: C33, C34) from 2008 to 2018 in the TCR database. Exclusion criteria included individuals with ages under 20, SCLC patients, and patients who did not have any medical history in the three hospitals (TMUH, WFH, SHH). These 3,714 patients were included in this study, including 960 patients from TMUH, 1,320 from WFH, and 1,434 from SHH.

Selected primary lung cancer patients in Taiwan Cancer Registry (TCR) database from 2008 to 2018 [n = 4,464]

Excluded
- n = 51 – Patients aged less than 20 years
- n = 303 – Patients with small-cell lung cancer (SCLC)

NSCLC patients from 2008 to 2018 in TCR [n = 4,110]

- n = 396 - Patients did not have any medical history in hospitals

NSCLC patients from 2008 to 2018 and link to TMUCRD database [n = 3,714]

Taipei Medical University Hospital (TMUH) [n = 960]
Wan-Fang Hospital (WFH) [n = 1,320]
Shuang-Ho Hospital (SHH) [n = 1,434]

Figure 1. Cohort Selection Process

**Outcome Measurement**
We ascertained the study outcomes using TMUCRD EHR and vital status data from the Taiwan Death Registry (TDR). We used the diagnosis date of NSCLC as the index date, and the outcome of this study was death within two years following diagnosis. Data were censored at the date of death or loss to follow-up, insurance termination, or the study's end on December 31, 2018.

### Methods

**Feature Selection**
Based on a literature review and consultation with clinicians, we selected features that may lead to the mortality of NSCLC patients to build prediction models. Those features consisted of: (1) Demographic information, (2) Cancer conditions, (3) Comorbidities, (4) Medications, (5) Laboratory tests, (6) Genomic tests.

**Development of the Algorithms**
This study established prediction models based on four modes and different algorithms.

(1) The primary mode (e.g., mode 1) included demographic information, cancer conditions, comorbidities, and medications.
(2) The second mode (mode 2) included the data of mode one and the laboratory tests.
(3) The third mode (mode 3) included the data of mode one and genomic tests.
(4) The fourth mode (mode 4) considered all the above features.

The study aims to predict the survival of lung cancer patients; therefore, the problem can be formulated as a classification model as it could occur in the same patients. We used those possible machine learning techniques such as logistic regression (LR), linear discriminant analysis (LDA), light gradient boosting machine (LGBM), gradient boosting machine (GBM), extreme gradient boosting (XGBoost), random forest (RF), AdaBoost, support vector machine (SVC), and artificial neural network (ANN). These methods were briefly introduced as follows.

**Evaluating the Algorithms**
The training dataset contained the data of patients from TMUH and WFH. The stratified 5-fold cross-validation was applied in the training set to assess the different machine learning models' performance and general errors. In words, patients in the training set were divided into five groups, each used repeatedly as the internal validation set. We recruited data from SHH and used it for external testing set for generalizing the model.

The performance of the algorithms was measured by the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity (Recall), specificity, positive predictive value (PPV, Precision), negative predictive value (NPV), and F1-score. We defined the best model using the highest AUC by comparing various models based on the external testing set. We, furthermore, analyzed the feature's contribution (i.e., features importance) of the best model using SHAP values (SHapley Additive exPlanations).

### Results

**Table 1. Performance of various Prediction Models by Modes**

| Modes | Models | AUC Training | AUC Testing | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|
| Mode 1 | LR | 0.70 | 0.72 | 0.65 | 0.88 | 0.64 | 0.75 |
| | LDA | 0.78 | 0.78 | 0.71 | 0.90 | 0.70 | 0.80 |
| | LGBM | 0.98 | 0.81 | 0.73 | 0.92 | 0.72 | 0.81 |
| | GBM | 0.96 | 0.83 | 0.75 | 0.91 | 0.76 | 0.84 |
| | XGBoost | 0.99 | 0.80 | 0.75 | 0.90 | 0.77 | 0.84 |
| | RF | 0.90 | 0.82 | 0.72 | 0.92 | 0.70 | 0.80 |
| | AdaBoost | 0.94 | 0.81 | 0.73 | 0.91 | 0.72 | 0.81 |
| | SVC | 0.78 | 0.78 | 0.71 | 0.89 | 0.72 | 0.79 |
| | ANN* | 0.89 | 0.88 | 0.82 | 0.90 | 0.75 | 0.64 |
| Mode 2 | LR | 0.74 | 0.75 | 0.60 | 0.93 | 0.53 | 0.67 |
| | LDA | 0.81 | 0.79 | 0.71 | 0.90 | 0.70 | 0.80 |
| | LGBM | 0.99 | 0.83 | 0.78 | 0.91 | 0.79 | 0.86 |
| | GBM | 0.96 | 0.84 | 0.78 | 0.91 | 0.80 | 0.87 |
| | XGBoost | 1.00 | 0.81 | 0.78 | 0.90 | 0.81 | 0.86 |
| | RF | 0.92 | 0.83 | 0.69 | 0.94 | 0.64 | 0.76 |
| | AdaBoost | 0.95 | 0.80 | 0.74 | 0.90 | 0.76 | 0.83 |
| | SVC | 0.81 | 0.79 | 0.70 | 0.91 | 0.68 | 0.78 |
| | ANN* | 0.89 | 0.89 | 0.80 | 0.91 | 0.75 | 0.64 |
| Mode 3 | LR | 0.70 | 0.73 | 0.65 | 0.88 | 0.63 | 0.74 |
| | LDA | 0.80 | 0.81 | 0.75 | 0.91 | 0.76 | 0.83 |
| | LGBM | 0.98 | 0.85 | 0.80 | 0.92 | 0.81 | 0.87 |
| | GBM | 0.96 | 0.85 | 0.79 | 0.92 | 0.79 | 0.86 |
| | XGBoost | 1.00 | 0.83 | 0.79 | 0.91 | 0.80 | 0.86 |
| | RF | 0.91 | 0.84 | 0.72 | 0.93 | 0.69 | 0.80 |
| | AdaBoost | 0.95 | 0.83 | 0.79 | 0.91 | 0.80 | 0.86 |
| | SVC | 0.80 | 0.81 | 0.75 | 0.90 | 0.75 | 0.83 |
| | ANN* | 0.89 | 0.89 | 0.83 | 0.89 | 0.81 | 0.64 |
| Mode 4 | LR | 0.74 | 0.75 | 0.61 | 0.93 | 0.53 | 0.67 |
| | LDA | 0.83 | 0.82 | 0.76 | 0.90 | 0.77 | 0.84 |
| | LGBM | 0.99 | 0.86 | 0.81 | 0.92 | 0.83 | 0.88 |
| | GBM | 0.97 | 0.85 | 0.79 | 0.92 | 0.81 | 0.87 |
| | XGBoost | 1.00 | 0.84 | 0.77 | 0.92 | 0.77 | 0.85 |
| | RF | 0.93 | 0.85 | 0.75 | 0.93 | 0.73 | 0.82 |
| | AdaBoost | 0.96 | 0.83 | 0.76 | 0.92 | 0.75 | 0.83 |
| | SVC | 0.83 | 0.81 | 0.75 | 0.90 | 0.76 | 0.84 |
| | ANN* | 0.89 | 0.89 | 0.82 | 0.91 | 0.75 | 0.65 |

### Results

**Figure 2.** The Performance of the Prediction Models in the Testing dataset by different Modes
**Note**: A, Mode 1; B, Mode 2 ; C, Mode 3 ; D, Mode 4

**Figure 3.** Feature Importance of the ANN Prediction Model in Mode 4
**Note**: BMI, Body mass index; EGFR, Epidermal growth factor receptor; WBC, White blood cell; PD-L1, Programmed death-ligand 1; COPD, Chronic obstructive pulmonary disease; CCI, Charlson comorbidity index

### Conclusions

In summary, to observe the expected survival of NSCLC patients during two years period, an artificial neural network model had high AUC, precision, and recall. Thus, integrating different data types (especially laboratory and genomic data) led to better predictive performance. Further research is necessary to determine the feasibility of applying the algorithm in the clinical setting and explore whether using this tool could improve care and outcomes.

