OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Meet the 2024 Titans

OHDSI Community Call
Nov. 5, 2024 • 11 am ET

Upcoming Community Calls

Date	Topic
Nov. 5	Meet The 2024 Titans
Nov. 12	Next Steps in Evidence Dissemination
Nov. 19	Evidence Network in Action: Semiglutide Study
Nov. 26	Collaborator Showcase Honorees
Dec. 3	Recent OHDSI Publications
Dec. 10	How Did We Do In 2024?
Dec. 17	Holiday-Themed Final Call of 2024

Three Stages of The Journey

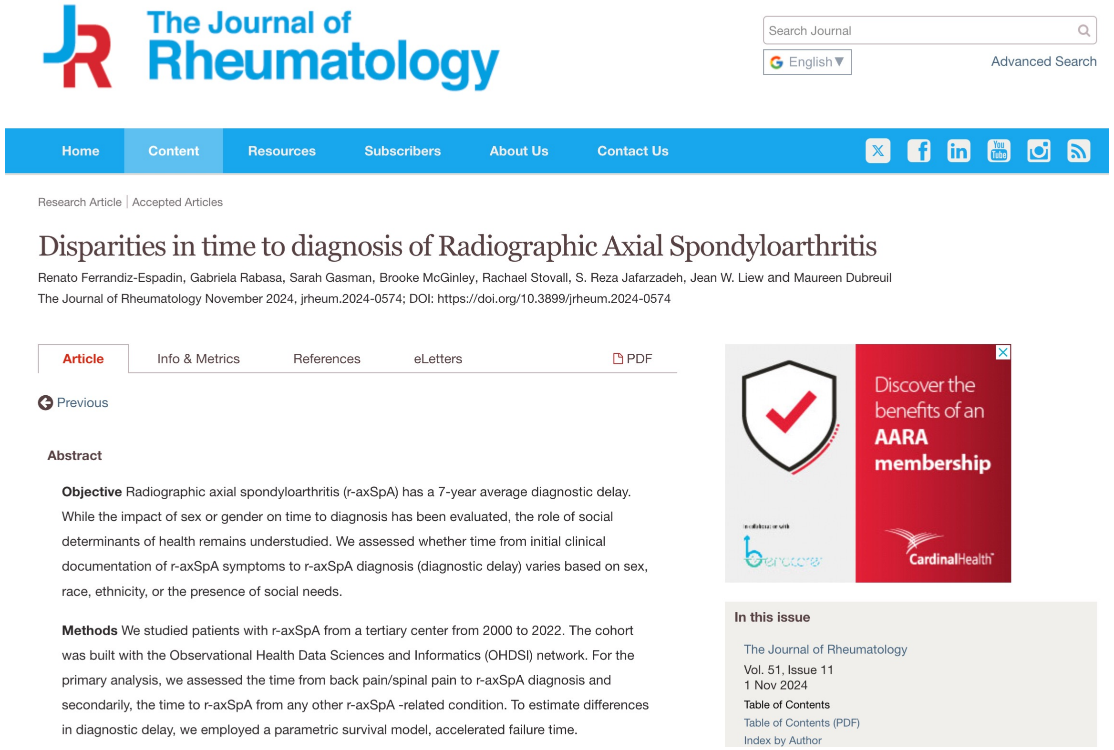
Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Renato Ferrandiz-Espadin, Gabriela Rabasa, Sarah Gasman, Brooke McGinley, Rachael Stovall, S. Reza Jafarzadeh, Jean W. Liew and Maureen Dubreuil** on the publication of **Disparities in time to diagnosis of Radiographic Axial Spondyloarthritis** in *The Journal of Rheumatology*.



The Journal of Rheumatology

Search Journal

English

Advanced Search

Home | Content | Resources | Subscribers | About Us | Contact Us

Research Article | Accepted Articles

Disparities in time to diagnosis of Radiographic Axial Spondyloarthritis

Renato Ferrandiz-Espadin, Gabriela Rabasa, Sarah Gasman, Brooke McGinley, Rachael Stovall, S. Reza Jafarzadeh, Jean W. Liew and Maureen Dubreuil

The Journal of Rheumatology November 2024, jrheum.2024-0574; DOI: https://doi.org/10.3899/jrheum.2024-0574

Article | Info & Metrics | References | eLetters | PDF

Previous

Abstract

Objective Radiographic axial spondyloarthritis (r-axSpA) has a 7-year average diagnostic delay. While the impact of sex or gender on time to diagnosis has been evaluated, the role of social determinants of health remains understudied. We assessed whether time from initial clinical documentation of r-axSpA symptoms to r-axSpA diagnosis (diagnostic delay) varies based on sex, race, ethnicity, or the presence of social needs.

Methods We studied patients with r-axSpA from a tertiary center from 2000 to 2022. The cohort was built with the Observational Health Data Sciences and Informatics (OHDSI) network. For the primary analysis, we assessed the time from back pain/spinal pain to r-axSpA diagnosis and secondarily, the time to r-axSpA from any other r-axSpA -related condition. To estimate differences in diagnostic delay, we employed a parametric survival model, accelerated failure time.

In this issue

The Journal of Rheumatology

Vol. 51, Issue 11

1 Nov 2024

Table of Contents

Table of Contents (PDF)

Index by Author

OHDSI Shoutouts!

Congratulations to the team of **Jiayi Tong, Lu Li, Jenna Marie Reps, Vitaly Lorman, Naimin Jing, Mackenzie Edmondson, Xiwei Lou, Ravi Jhaveri, Kelly J. Kelleher, Nathan M. Pajor, Christopher B. Forrest, Jiang Bian, Haitao Chu, and Yong Chen** on the publication of **Advancing Interpretable Regression Analysis for Binary Data: A Novel Distributed Algorithm Approach** in *Statistics in Medicine*.

Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Advancing Interpretable Regression Analysis for Binary Data: A Novel Distributed Algorithm Approach

Jiayi Tong[1,2] | Lu Li[1,3] | Jenna Marie Reps[4,5,6] | Vitaly Lorman[7] | Naimin Jing[8] | Mackenzie Edmondson[8] | Xiwei Lou[9] | Ravi Jhaveri[10] | Kelly J. Kelleher[11] | Nathan M. Pajor[12] | Christopher B. Forrest[7] | Jiang Bian[9] | Haitao Chu[13] | Yong Chen[1,2,14,15,16,17]

Correspondence: Yong Chen (ychen123@upenn.edu)

Received: 5 May 2023 | **Revised:** 4 August 2024 | **Accepted:** 2 October 2024

Funding: This work was supported by Patient-Centered Outcomes Research Institute, ME-2018C3-14899, ME-2019C3-18315 and National Institutes of Health, U01TR003709, U24MH136069, RF1AG077820, 1R01LM014344, 1R01AG077820, R01LM012607, R01AI130460, R01AG073435, R56AG074604, R01LM013519, R01DK128237, R56AG069880, R21AI167418, R21EY034179.