**Note**
This study used TMUCRD with data from 3 hospitals as the data source, the data were mapped to OHDSI OMOP CDM. It is expected to be developed into a multinational cooperative research using OHDSI tools and OMOP CDM in the future as well.

**References**

1. Lynch CM, Abdollahi B, Fuqua JD, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int J Med Inform.* 2017;108:1-8.
2. Bartholomai JA, Frieboes HB. Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. *Proc IEEE Int Symp Signal Proc Inf Tech.* 2018;2018:632-637.
3. Siah KW, Khozin S, Wong CH, Lo AW. Machine-Learning and Stochastic Tumor Growth Models for Predicting Outcomes in Patients With Advanced Non-Small-Cell Lung Cancer. *JCO Clin Cancer Inform.* 2019;3:1-11.
4. Cui L, Li H, Hui W, et al. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. *BMC Bioinformatics.* 2020;21(1):112.
5. He J, Zhang JX, Chen CT, et al. The Relative Importance of Clinical and Socio-demographic Variables in Prognostic Prediction in Non-Small Cell Lung Cancer: A Variable Importance Approach. *Med Care.* 2020;58(5):461-467.
6. She Y, Jin Z, Wu J, et al. Development and Validation of a Deep Learning Model for Non-Small Cell Lung Cancer Survival. *JAMA Netw Open.* 2020;3(6):e205842.
7. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics.* 2002;35(5):352-359.
8. Izenman AJ. Linear discriminant analysis. In: *Modern multivariate statistical techniques.* Springer 2013:237-280.
9. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst.* 2017;30.
10. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of statistics.* 2001:1189-1232.
11. Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting. *R package version 04-2.* 2015;1(4):1-4.
12. Ho TK. Random decision forests. Paper presented at: Proceedings of 3rd international conference on document analysis and recognition1995.
13. Hastie T, Rosset S, Zhu J, Zou H. Multi-class adaboost. *Statistics and its Interface.* 2009;2(3):349-360.
14. Gunn SR. Support vector machines for classification and regression. *ISIS technical report.* 1998;14(1):5-16.
15. Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis.* 2000;22(5):717-727.

**FRIDAY**

**Development of Lung Cancer Survival Prediction Models Based on Real-world Data and Machine Learning** (Jason C. Hsu, Phung-Anh Nguyen, Phan Thanh Phuc, Tsai-Chih Lo, Min-Huei Hsu, Chi-Tsun Cheng, Tzu-Hao Chang, Cheng-Yu Chen)

# Openings

FDA/CDER’s Division of Hepatology and Nutrition is seeking a clinician with bioinformatics or biostatistics training to work with the Drug-Induced Liver Injury (DILI) Team to evaluate large datasets of liver-related data, collaborate on the Team’s review of drugs with hepatotoxicity signals, and help develop informatics-based processes in DILI evaluation across the Agency.

Contact **Judy Racoosin** at **judith.racoosin@fda.hhs.gov** for information about the application process (that will be through USAJOBS).

# Openings

**Andrew Williams** recently announced two exciting new openings at Tufts Medicine.

1) Senior Project Manager for a multisite multiyear grant standardizing critical care EHR and waveform data. (CHoRUS Bridge2AI)

2) Lead software developer and research data warehouse manager for Tufts Medicine's OMOP instance and related services.

Remote work is possible for both positions.

1. Link for Senior Project Manager position: https://smrtr.io/bBVzh
2. Link for Lead Software Developer and Research Data Warehouse Manager position: https://jobs.smartrecruiters.com/TuftsMedicalCenter1/743999857980631-software-development-lead-res-g-c-ctsi

Andrew's email: awilliams15@tuftsmedicalcenter.org

# Openings

## Research Associate (Data Scientist/Statistical Engineer), Johns Hopkins inHealth and Biostatistics Center

- Execute OHDSI studies (e.g. for cohort characterizations and comparative effectiveness) on Johns Hopkins's EHR data to support clinicians;
- Collaborate with statisticians and clinicians to continuously integrate state-of-the-art statistical tools to the inHealth/OHDSI tool stack for deployment;
- Mentor trainees on data science and software development skills;
- Co-teach courses on observational health data analytics and data science skills at School of Medicine and Public Health;
- Facilitate adoption of the inHealth tools among the broader OHDSI community by  contributing to OHDSI's Health Analytics Data-to-Evidence Suite.
- https://apply.interfolio.com/114436

# Where Are We Going?

Any other announcements of upcoming work, events, deadlines, etc?

# Three Stages of The Journey

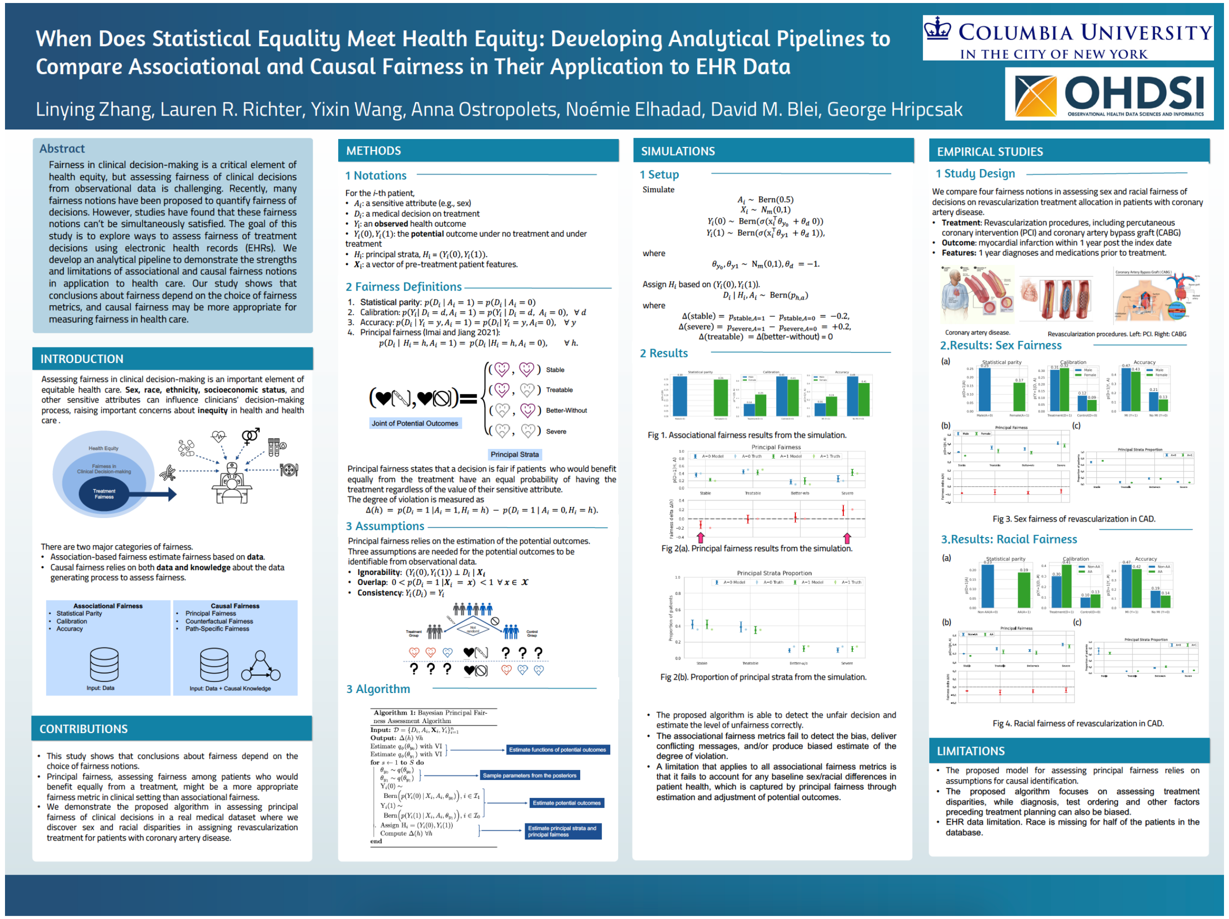
## Where Have We Been?