Keywords: binary data | distributed algorithm | modified Poisson regression | relative risk

ABSTRACT

Sparse data bias, where there is a lack of sufficient cases, is a common problem in data analysis, particularly when studying rare binary outcomes. Although a two-step meta-analysis approach may be used to lessen the bias by combining the summary statistics to increase the number of cases from multiple studies, this method does not completely eliminate bias in effect estimation. In this paper, we propose a one-shot distributed algorithm for estimating relative risk using a modified Poisson regression for binary data, named ODAP-B. We evaluate the performance of our method through both simulation studies and real-world case analyses of postacute sequelae of SARS-CoV-2 infection in children using data from 184 501 children across eight national academic medical centers. Compared with the meta-analysis method, our method provides closer estimates of the relative risk for all outcomes considered including syndromic and systemic outcomes. Our method is communication-efficient and privacy-preserving, requiring only aggregated data to obtain relatively unbiased effect estimates compared with two-step meta-analysis methods. Overall, ODAP-B is an effective distributed learning algorithm for Poisson regression to study rare binary outcomes. The method provides inference on adjusted relative risk with a robust variance estimator.

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

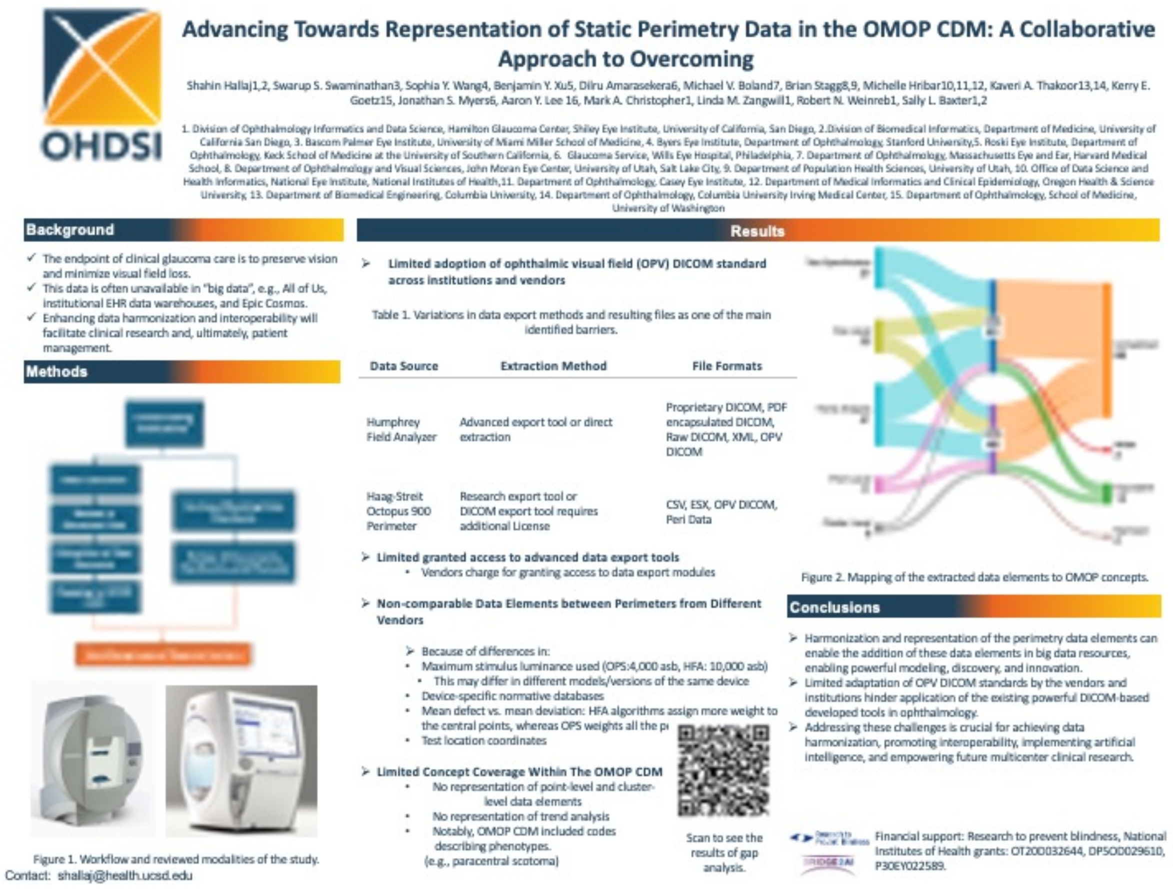
Date	Time (ET)	Meeting
Wednesday	8 am	Psychiatry
Wednesday	4 pm	Joint Vulcan/OHDSI Meeting
Thursday	9:30 am	Themis
Thursday	11 am	Industry
Thursday	12 pm	Methods Research
Thursday	7 pm	Dentistry
Friday	9 am	Phenotype Development & Evaluation
Friday	10 am	GIS-Geographic Information System
Friday	11:30 am	Steering
Friday	11:30 am	Clinical Trials
Friday	11 pm	China Chapter
Monday	10 am	CDM Survey Subgroup
Monday	10 am	Africa Chapter

#OHDSI2024 Showcase Honors

Observational Data Standards and Management

Gap Analysis of Static Automated Perimetry Concept Representation in OMOP CDM

(**Shahin Hallaj**, William Halfpenny, Niloofar Radgoudarzi, Michael V. Boland, Swarup S. Swaminathan, Sophia Y. Wang, Benjamin Y. Xu, Dilru C. Amarasekera, Brian Stagg, Michelle Hribar, Kaveri A. Thakoor, Kerry E. Goetz, Jonathan S. Myers, Aaron Y. Lee, Mark A. Christopher, Linda M. Zangwill, Robert N. Weinreb, Sally L. Baxter)