## Where Are We Now?

## Where Are We Going?

# Best Community Contribution Awards

Methods Research



## When Does Statistical Equality Meet Health Equity: Developing Analytical Pipelines to Compare Associational and Causal Fairness in Their Application to EHR Data

Linying Zhang, Lauren R. Richter, Yixin Wang, Anna Ostropolets, Noémie Elhadad, David M. Blei, George Hripcsak

Columbia University in the City of New York

OHDSI Observational Health Data Sciences and Informatics

### Abstract

Fairness in clinical decision-making is a critical element of health equity, but assessing fairness of clinical decisions from observational data is challenging. Recently, many fairness notions have been proposed to quantify fairness of decisions. However, studies have found that these fairness notions can't be simultaneously satisfied. The goal of this study is to explore ways to assess fairness of treatment decisions using electronic health records (EHRs). We develop an analytical pipeline to demonstrate the strengths and limitations of associational and causal fairness notions in application to health care. Our study shows that conclusions about fairness depend on the choice of fairness metrics, and causal fairness may be more appropriate for measuring fairness in health care.

### INTRODUCTION

Assessing fairness in clinical decision-making is an important element of equitable health care. **Sex, race, ethnicity, socioeconomic status**, and other sensitive attributes can influence clinicians' decision-making process, raising important concerns about **inequity** in health and health care .

There are two major categories of fairness.

- Association-based fairness estimate fairness based on **data**.
- Causal fairness relies on both **data and knowledge** about the data generating process to assess fairness.

**Associational Fairness**
- Statistical Parity
- Calibration
- Accuracy

Input: Data

**Causal Fairness**
- Principal Fairness
- Counterfactual Fairness
- Path-Specific Fairness

Input: Data + Causal Knowledge

### CONTRIBUTIONS

- This study shows that conclusions about fairness depend on the choice of fairness notions.
- Principal fairness, assessing fairness among patients who would benefit equally from a treatment, might be a more appropriate fairness metric in clinical setting than associational fairness.
- We demonstrate the proposed algorithm in assessing principal fairness of clinical decisions in a real medical dataset where we discover sex and racial disparities in assigning revascularization treatment for patients with coronary artery disease.

### METHODS

#### 1 Notations

For the *i*-th patient,

- $A_i$: a sensitive attribute (e.g., sex)
- $D_i$: a medical decision on treatment
- $Y_i$: an **observed** health outcome
- $Y_i(0), Y_i(1)$: the **potential** outcome under no treatment and under treatment
- $H_i$: principal strata, $H_i = (Y_i(0), Y_i(1))$.
- $\boldsymbol{X}_i$: a vector of pre-treatment patient features.

#### 2 Fairness Definitions

1. Statistical parity: $p(D_i \mid A_i = 1) = p(D_i \mid A_i = 0)$
2. Calibration: $p(Y_i \mid D_i = d, A_i = 1) = p(Y_i \mid D_i = d,\ A_i = 0),\ \ \forall d$
3. Accuracy: $p(D_i \mid Y_i = y, A_i = 1) = p(D_i \mid Y_i = y, A_i = 0),\ \ \forall y$
4. Principal fairness (Imai and Jiang 2021):
$$p(D_i \mid H_i = h, A_i = 1) = p(D_i \mid H_i = h, A_i = 0), \qquad \forall h.$$

Principal fairness states that a decision is fair if patients who would benefit equally from the treatment have an equal probability of having the treatment regardless of the value of their sensitive attribute.

The degree of violation is measured as
$$\Delta(h) = p(D_i = 1 \mid A_i = 1, H_i = h) - p(D_i = 1 \mid A_i = 0, H_i = h).$$

#### 3 Assumptions

Principal fairness relies on the estimation of the potential outcomes. Three assumptions are needed for the potential outcomes to be identifiable from observational data.

- **Ignorability**: $(Y_i(0), Y_i(1)) \perp D_i \mid \boldsymbol{X_i}$
- **Overlap**: $0 < p(D_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}) < 1\ \forall \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$
- **Consistency**: $Y_i(D_i) = Y_i$

#### 3 Algorithm

```
Algorithm 1: Bayesian Principal Fairness Assessment Algorithm
Input: $\mathcal{D} = \{D_i, A_i, \mathbf{X}_i, Y_i\}_{i=1}^n$
Output: $\Delta(h)\ \forall h$
Estimate $q_\phi(\theta_{y_0})$ with VI  }  → Estimate functions of potential outcomes
Estimate $q_\phi(\theta_{y_1})$ with VI  }
for $s \leftarrow 1$ to $S$ do
    $\theta_{y_0} \sim q(\theta_{y_0})$  }  → Sample parameters from the posteriors
    $\theta_{y_1} \sim q(\theta_{y_1})$  }
    $Y_i(0) \sim$
      $\mathrm{Bern}\big(p(Y_i(0) \mid X_i, A_i, \theta_{y_0})\big), i \in \mathcal{I}_1$  }  → Estimate potential outcomes
    $Y_i(1) \sim$
      $\mathrm{Bern}\big(p(Y_i(1) \mid X_i, A_i, \theta_{y_1})\big), i \in \mathcal{I}_0$  }
    Assign $H_i = (Y_i(0), Y_i(1))$  }  → Estimate principal strata and principal fairness
    Compute $\Delta(h)\ \forall h$
end
```

### SIMULATIONS

#### 1 Setup

Simulate
$$A_i \sim \mathrm{Bern}(0.5)$$
$$X_i \sim N_m(0,1)$$
$$Y_i(0) \sim \mathrm{Bern}(\sigma(x_i^T \theta_{y_0} + \theta_d\, 0))$$
$$Y_i(1) \sim \mathrm{Bern}(\sigma(x_i^T \theta_{y1} + \theta_d\, 1)),$$

where
$$\theta_{y_0}, \theta_{y1} \sim \mathrm{N}_m(0,1), \theta_d = -1.$$

Assign $H_i$ based on $(Y_i(0), Y_i(1))$.
$$D_i \mid H_i, A_i \sim \mathrm{Bern}(p_{h,a})$$
where
$$\Delta(\text{stable}) = p_{\text{stable},A=1} - p_{\text{stable},A=0} = -0.2,$$
$$\Delta(\text{severe}) = p_{\text{severe},A=1} - p_{\text{severe},A=0} = +0.2,$$
$$\Delta(\text{treatable}) = \Delta(\text{better-without}) = 0$$

#### 2 Results

Fig 1. Associational fairness results from the simulation.

Fig 2(a). Principal fairness results from the simulation.

Fig 2(b). Proportion of principal strata from the simulation.

- The proposed algorithm is able to detect the unfair decision and estimate the level of unfairness correctly.
- The associational fairness metrics fail to detect the bias, deliver conflicting messages, and/or produce biased estimate of the degree of violation.
- A limitation that applies to all associational fairness metrics is that it fails to account for any baseline sex/racial differences in patient health, which is captured by principal fairness through estimation and adjustment of potential outcomes.

### EMPIRICAL STUDIES

#### 1 Study Design

We compare four fairness notions in assessing sex and racial fairness of decisions on revascularization treatment allocation in patients with coronary artery disease.

- **Treatment**: Revascularization procedures, including percutaneous coronary intervention (PCI) and coronary artery bypass graft (CABG)
- **Outcome**: myocardial infarction within 1 year post the index date
- **Features:** 1 year diagnoses and medications prior to treatment.

Coronary artery disease.

Revascularization procedures. Left: PCI. Right: CABG

#### 2.Results: Sex Fairness

Fig 3. Sex fairness of revascularization in CAD.

#### 3.Results: Racial Fairness

Fig 4. Racial fairness of revascularization in CAD.

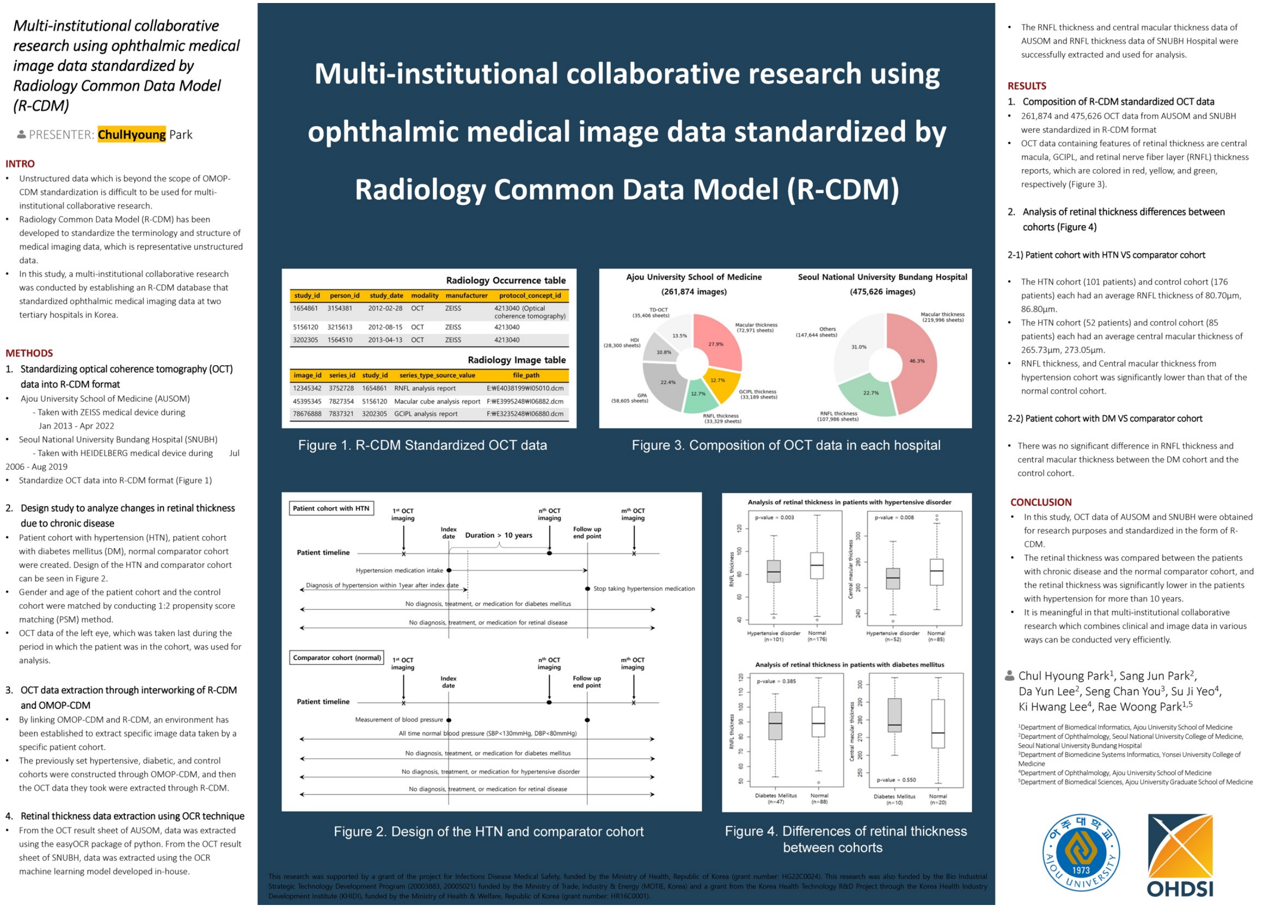
### LIMITATIONS

- The proposed model for assessing principal fairness relies on assumptions for causal identification.
- The proposed algorithm focuses on assessing treatment disparities, while diagnosis, test ordering and other factors preceding treatment planning can also be biased.
- EHR data limitation. Race is missing for half of the patients in the database.

Assessing Racial Fairness of Dialysis Allocation in End-Stage Renal Disease (**Linying Zhang**, Lauren R. Richter, David M. Blei, Yixin Wang, Anna Ostropolets, Noemie Elhadad, George Hripcsak)

# Best Community Contribution Awards

Data Standards



*Multi-institutional collaborative research using ophthalmic medical image data standardized by Radiology Common Data Model (R-CDM)*

PRESENTER: **ChulHyoung** Park

## INTRO

- Unstructured data which is beyond the scope of OMOP-CDM standardization is difficult to be used for multi-institutional collaborative research.
- Radiology Common Data Model (R-CDM) has been developed to standardize the terminology and structure of medical imaging data, which is representative unstructured data.
- In this study, a multi-institutional collaborative research was conducted by establishing an R-CDM database that standardized ophthalmic medical imaging data at two tertiary hospitals in Korea.

## METHODS

1. Standardizing optical coherence tomography (OCT) data into R-CDM format

- Ajou University School of Medicine (AUSOM)
  - Taken with ZEISS medical device during Jan 2013 - Apr 2022
- Seoul National University Bundang Hospital (SNUBH)
  - Taken with HEIDELBERG medical device during Jul 2006 - Aug 2019
- Standardize OCT data into R-CDM format (Figure 1)

2. Design study to analyze changes in retinal thickness due to chronic disease

- Patient cohort with hypertension (HTN), patient cohort with diabetes mellitus (DM), normal comparator cohort were created. Design of the HTN and comparator cohort can be seen in Figure 2.
- Gender and age of the patient cohort and the control cohort were matched by conducting 1:2 propensity score matching (PSM) method.
- OCT data of the left eye, which was taken last during the period in which the patient was in the cohort, was used for analysis.