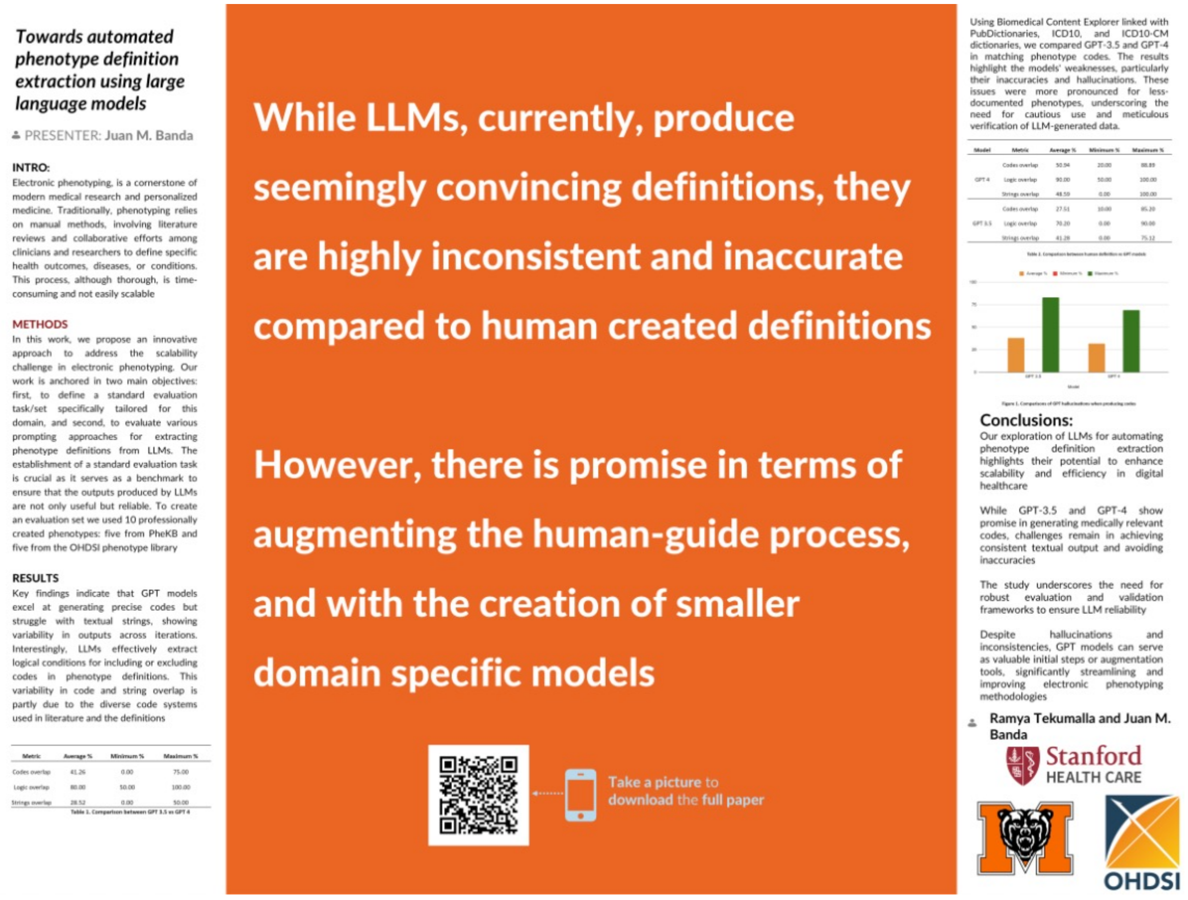


#OHDSI2024 Showcase Honors

Methodological Research

Towards automated phenotype definition extraction using large language models

(Ramya Tekumalla, Juan M. Banda)



Towards automated phenotype definition extraction using large language models

PRESENTER: Juan M. Banda

INTRO:
Electronic phenotyping, is a cornerstone of modern medical research and personalized medicine. Traditionally, phenotyping relies on manual methods, involving literature reviews and collaborative efforts among clinicians and researchers to define specific health outcomes, diseases, or conditions. This process, although thorough, is time-consuming and not easily scalable

METHODS
In this work, we propose an innovative approach to address the scalability challenge in electronic phenotyping. Our work is anchored in two main objectives: first, to define a standard evaluation task/set specifically tailored for this domain, and second, to evaluate various prompting approaches for extracting phenotype definitions from LLMs. The establishment of a standard evaluation task is crucial as it serves as a benchmark to ensure that the outputs produced by LLMs are not only useful but reliable. To create an evaluation set we used 10 professionally created phenotypes: five from PheKB and five from the OHDSI phenotype library

RESULTS
Key findings indicate that GPT models excel at generating precise codes but struggle with textual strings, showing variability in outputs across iterations. Interestingly, LLMs effectively extract logical conditions for including or excluding codes in phenotype definitions. This variability in code and string overlap is partly due to the diverse code systems used in literature and the definitions

Metric	Average %	Minimum %	Maximum %
Codes overlap	41.26	0.00	75.00
Logic overlap	80.00	50.00	100.00
Strings overlap	28.52	0.00	50.00

Table 1. Comparison between GPT 3.5 vs GPT 4

While LLMs, currently, produce seemingly convincing definitions, they are highly inconsistent and inaccurate compared to human created definitions

However, there is promise in terms of augmenting the human-guide process, and with the creation of smaller domain specific models

Take a picture to download the full paper

Using Biomedical Content Explorer linked with PubDictionaries, ICD10, and ICD10-CM dictionaries, we compared GPT-3.5 and GPT-4 in matching phenotype codes. The results highlight the models' weaknesses, particularly their inaccuracies and hallucinations. These issues were more pronounced for less-documented phenotypes, underscoring the need for cautious use and meticulous verification of LLM-generated data.

Model	Metric	Average %	Minimum %	Maximum %
GPT 4	Codes overlap	50.94	20.00	88.89
	Logic overlap	90.00	50.00	100.00
	Strings overlap	48.59	0.00	100.00
GPT 3.5	Codes overlap	27.51	10.00	85.20
	Logic overlap	70.20	0.00	90.00
	Strings overlap	41.28	0.00	75.12

Table 2. Comparison between human definition vs GPT models

Figure 1. Comparisons of GPT hallucinations when producing codes

Conclusions:
Our exploration of LLMs for automating phenotype definition extraction highlights their potential to enhance scalability and efficiency in digital healthcare

While GPT-3.5 and GPT-4 show promise in generating medically relevant codes, challenges remain in achieving consistent textual output and avoiding inaccuracies

The study underscores the need for robust evaluation and validation frameworks to ensure LLM reliability

Despite hallucinations and inconsistencies, GPT models can serve as valuable initial steps or augmentation tools, significantly streamlining and improving electronic phenotyping methodologies

Ramya Tekumalla and Juan M. Banda

Stanford HEALTH CARE

OHDSI

#OHDSI2024 Showcase Honors

Open-Source Anayltics Development

Bridging the Language Gap: Generative Models for Efficient Medical Concept Discovery

(**Alvaro A Alvarez**, Priya Desai, Somalee Datta)

Bridging the Language Gap
Generative Models for Efficient Medical Concept Discovery

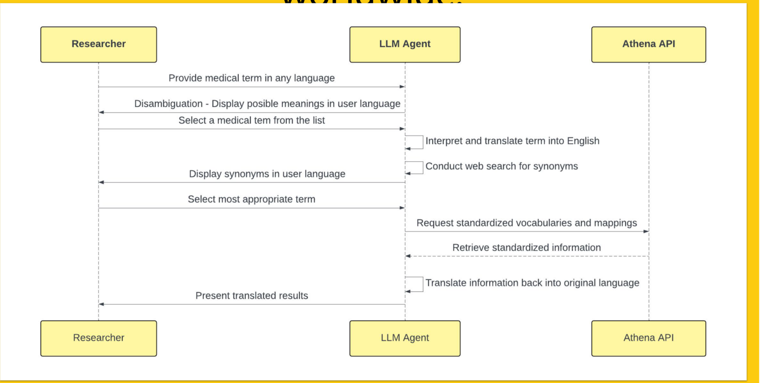
PRESENTER: **Alvaro A. Alvarez**

INTRO:
- Athena is crucial for OHDSI researchers, providing access to medical vocabularies.
- Researchers struggle to find correct medical concepts, especially with language barriers.
- Direct translations of medical terms can be ambiguous. For example, the Polish word "zawał" can mean either myocardial infarction or cerebral infarction, while the Spanish word "constipado" can refer to either a cold or constipation.
- **Lack of multilingual support hinders accessibility for non-English speaking researchers.**

METHODS
1. Developed a modular, AI-powered solution for medical concept discovery.
2. Uses Gpt 4o model (model-agnostic design for future upgrades).
3. Interprets input terms considering context and language-specific nuances.
4. Conducts web search for definitions and synonyms.
5. Communicates with Athena API to retrieve relevant medical concepts.
6. Translates results back to the user's original language.