3. OCT data extraction through interworking of R-CDM and OMOP-CDM

- By linking OMOP-CDM and R-CDM, an environment has been established to extract specific image data taken by a specific patient cohort.
- The previously set hypertensive, diabetic, and control cohorts were constructed through OMOP-CDM, and then the OCT data they took were extracted through R-CDM.

4. Retinal thickness data extraction using OCR technique

- From the OCT result sheet of AUSOM, data was extracted using the easyOCR package of python. From the OCT result sheet of SNUBH, data was extracted using the OCR machine learning model developed in-house.

## Multi-institutional collaborative research using ophthalmic medical image data standardized by Radiology Common Data Model (R-CDM)

Radiology Occurrence table

| study_id | person_id | study_date | modality | manufacturer | protocol_concept_id |
|---|---|---|---|---|---|
| 1654861 | 3154381 | 2012-02-28 | OCT | ZEISS | 4213040 (Optical coherence tomography) |
| 5156120 | 3215613 | 2012-08-15 | OCT | ZEISS | 4213040 |
| 3202305 | 1564510 | 2013-04-13 | OCT | ZEISS | 4213040 |

Radiology Image table

| image_id | series_id | study_id | series_type_source_value | file_path |
|---|---|---|---|---|
| 12345342 | 3752728 | 1654861 | RNFL analysis report | E:₩E4038199₩I05010.dcm |
| 45395345 | 7827354 | 5156120 | Macular cube analysis report | F:₩E3995248₩I06882.dcm |
| 78676888 | 7837321 | 3202305 | GCIPL analysis report | F:₩E3235248₩I06880.dcm |

Figure 1. R-CDM Standardized OCT data

Ajou University School of Medicine
(261,874 images)
TD-OCT (35,406 sheets)
13.5%
Macular thickness (72,971 sheets)
27.9%
HDI (28,300 sheets)
10.8%
GPA (58,605 sheets)
22.4%
RNFL thickness (33,329 sheets)
12.7%
GCIPL thickness (33,189 sheets)
12.7%
Seoul National University Bundang Hospital
(475,626 images)
Others (147,644 sheets)
31.0%
Macular thickness (219,996 sheets)
46.3%
RNFL thickness (107,986 sheets)
22.7%

Figure 3. Composition of OCT data in each hospital

Patient cohort with HTN
1st OCT imaging
Index date
Duration > 10 years
nth OCT imaging
Follow up end point
mth OCT imaging
Patient timeline
Hypertension medication intake
Diagnosis of hypertension within 1year after index date
Stop taking hypertension medication
No diagnosis, treatment, or medication for diabetes mellitus
No diagnosis, treatment, or medication for retinal disease
Comparator cohort (normal)
1st OCT imaging
Index date
nth OCT imaging
Follow up end point
mth OCT imaging
Patient timeline
Measurement of blood pressure
All time normal blood pressure (SBP<130mmHg, DBP<80mmHg)
No diagnosis, treatment, or medication for diabetes mellitus
No diagnosis, treatment, or medication for hypertensive disorder
No diagnosis, treatment, or medication for retinal disease

Figure 2. Design of the HTN and comparator cohort

Figure 4. Differences of retinal thickness between cohorts

This research was supported by a grant of the project for Infections Disease Medical Safety, funded by the Ministry of Health, Republic of Korea (grant number: HG22C0024). This research was also funded by the Bio Industrial Strategic Technology Development Program (20003883, 20005021) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR16C0001).

- The RNFL thickness and central macular thickness data of AUSOM and RNFL thickness data of SNUBH Hospital were successfully extracted and used for analysis.

## RESULTS

1. Composition of R-CDM standardized OCT data

- 261,874 and 475,626 OCT data from AUSOM and SNUBH were standardized in R-CDM format
- OCT data containing features of retinal thickness are central macula, GCIPL, and retinal nerve fiber layer (RNFL) thickness reports, which are colored in red, yellow, and green, respectively (Figure 3).

2. Analysis of retinal thickness differences between cohorts (Figure 4)

2-1) Patient cohort with HTN VS comparator cohort

- The HTN cohort (101 patients) and control cohort (176 patients) each had an average RNFL thickness of 80.70μm, 86.80μm.
- The HTN cohort (52 patients) and control cohort (85 patients) each had an average central macular thickness of 265.73μm, 273.05μm.
- RNFL thickness, and Central macular thickness from hypertension cohort was significantly lower than that of the normal control cohort.

2-2) Patient cohort with DM VS comparator cohort

- There was no significant difference in RNFL thickness and central macular thickness between the DM cohort and the control cohort.

## CONCLUSION

- In this study, OCT data of AUSOM and SNUBH were obtained for research purposes and standardized in the form of R-CDM.
- The retinal thickness was compared between the patients with chronic disease and the normal comparator cohort, and the retinal thickness was significantly lower in the patients with hypertension for more than 10 years.
- It is meaningful in that multi-institutional collaborative research which combines clinical and image data in various ways can be conducted very efficiently.

Chul Hyoung Park[1], Sang Jun Park[2], Da Yun Lee[2], Seng Chan You[3], Su Ji Yeo[4], Ki Hwang Lee[4], Rae Woong Park[1,5]

[1]Department of Biomedical Informatics, Ajou University School of Medicine
[2]Department of Ophthalmology, Seoul National University College of Medicine, Seoul National University Bundang Hospital
[3]Department of Biomedicine Systems Informatics, Yonsei University College of Medicine
[4]Department of Ophthalmology, Ajou University School of Medicine
[5]Department of Biomedical Sciences, Ajou University Graduate School of Medicine

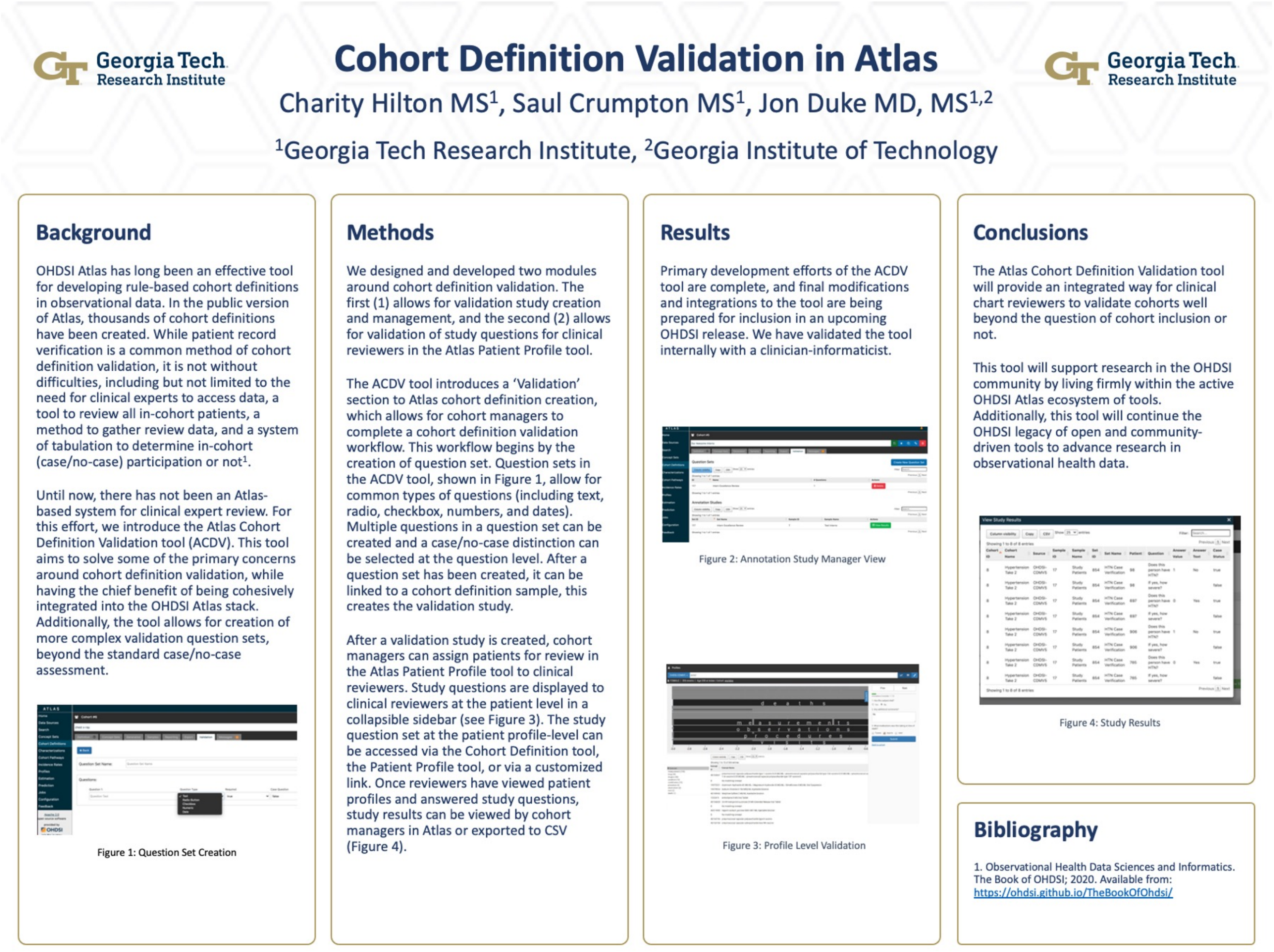
아주대학교 AJOU UNIVERSITY 1973

OHDSI

**Analyzing the Effect of Hypertension on Retinal Thickness Using Radiology Common Data Model (R-CDM) (Chul Hyoung Park, Rae Woong Park, Sang Jun Park, Da Yun Lee, Seng Chan You, Ki Hwang Lee)**

# Best Community Contribution Awards

Open-Source Analytics



Georgia Tech Research Institute

## Cohort Definition Validation in Atlas

Charity Hilton MS[1], Saul Crumpton MS[1], Jon Duke MD, MS[1,2]

[1]Georgia Tech Research Institute, [2]Georgia Institute of Technology

Georgia Tech Research Institute

### Background

OHDSI Atlas has long been an effective tool for developing rule-based cohort definitions in observational data. In the public version of Atlas, thousands of cohort definitions have been created. While patient record verification is a common method of cohort definition validation, it is not without difficulties, including but not limited to the need for clinical experts to access data, a tool to review all in-cohort patients, a method to gather review data, and a system of tabulation to determine in-cohort (case/no-case) participation or not[1].

Until now, there has not been an Atlas-based system for clinical expert review. For this effort, we introduce the Atlas Cohort Definition Validation tool (ACDV). This tool aims to solve some of the primary concerns around cohort definition validation, while having the chief benefit of being cohesively integrated into the OHDSI Atlas stack. Additionally, the tool allows for creation of more complex validation question sets, beyond the standard case/no-case assessment.

Figure 1: Question Set Creation

### Methods

We designed and developed two modules around cohort definition validation. The first (1) allows for validation study creation and management, and the second (2) allows for validation of study questions for clinical reviewers in the Atlas Patient Profile tool.

The ACDV tool introduces a 'Validation' section to Atlas cohort definition creation, which allows for cohort managers to complete a cohort definition validation workflow. This workflow begins by the creation of question set. Question sets in the ACDV tool, shown in Figure 1, allow for common types of questions (including text, radio, checkbox, numbers, and dates). Multiple questions in a question set can be created and a case/no-case distinction can be selected at the question level. After a question set has been created, it can be linked to a cohort definition sample, this creates the validation study.

After a validation study is created, cohort managers can assign patients for review in the Atlas Patient Profile tool to clinical reviewers. Study questions are displayed to clinical reviewers at the patient level in a collapsible sidebar (see Figure 3). The study question set at the patient profile-level can be accessed via the Cohort Definition tool, the Patient Profile tool, or via a customized link. Once reviewers have viewed patient profiles and answered study questions, study results can be viewed by cohort managers in Atlas or exported to CSV (Figure 4).

### Results

Primary development efforts of the ACDV tool are complete, and final modifications and integrations to the tool are being prepared for inclusion in an upcoming OHDSI release. We have validated the tool internally with a clinician-informaticist.

Figure 2: Annotation Study Manager View

Figure 3: Profile Level Validation

### Conclusions

The Atlas Cohort Definition Validation tool will provide an integrated way for clinical chart reviewers to validate cohorts well beyond the question of cohort inclusion or not.

This tool will support research in the OHDSI community by living firmly within the active OHDSI Atlas ecosystem of tools. Additionally, this tool will continue the OHDSI legacy of open and community-driven tools to advance research in observational health data.