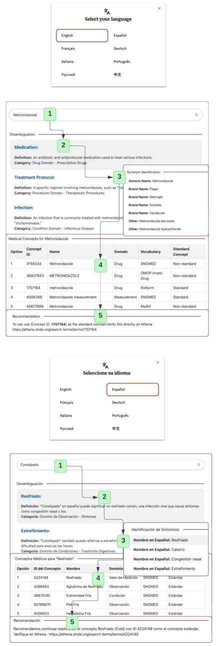
RESULTS
- Enhanced efficiency in locating relevant medical concepts.
- Improved multilingual support and handling of language-specific ambiguities.
- More equitable access for researchers with limited English proficiency.
- Seamless integration with OHDSI's ecosystem.

Generative models can **bridge the language gap** in medical concept discovery, making OHDSI tools more **accessible and efficient** for researchers worldwide.



Try the app here !

Source Code



Alvaro A Alvarez, Priya Desai, Somalee Datta

Stanford MEDICINE
Technology & Digital Solutions

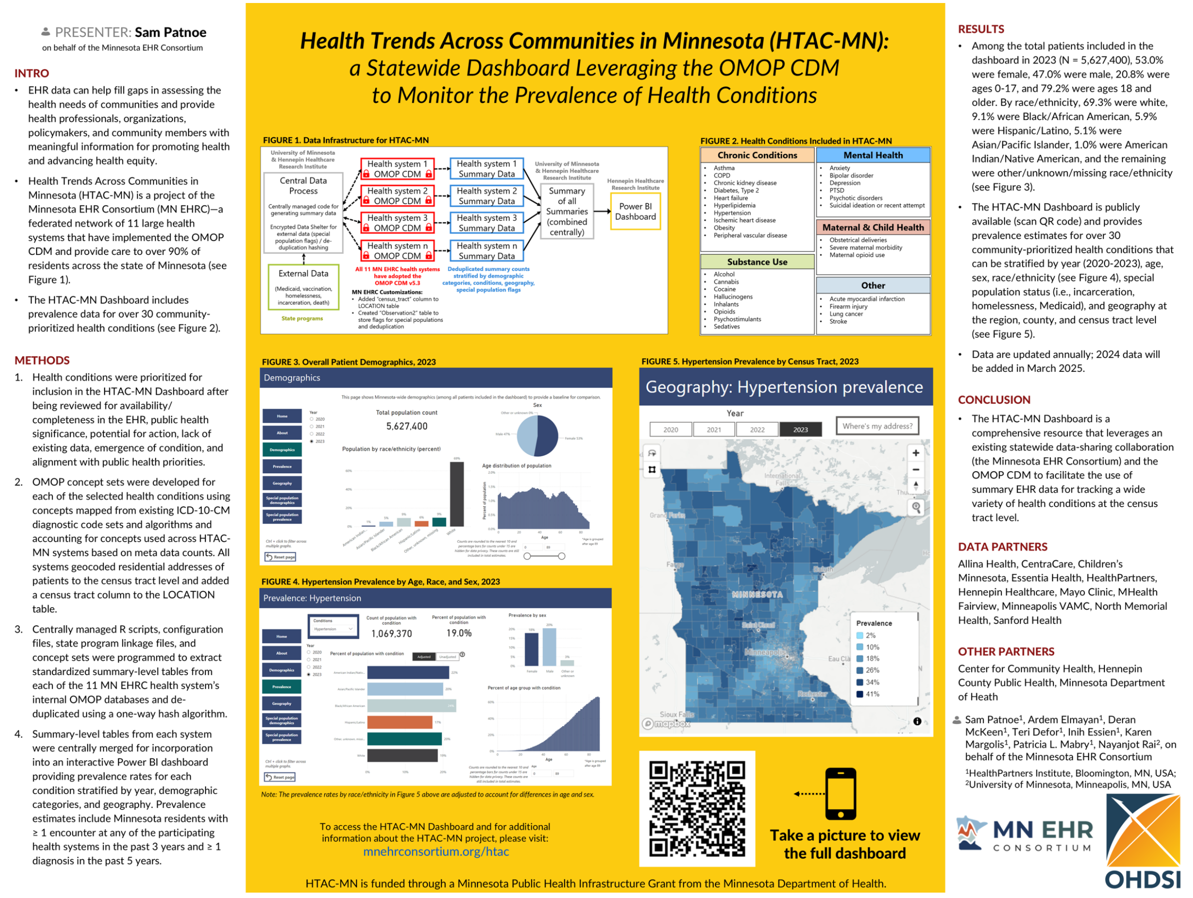
OHDSI

#OHDSI2024 Showcase Honors

Clinical Applications

Health Trends Across Communities in Minnesota: a Statewide Dashboard Leveraging the OMOP CDM to Monitor the Prevalence of Health Conditions

(Samuel T. Patnoe, Ardem S. Elmayan, Deran A. McKeen, Terese A. DeFor, Inih J. Essien, Karen L. Margolis, Patricia L. Mabry, Bjorn C. Westgard, Anna R. Bergdall, Renee Van Siclen, Peter J. Bodurtha, Daniel Muldoon, Tyler NA Winkelman, Nayanjot K. Rai, Paul E. Drawz, R. Adams Dudley, Steven G. Johnson, Stephen C. Waring, Alanna M. Chamberlain, Amy Leite Bennett, Abby Jessen, David Johnson, on behalf of the Minnesota Electronic Health Record Consortium



PRESENTER: Sam Patnoe
on behalf of the Minnesota EHR Consortium

INTRO

- EHR data can help fill gaps in assessing the health needs of communities and provide health professionals, organizations, policymakers, and community members with meaningful information for promoting health and advancing health equity.
- Health Trends Across Communities in Minnesota (HTAC-MN) is a project of the Minnesota EHR Consortium (MN EHRC)—a federated network of 11 large health systems that have implemented the OMOP CDM and provide care to over 90% of residents across the state of Minnesota (see Figure 1).
- The HTAC-MN Dashboard includes prevalence data for over 30 community-prioritized health conditions (see Figure 2).

METHODS

1. Health conditions were prioritized for inclusion in the HTAC-MN Dashboard after being reviewed for availability/completeness in the EHR, public health significance, potential for action, lack of existing data, emergence of condition, and alignment with public health priorities.
2. OMOP concept sets were developed for each of the selected health conditions using concepts mapped from existing ICD-10-CM diagnostic code sets and algorithms and accounting for concepts used across HTAC-MN systems based on meta data counts. All systems geocoded residential addresses of patients to the census tract level and added a census tract column to the LOCATION table.
3. Centrally managed R scripts, configuration files, state program linkage files, and concept sets were programmed to extract standardized summary-level tables from each of the 11 MN EHRC health system's internal OMOP databases and de-duplicated using a one-way hash algorithm.
4. Summary-level tables from each system were centrally merged for incorporation into an interactive Power BI dashboard providing prevalence rates for each condition stratified by year, demographic categories, and geography. Prevalence estimates include Minnesota residents with ≥ 1 encounter at any of the participating health systems in the past 3 years and ≥ 1 diagnosis in the past 5 years.

Health Trends Across Communities in Minnesota (HTAC-MN): a Statewide Dashboard Leveraging the OMOP CDM to Monitor the Prevalence of Health Conditions

FIGURE 1. Data Infrastructure for HTAC-MN

FIGURE 2. Health Conditions Included in HTAC-MN

Chronic Conditions	Mental Health
Asthma, COPD, Chronic kidney disease, Diabetes, Type 2, Heart failure, Hyperlipidemia, Hypertension, Ischemic heart disease, Obesity, Peripheral vascular disease	Anxiety, Bipolar disorder, Depression, PTSD, Psychotic disorders, Suicidal ideation or recent attempt
Substance Use	**Maternal & Child Health**
Alcohol, Cannabis, Cocaine, Hallucinogens, Inhalants, Opioids, Psychostimulants, Sedatives	Obstetrical deliveries, Severe maternal morbidity, Maternal opioid use
	Other
	Acute myocardial infarction, Firearm injury, Lung cancer, Stroke

FIGURE 3. Overall Patient Demographics, 2023

FIGURE 4. Hypertension Prevalence by Age, Race, and Sex, 2023

FIGURE 5. Hypertension Prevalence by Census Tract, 2023

Note: The prevalence rates by race/ethnicity in Figure 5 above are adjusted to account for differences in age and sex.

To access the HTAC-MN Dashboard and for additional information about the HTAC-MN project, please visit: mnehrconsortium.org/htac

Take a picture to view the full dashboard

HTAC-MN is funded through a Minnesota Public Health Infrastructure Grant from the Minnesota Department of Health.

RESULTS

- Among the total patients included in the dashboard in 2023 (N = 5,627,400), 53.0% were female, 47.0% were male, 20.8% were ages 0-17, and 79.2% were ages 18 and older. By race/ethnicity, 69.3% were white, 9.1% were Black/African American, 5.9% were Hispanic/Latino, 5.1% were Asian/Pacific Islander, 1.0% were American Indian/Native American, and the remaining were other/unknown/missing race/ethnicity (see Figure 3).
- The HTAC-MN Dashboard is publicly available (scan QR code) and provides prevalence estimates for over 30 community-prioritized health conditions that can be stratified by year (2020-2023), age, sex, race/ethnicity (see Figure 4), special population status (i.e., incarceration, homelessness, Medicaid), and geography at the region, county, and census tract level (see Figure 5).
- Data are updated annually; 2024 data will be added in March 2025.

CONCLUSION

- The HTAC-MN Dashboard is a comprehensive resource that leverages an existing statewide data-sharing collaboration (the Minnesota EHR Consortium) and the OMOP CDM to facilitate the use of summary EHR data for tracking a wide variety of health conditions at the census tract level.

DATA PARTNERS

Allina Health, CentraCare, Children's Minnesota, Essentia Health, HealthPartners, Hennepin Healthcare, Mayo Clinic, MHealth Fairview, Minneapolis VAMC, North Memorial Health, Sanford Health

OTHER PARTNERS

Center for Community Health, Hennepin County Public Health, Minnesota Department of Heath

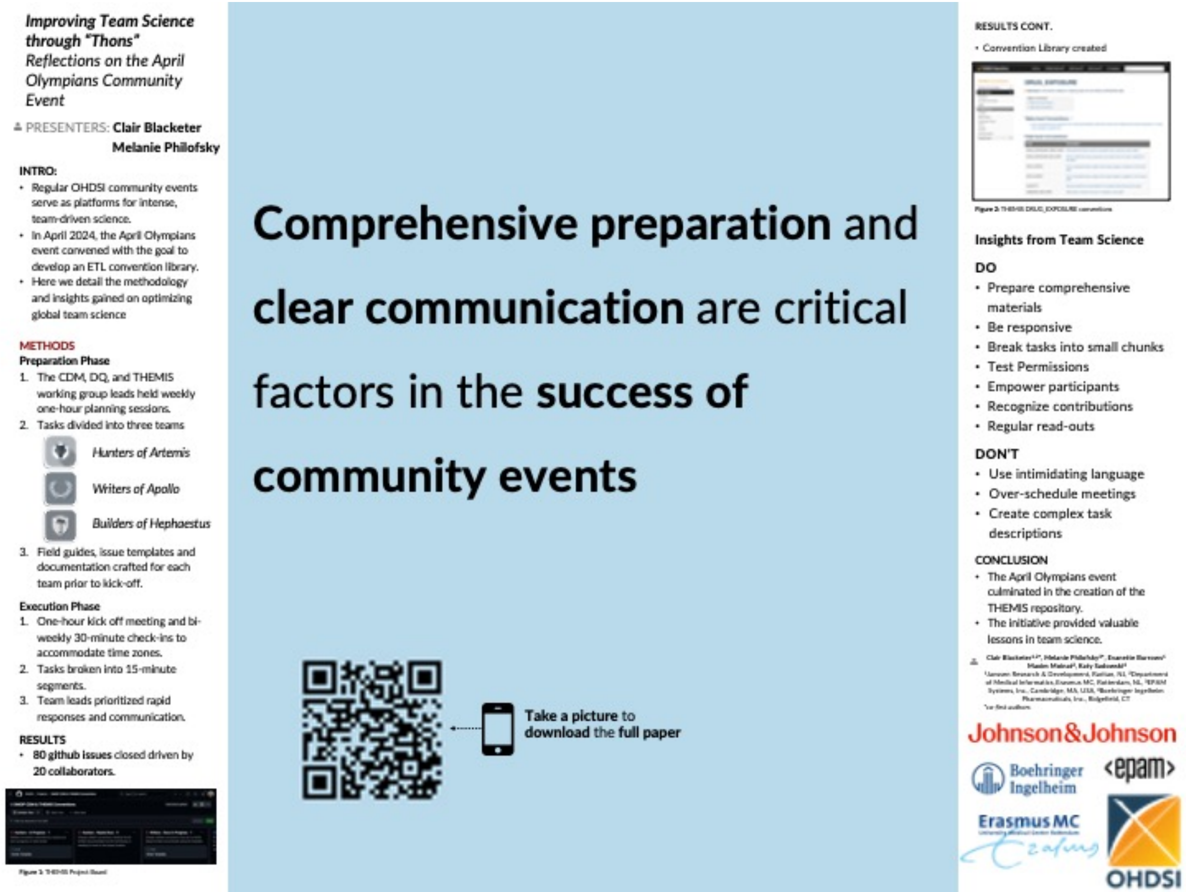
Sam Patnoe[1], Ardem Elmayan[1], Deran McKeen[1], Teri Defor[1], Inih Essien[1], Karen Margolis[1], Patricia L. Mabry[1], Nayanjot Rai[2], on behalf of the Minnesota EHR Consortium
[1]HealthPartners Institute, Bloomington, MN, USA; [2]University of Minnesota, Minneapolis, MN, USA

#OHDSI2024 Showcase Honors

Community

Improving Team Science Through “Thons” Reflections on the April Olympians Community Event

(Clair Blacketer, Melanie Philofsky, Evanette Burrows, Maxim Moinat, Katy Sadowski)



Next CBER Best Seminar: Nov. 20

Topic: Statistical methods for improving post-licensure vaccine safety surveillance

Presenter: Jennifer Clark Nelson, PhD, Director of Biostatistics & Senior Investigator, Biostatistics Division, Kaiser Permanente Washington Health Research Institute.