Figure 4: Study Results

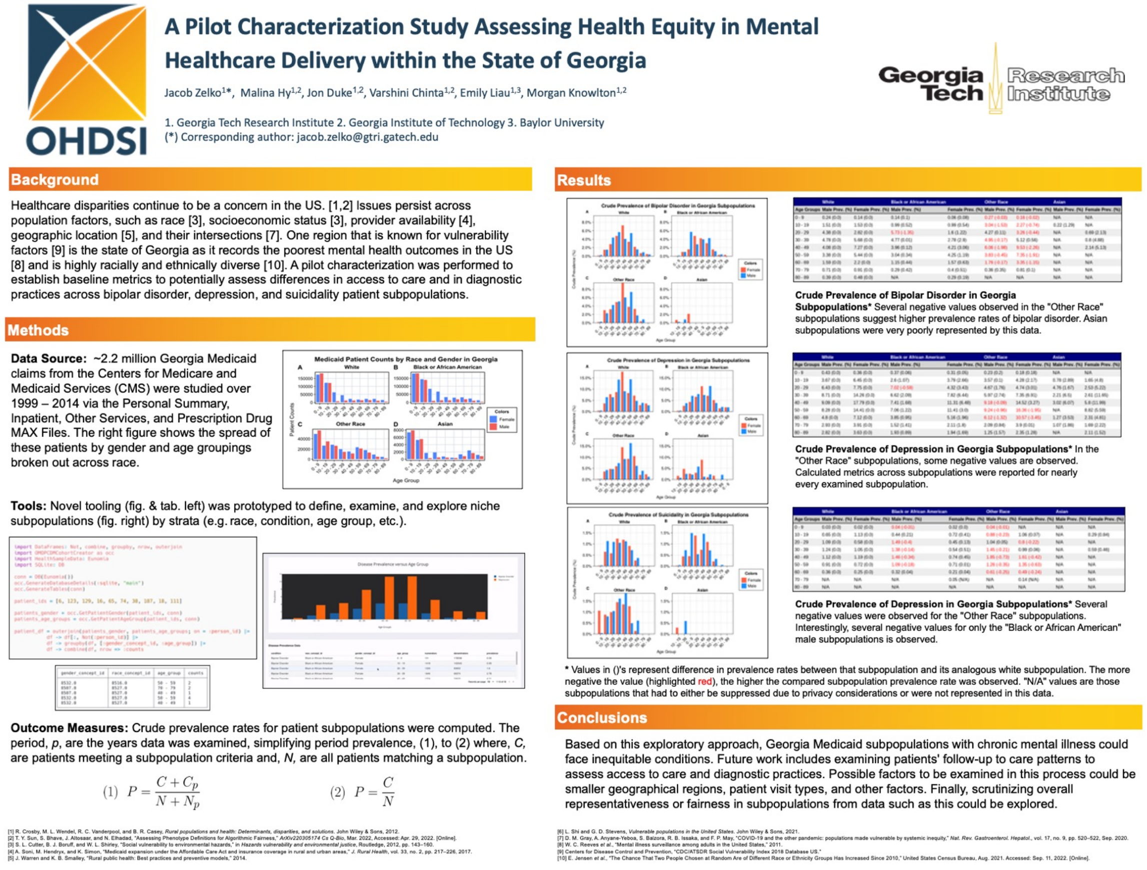
### Bibliography

1. Observational Health Data Sciences and Informatics. The Book of OHDSI; 2020. Available from: https://ohdsi.github.io/TheBookOfOhdsi/

Cohort Definition Validation in Atlas (Charity Hilton, Saul Crumpton, Jon Duke)

# Best Community Contribution Awards

Clinical Applications



## A Pilot Characterization Study Assessing Health Equity in Mental Healthcare Delivery within the State of Georgia

Jacob Zelko[1*], Malina Hy[1,2], Jon Duke[1,2], Varshini Chinta[1,2], Emily Liau[1,3], Morgan Knowlton[1,2]

1. Georgia Tech Research Institute 2. Georgia Institute of Technology 3. Baylor University
(*) Corresponding author: jacob.zelko@gtri.gatech.edu

Georgia Tech Research Institute

### Background

Healthcare disparities continue to be a concern in the US. [1,2] Issues persist across population factors, such as race [3], socioeconomic status [3], provider availability [4], geographic location [5], and their intersections [7]. One region that is known for vulnerability factors [9] is the state of Georgia as it records the poorest mental health outcomes in the US [8] and is highly racially and ethnically diverse [10]. A pilot characterization was performed to establish baseline metrics to potentially assess differences in access to care and in diagnostic practices across bipolar disorder, depression, and suicidality patient subpopulations.

### Methods

**Data Source:** ~2.2 million Georgia Medicaid claims from the Centers for Medicare and Medicaid Services (CMS) were studied over 1999 – 2014 via the Personal Summary, Inpatient, Other Services, and Prescription Drug MAX Files. The right figure shows the spread of these patients by gender and age groupings broken out across race.

Medicaid Patient Counts by Race and Gender in Georgia
A White
B Black or African American
C Other Race
D Asian
Patient Counts
Age Group
Colors
Female
Male

**Tools:** Novel tooling (fig. & tab. left) was prototyped to define, examine, and explore niche subpopulations (fig. right) by strata (e.g. race, condition, age group, etc.).

```
import DataFrames: Not, combine, groupby, nrow, outerjoin
import OMOPCDMCohortCreator as occ
import HealthSampleData: Eunomia
import SQLite: DB

conn = DB(Eunomia())
occ.GenerateDatabaseDetails(:sqlite, "main")
occ.GenerateTables(conn)

patient_ids = [6, 123, 129, 16, 65, 74, 38, 107, 18, 111]

patients_gender = occ.GetPatientGender(patient_ids, conn)
patients_age_groups = occ.GetPatientAgeGroup(patient_ids, conn)

patient_df = outerjoin(patients_gender, patients_age_groups; on = :person_id) |>
             df -> df[:, Not(:person_id)] |>
             df -> groupby(df, [:gender_concept_id, :age_group]) |>
             df -> combine(df, nrow => :counts
```

| gender_concept_id | race_concept_id | age_group | counts |
|---|---|---|---|
| 8532.0 | 8516.0 | 50 - 59 | 2 |
| 8507.0 | 8527.0 | 70 - 79 | 2 |
| 8507.0 | 8527.0 | 40 - 49 | 1 |
| 8532.0 | 8527.0 | 50 - 59 | 4 |
| 8532.0 | 8527.0 | 40 - 49 | 1 |

**Outcome Measures:** Crude prevalence rates for patient subpopulations were computed. The period, *p*, are the years data was examined, simplifying period prevalence, (1), to (2) where, *C*, are patients meeting a subpopulation criteria and, *N*, are all patients matching a subpopulation.

$$(1)\quad P = \frac{C + C_p}{N + N_p} \qquad\qquad (2)\quad P = \frac{C}{N}$$

### Results

Crude Prevalence of Bipolar Disorder in Georgia Subpopulations
A White
B Black or African American
C Other Race
D Asian
Crude Prevalence (%)
Age Group
Colors
Female
Male

| | White | | Black or African American | | Other Race | | Asian | |
|---|---|---|---|---|---|---|---|---|
| Age Groups | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) |
| 0 - 9 | 0.24 (0.0) | 0.14 (0.0) | 0.14 (0.1) | 0.06 (0.08) | 0.27 (-0.03) | 0.16 (-0.02) | N/A | N/A |
| 10 - 19 | 1.51 (0.0) | 1.53 (0.0) | 0.99 (0.52) | 0.99 (0.54) | 3.04 (-1.53) | 2.27 (-0.74) | 0.22 (1.29) | N/A |
| 20 - 29 | 4.38 (0.0) | 2.82 (0.0) | 5.73 (-1.35) | 1.6 (1.22) | 4.27 (0.11) | 3.26 (-0.44) | N/A | 0.69 (2.13) |
| 30 - 39 | 4.78 (0.0) | 5.68 (0.0) | 4.77 (0.01) | 2.78 (2.9) | 4.95 (-0.17) | 5.12 (0.56) | N/A | 0.8 (4.88) |
| 40 - 49 | 4.08 (0.0) | 7.27 (0.0) | 3.96 (0.12) | 4.21 (3.06) | 6.06 (-1.98) | 9.53 (-2.26) | N/A | 2.14 (5.13) |
| 50 - 59 | 3.38 (0.0) | 5.44 (0.0) | 3.04 (0.34) | 4.25 (1.19) | 3.83 (-0.45) | 7.35 (-1.91) | N/A | N/A |
| 60 - 69 | 1.59 (0.0) | 2.2 (0.0) | 1.15 (0.44) | 1.57 (0.63) | 1.76 (-0.17) | 3.35 (-1.15) | N/A | N/A |
| 70 - 79 | 0.71 (0.0) | 0.91 (0.0) | 0.29 (0.42) | 0.4 (0.51) | 0.36 (0.35) | 0.81 (0.1) | N/A | N/A |
| 80 - 89 | 0.39 (0.0) | 0.48 (0.0) | N/A | 0.29 (0.19) | N/A | N/A | N/A | N/A |

**Crude Prevalence of Bipolar Disorder in Georgia Subpopulations*** Several negative values observed in the "Other Race" subpopulations suggest higher prevalence rates of bipolar disorder. Asian subpopulations were very poorly represented by this data.