Date/Time: Nov. 20, 11 am ET

ohdsi.org/cber-best-seminar-series

The Center for Advanced Healthcare Research Informatics (CAHRI) at Tufts Medicine welcomes:

Agnes Kiragga, PhD

Lead - Data Science Program, African Population and Health Research Center (APHRC)

‘Promoting Data Science and Data Harmonization in Africa ’

November 21, 2024, 11am-12pm EST
Virtually via Zoom

Please contact Marty Alvarez at malvarez2@tuftsmedicalcenter.org for calendar invite or questions.

TuftsMedicine
Tufts Medical Center

NEI Eye Care and Ocular Imaging Challenge

NIH

NEI Expand OHDSI Initiative for Eye Care and Ocular Imaging Challenge

Submit your innovative ideas related to eye care and vision research for leveraging OHDSI.

This challenge seeks to expand the OHDSI network for vision research by incentivizing innovative ideas for leveraging real-world evidence. Prizes can support winner's integration into the network.

Key Dates and Challenge Timeline:

- Registration Period Open: August 26, 2024
- Mandatory Registration (intent to participate) Due: November 12, 2024
- Submission Period Open: December 1, 2024
- Submission Deadline: January 31, 2025
- Judging Start: February 10, 2025
- Judging End: March 24, 2025
- Winners Announced: April 2025

2024 APAC Symposium

Dec. 4-8 • Marina Bay Sands & National University of Singapore (NUS)

Dec. 4: Tutorial at NUS

Dec. 5-6: Main Conference at Marina Bay Sands

Dec. 7-8: Datathon at NUS

ohdsi.org/APAC2024

#OHDSISocialShowcase This Week

Monday

Dynamic Mapping Tools: Keeping Up to Date with Vocabulary Changes

(Melanie Philofsky, Hanan Shorrosh)

Dynamic Mapping Tools: Keeping Up to Date with Vocabulary Changes

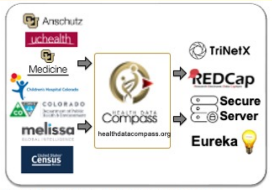
Melanie Philofsky, RN, MS[1,2*], Hanan Shorrosh[2*], Margaret Izzie Clinton[2], Jue Wang, MFM[2], Krista Miller, MS, MHA[2], Michael G. Kahn, MD, PhD[2], Michelle N. Edelmann, PhD[2], Ian M. Brooks, PhD[2]
[1]Odysseus EPAM [2]Health Data Compass, University of Colorado Anschutz Medical Campus *co-first authors

OHDSI | University of Colorado Anschutz Medical Campus | Health Data Compass | <epam> | ODYSSEUS DATA SERVICES

Background

Health Data Compass (HDC) is the enterprise clinical data warehouse at the University of Colorado Anschutz Medical Center (CU AMC), integrating patient data from a variety of hospital, state and public data sources.

HDC has identified the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) as its main data mart to utilize in the delivery of datasets to researchers.



With each new release of the OHDSI vocabulary data:

1. Concept IDs can change from "standard" to "non-standard" or vice versa. This results in data moving to a new field location.
2. Concept IDs can change domains which changes the table where data are located.

Problem: How do we account for the dynamic nature of concepts, especially for recurring reports?

Solution: We created two tools, the **concept mapping stored procedure** and the **concept mapping table.**

Methods

Call the mapping stored procedure

Declare output table variable
Input table name
Input table field names: Concept IDs, Codes, Vocabularies
Return descendent concepts? True/False
Output table variable

```
Declare vOutputTable string;

Call project.dataset.sp_concept_mapping(
'project.dataset.concepts_input',
'ID',
'Code',
'Vocabulary',
true,
vOutputTable);
```

Join to the mapping table

1. Create input table of concept IDs.
   a. Use stored procedure to check for errors before proceeding

Join input table to mapping table on concept ID

Join mapping table to OMOP CDM table on standard concept ID or source concept ID

```
with concept_mapped as (
  select distinct
   cm.concept_id,
   cm.standard_concept_id,
   cm.source_concept_id
  from `project.dataset.concept_mapped` cm
    inner join `project.dataset.concepts_input` c on
cm.concept_id = c.concept_id )
Select o.* from `project.dataset.observation` o
   cross join concept_mapped cm
WHERE (o.observation_concept_id = cm.standard_concept_id or
o.observation_source_concept_id = cm.source_concept_id)
```

Results

Mapping stored procedure

Error checking examples

ERROR 002: Icd-10 will be translated to ICD10CM; Code does not match vocab; Id does not match code / vocab; Missing vocab: Need Id or code & vocab; Id does not exist
ERROR 003: Classifications / ingredients: error if option to return descendent concepts is set to false
ERROR 004: Deprecated code; does not map

Id	Code	Vocabulary
	E08.10	Icd-10
	99214	HCPCS
44832532	648	ICD9CM
	J05AH01	
92616370		
1708748	69722	RxNorm
45889789	1007534	CPT4
35224711	Y60	ICD10CM

Output table

- Input concepts mapped to current standard concepts and their domains
- Input codes case-corrected and vocabularies translated
- Ingredients and classifications identified
- Descendent concepts mapped to current standard concepts and their domains (optional)

Mapping table

- All concepts mapped to current standard concepts and their domains
- If table is maintained with regular updates, reports joined to the table will always return current results

Conclusions

Early results:

- The team found the tools to be helpful in preventing errors of omission when delivering datasets to customers.

Example of issue	Consequence without tool	Tool solution
Concept ID changed from standard to non-standard	Data are omitted from the concept set, cohort, and study.	Concept mapping table and concept mapping stored procedure return current concept mappings and locations
Concept ID changed domains		
List of requested concepts contains a typo in the code or concept ID; variable of interest not found in concept table	Errors need to be manually identified and fixed. If errors not identified, data are omitted from the concept set, cohort, and study.	Concept mapping stored procedure identifies errors for correction
List of requested concepts contains slight errors in vocabulary name		Concept mapping stored procedure translates common variations of vocabularies to standardized versions
Descendent concepts are requested from list of classification concepts	Descendent concepts may be lost from the result set unless manually added	Concept mapping stored procedure returns each descendent concept in a separate row in the output table

- Future work includes measuring time saved and errors prevented by utilizing the two tools.

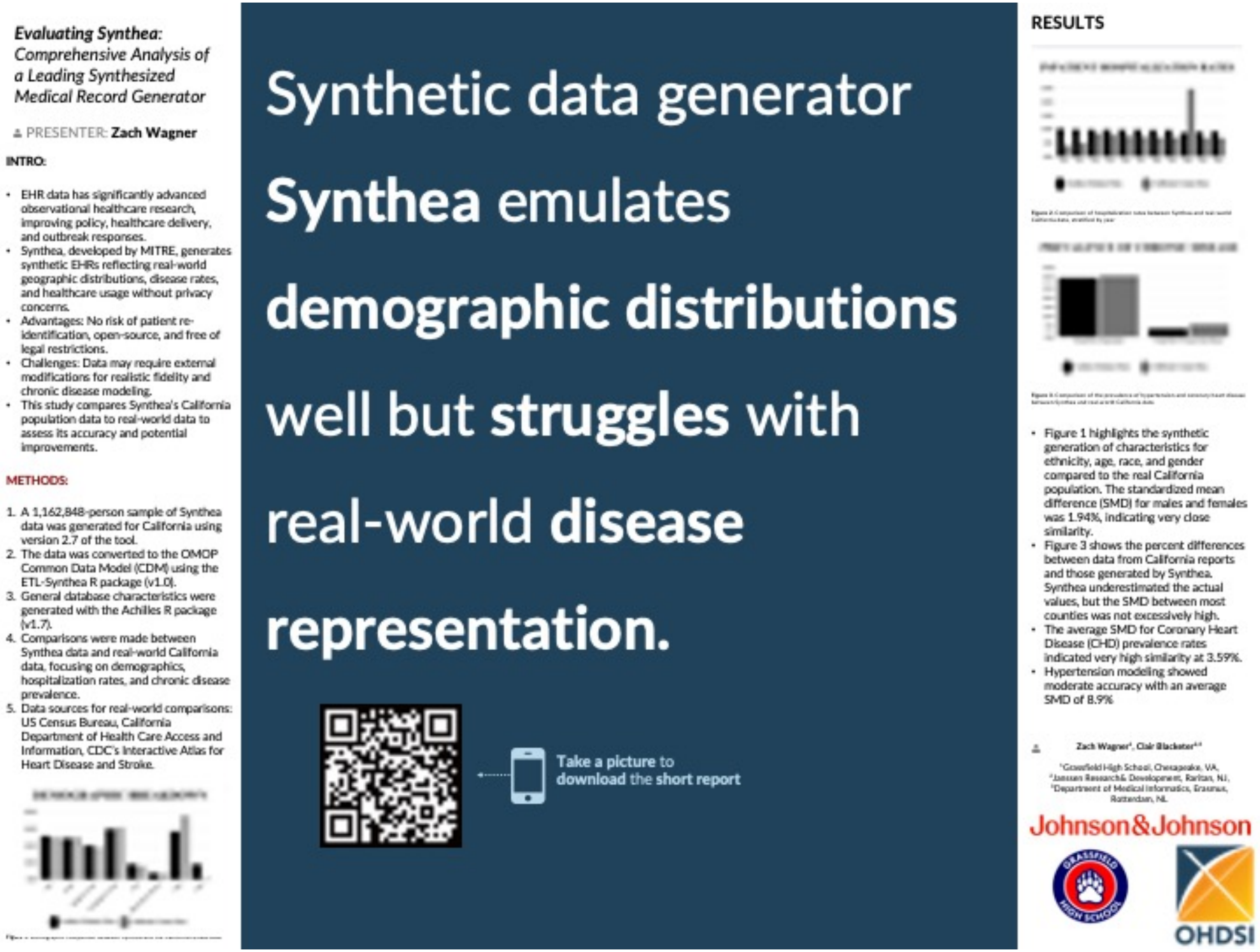
Contact: philofsky@ohdsi.org; healthdatacompass@ucdenver.edu

#OHDSISocialShowcase This Week

Tuesday

Evaluating Synthea: Comprehensive Analysis of a Leading Synthesized Medical Record Generator

(Zach Wagner, Clair Blacketer)



#OHDSISocialShowcase This Week

Wednesday

Estimation of Causal Effects under Treatment Misclassification: A Semi-Parametric Bias Correction Framework with Application to Vaccine Effectiveness Study

(Qiong Wu, Huiyuan Wang, Yong Chen)

OHDSI

Estimation of Causal Effects under Treatment Misclassification: A Semi-Parametric Bias Correction Framework with Application to Vaccine Effectiveness Study

Qiong Wu[a,b,c], Huiyuan Wang[b,c], and Yong Chen[b,c]

a. Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, PA, USA
b. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
c. The Center for Health Analytics and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA

Contact: qiongwu@pitt.edu and ychen123@pennmedicine.upenn.edu

University of Pittsburgh School of Public Health

Perelman School of Medicine University of Pennsylvania

PennCIL A Computing · Inference · Learning lab at University of Pennsylvania

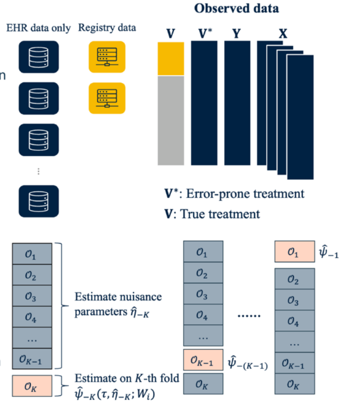
CHASE Center for Health Analytics & Synthesis of Evidence

Background

- **Clinical Inquiry**:
  - How does a COVID-19 vaccine administered prior to infection impact long COVID risks
- **Existing Knowledge**:
  - Studies focus primarily on *adult populations.*
  - *Inconsistent findings*: Range from significant protective effects, mixed outcomes, to counter-protective effects.
  - *Scope limitation*: Most studies assess effectiveness only in the infected population, by conditioning on infection status
- **Methodological Challenges**:
  - *Real-world effectiveness* using electronic health record (EHR) data
  - *Incomplete vaccination status* documentation within U.S. health systems

Methods

- Observed data $(V^*, RV, R, Y, \boldsymbol{X})$: :
  - Main dataset $(V^*, Y, \boldsymbol{X})$:
    - e.g., EHR data from PEDSnet
  - Internal validation dataset $(V^*, V, Y, \boldsymbol{X})$:
    - e.g., linked data with immunization registration database
- Estimand
  - Average treatment effect (ATE):
  $$\tau_0 = E(Y_1) - E(Y_0)$$
  where Y_1 and Y_0 are potential outcomes
- Goal:
  - Unbiased estimation of the estimand of interest
  - Use entire dataset for greater statistical efficiency
  - Minimize parametric assumptions