Crude Prevalence of Depression in Georgia Subpopulations
A White
B Black or African American
C Other Race
D Asian
Crude Prevalence (%)
Age Group
Colors
Female
Male

| | White | | Black or African American | | Other Race | | Asian | |
|---|---|---|---|---|---|---|---|---|
| Age Groups | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) |
| 0 - 9 | 0.43 (0.0) | 0.36 (0.0) | 0.37 (0.06) | 0.31 (0.05) | 0.23 (0.2) | 0.18 (0.18) | N/A | N/A |
| 10 - 19 | 3.67 (0.0) | 6.45 (0.0) | 2.6 (1.07) | 3.79 (2.66) | 3.57 (0.1) | 4.28 (2.17) | 0.78 (2.89) | 1.65 (4.8) |
| 20 - 29 | 6.43 (0.0) | 7.75 (0.0) | 7.02 (-0.59) | 4.32 (3.43) | 4.67 (1.76) | 4.74 (3.01) | 4.76 (1.67) | 2.53 (5.22) |
| 30 - 39 | 8.71 (0.0) | 14.28 (0.0) | 6.62 (2.09) | 7.82 (6.44) | 5.97 (2.74) | 7.35 (6.91) | 2.21 (6.5) | 2.61 (11.65) |
| 40 - 49 | 9.09 (0.0) | 17.79 (0.0) | 7.41 (1.68) | 11.31 (6.48) | 9.18 (-0.09) | 14.52 (3.27) | 3.02 (6.07) | 5.8 (11.99) |
| 50 - 59 | 8.29 (0.0) | 14.41 (0.0) | 7.06 (1.22) | 11.41 (3.0) | 9.24 (-0.96) | 16.36 (-1.95) | N/A | 8.82 (5.59) |
| 60 - 69 | 4.8 (0.0) | 7.12 (0.0) | 3.85 (0.95) | 5.16 (1.96) | 6.12 (-1.32) | 10.57 (-3.45) | 1.27 (3.53) | 2.31 (4.81) |
| 70 - 79 | 2.93 (0.0) | 3.91 (0.0) | 1.52 (1.41) | 2.11 (1.8) | 2.09 (0.84) | 3.9 (0.01) | 1.07 (1.86) | 1.69 (2.22) |
| 80 - 89 | 2.82 (0.0) | 3.63 (0.0) | 1.93 (0.89) | 1.94 (1.69) | 1.25 (1.57) | 2.35 (1.28) | N/A | 2.11 (1.52) |

**Crude Prevalence of Depression in Georgia Subpopulations*** In the "Other Race" subpopulations, some negative values are observed. Calculated metrics across subpopulations were reported for nearly every examined subpopulation.

Crude Prevalence of Suicidality in Georgia Subpopulations
A White
B Black or African American
C Other Race
D Asian
Crude Prevalence (%)
Age Group
Colors
Female
Male

| | White | | Black or African American | | Other Race | | Asian | |
|---|---|---|---|---|---|---|---|---|
| Age Groups | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) | Male Prev. (%) | Female Prev. (%) |
| 0 - 9 | 0.03 (0.0) | 0.02 (0.0) | 0.04 (-0.01) | 0.02 (0.0) | 0.04 (-0.01) | N/A | N/A | N/A |
| 10 - 19 | 0.65 (0.0) | 1.13 (0.0) | 0.44 (0.21) | 0.72 (0.41) | 0.88 (-0.23) | 1.06 (0.07) | N/A | 0.29 (0.84) |
| 20 - 29 | 1.09 (0.0) | 0.58 (0.0) | 1.49 (-0.4) | 0.45 (0.13) | 1.04 (0.05) | 0.8 (-0.22) | N/A | N/A |
| 30 - 39 | 1.24 (0.0) | 1.05 (0.0) | 1.38 (-0.14) | 0.54 (0.51) | 1.45 (-0.21) | 0.99 (0.06) | N/A | 0.59 (0.46) |
| 40 - 49 | 1.12 (0.0) | 1.19 (0.0) | 1.46 (-0.34) | 0.74 (0.45) | 1.85 (-0.73) | 1.61 (-0.42) | N/A | N/A |
| 50 - 59 | 0.91 (0.0) | 0.72 (0.0) | 1.09 (-0.18) | 0.71 (0.01) | 1.26 (-0.35) | 1.35 (-0.63) | N/A | N/A |
| 60 - 69 | 0.36 (0.0) | 0.25 (0.0) | 0.32 (0.04) | 0.21 (0.04) | 0.61 (-0.25) | 0.49 (-0.24) | N/A | N/A |
| 70 - 79 | N/A | N/A | N/A | 0.05 (N/A) | N/A | 0.14 (N/A) | N/A | N/A |
| 80 - 89 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |

**Crude Prevalence of Depression in Georgia Subpopulations*** Several negative values were observed for the "Other Race" subpopulations. Interestingly, several negative values for only the "Black or African American" male subpopulations is observed.

* Values in ()'s represent difference in prevalence rates between that subpopulation and its analogous white subpopulation. The more negative the value (highlighted red), the higher the compared subpopulation prevalence rate was observed. "N/A" values are those subpopulations that had to either be suppressed due to privacy considerations or were not represented in this data.

### Conclusions

Based on this exploratory approach, Georgia Medicaid subpopulations with chronic mental illness could face inequitable conditions. Future work includes examining patients' follow-up to care patterns to assess access to care and diagnostic practices. Possible factors to be examined in this process could be smaller geographical regions, patient visit types, and other factors. Finally, scrutinizing overall representativeness or fairness in subpopulations from data such as this could be explored.

[1] R. Crosby, M. L. Wendel, R. C. Vanderpool, and B. R. Casey, *Rural populations and health: Determinants, disparities, and solutions*. John Wiley & Sons, 2012.
[2] T. Y. Sun, S. Bhave, J. Altosaar, and N. Elhadad, "Assessing Phenotype Definitions for Algorithmic Fairness," *ArXiv220305174 Cs Q-Bio*, Mar. 2022, Accessed: Apr. 29, 2022. [Online].
[3] S. L. Cutter, B. J. Boruff, and W. L. Shirley, "Social vulnerability to environmental hazards," in *Hazards vulnerability and environmental justice*, Routledge, 2012, pp. 143–160.
[4] A. Soni, M. Hendryx, and K. Simon, "Medicaid expansion under the Affordable Care Act and insurance coverage in rural and urban areas," *J. Rural Health*, vol. 33, no. 2, pp. 217–226, 2017.
[5] J. Warren and K. B. Smalley, "Rural public health: Best practices and preventive models," 2014.
[6] L. Shi and G. D. Stevens, *Vulnerable populations in the United States*. John Wiley & Sons, 2021.
[7] D. M. Gray, A. Anyane-Yeboa, S. Balzora, R. B. Issaka, and F. P. May, "COVID-19 and the other pandemic: populations made vulnerable by systemic inequity," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 17, no. 9, pp. 520–522, Sep. 2020.
[8] W. C. Reeves *et al.*, "Mental illness surveillance among adults in the United States," 2011.
[9] Centers for Disease Control and Prevention, "CDC/ATSDR Social Vulnerability Index 2018 Database US."
[10] E. Jensen *et al.*, "The Chance That Two People Chosen at Random Are of Different Race or Ethnicity Groups Has Increased Since 2010," United States Census Bureau, Aug. 2021. Accessed: Sep. 11, 2022. [Online].

A Pilot Characterization Study Assessing Health Equity in Mental Healthcare Delivery within the State of Georgia (Jacob Zelko, Malina Hy, Varshini Chinta, Emily Liau, Morgan Knowlton, Jon Duke)