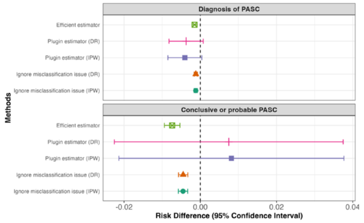


- Identification
$$E\left\{\frac{E(Y|\boldsymbol{X}, V^*=1) - E(Y|\boldsymbol{X}, V^*=0)}{P(V=1|\boldsymbol{X}, V^*=1) - P(V=1|\boldsymbol{X}, V^*=0)}\right\}.$$
  (weighted version of ATE based on V^*)
  - *Plugin estimator*:
    - Step 1: Estimate the misclassification model using internal validation data only.
    - Step 2: Sample weighting in estimating the ATE based on misclassified treatment status V^*
  - ***Efficient estimator***:
$$\begin{aligned}&\psi_{EIF}(\tau, \eta; O)\\ &= \frac{\tau^*(\boldsymbol{X})}{\delta(\boldsymbol{X})} + \frac{1}{\delta(\boldsymbol{X})}\left[\frac{V^*}{p(\boldsymbol{X})}\{Y - \mu_1(\boldsymbol{X})\} - \frac{1-V^*}{1-p(\boldsymbol{X})}\{Y - \mu_0(\boldsymbol{X})\}\right]\\ &- \frac{\tau^*(\boldsymbol{X})}{\delta(\boldsymbol{X})}\frac{1}{\delta(\boldsymbol{X})}\left[\frac{R}{\pi_1(\boldsymbol{X})}\frac{V^*}{p(\boldsymbol{X})}\{V - \alpha_1(\boldsymbol{X})\} - \frac{R}{\pi_0(\boldsymbol{X})}\frac{1-V^*}{1-p(\boldsymbol{X})}\{V - \alpha_0(\boldsymbol{X})\}\right] - \tau_0\end{aligned}$$
  ($\eta = \{p(\boldsymbol{X}), \mu_v(\boldsymbol{X}, V^*), \alpha_v(\boldsymbol{X}), \pi_v(\boldsymbol{X}): v = 0{,}1\}$ are nuisance parameters which can be estimated using machine learning methods.)

Results

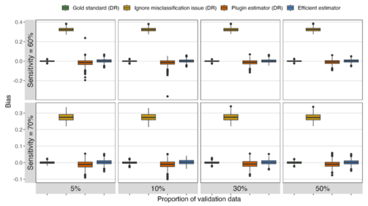
- **Data source**: Synthetic EHR data from summary statistics in Wu et. al., (2024)
- **Eligibility criteria**:
  - In the age group at the study start
  - No previous COVID-19 vaccination
  - No previous SARS-CoV-2 infection
  - User of the healthcare system of PEDSnet (i.e., having a primary care visit in the past 18 months)
- **Intervention**: BNT162b2 vaccine vs. no receipt of any type of COVID-19 vaccine
- **Outcomes**:
  - *Diagnosis of PASC*: Diagnosis based on ICD code U09.9.
  - *Conclusive or probable PASC*: Diagnosis from a computable phenotype algorithm based on ICD codes of PASC and long-COVID symptoms defined by clinicians.
- **Confounding variables**: demographic factors, clinical factors, and healthcare utilization factors (including the number of negative COVID-19 tests prior to the cohort entry).

	Misclassified treatment $V^* = 1$			
	True treatment $V = 1$			
	Immunization Information System (IIS)	Internal Administration	Patient reported	Other
Cincinnati	-	631	42	1510
CHOP	-	2135	3183	587
Colorado	1094	131	6	639
Lurie	-	164	488	0
Nationwide	-	785	1004	327
Nemours	-	250	0	3563
Seattle	-	704	0	0
Stanford	-	452	957	7



Simulation studies

- Misclassification setting
  - Differential misclassification
    - $V^*|V$ depends on part of covariates $\boldsymbol{X}$
  - Varied overall misclassification rates
    - The sensitivity ranges from 60% - 90%
    - The specificity ranges from 90% - 100%
- Size of validation data:
  - The proportion of internal validation data ranges from 5%-50%
- Evaluation metrics:
  - Bias, empirical standard error, coverage



Conclusions

- The novel pipeline produces a robust estimation of vaccine effectiveness while addressing incomplete vaccine records in EHR data due to the lack of immunization registry linkage.
- The research suggests a significant protective effectiveness of the BNT162b2 vaccine on long COVID risks during Omicron period based on a national pediatric cohort in the U.S.

Reference: Wu, Q., Tong, J., Zhang, B., Zhang, D., Chen, J., Lei, Y., … & Chen, Y. (2024). Real-world effectiveness of BNT162b2 against infection and severe diseases in children and adolescents. Annals of Internal Medicine, 177(2), 165-176.

#OHDSISocialShowcase This Week

Thursday

Predicting the risk of new onset of type 2 diabetes following exposure of Statin within patient with coronary artery disease

(Septi Melisa, Christianus Heru Setiawan, Muhammad Solihuddin Muhtar, Phan Thanh-Phuc, Nguyen Phung-Anh, Jason C. Hsu)

Predicting the risk of new onset of type 2 diabetes following exposure of Statin within patient with coronary artery disease

PRESENTER: **Septi Melisa**
d931111003@tmu.edu.tw

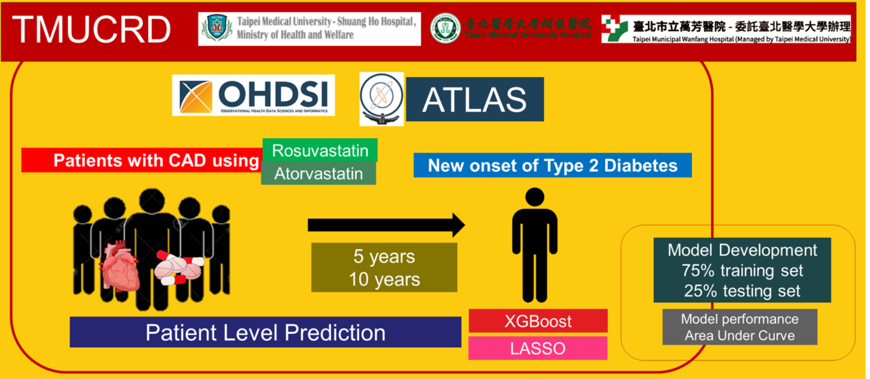
INTRO:

- Statins are essential in preventing atherosclerosis progression but are linked to an increased risk of type 2 diabetes.
- Understanding the balance between benefits and potential risks of statins is crucial.
- Aim: Develop a prediction model for new-onset type 2 diabetes in coronary artery disease patients on statins.

METHODS

1. Study Design: Retrospective cohort using Taipei Medical University Clinical Research Database (TMUCRD), which has been mapped into OMOP-CDM.
2. Population: Patients on rosuvastatin or atorvastatin, with coronary artery disease, aged >18, and no prior type 2 diabetes.
3. Data: 31,657 patients from 3 hospitals in Northern Taiwan.
   - Rosuvastatin: 11,084 patients.
   - Atorvastatin: 20,573 patients.
   - Exclusion: Patients with pre-existing diabetes.
4. Outcome: New-onset of type 2 diabetes, diagnosed 30+ days post-statin initiation.
5. Follow-up: 5 and 10 years.
6. Models development: 75% of training set and 25% of testing set to develop the model using Lasso Logistic Regression, Gradient Boosting Machine (XGBoost).
7. Tools: Atlas 2.13.0 and patient-level prediction package.

Statin therapy in coronary artery disease patients may increase the risk of new-onset type 2 diabetes; our model helps predict this risk, enabling early preventive interventions and personalized patient care.



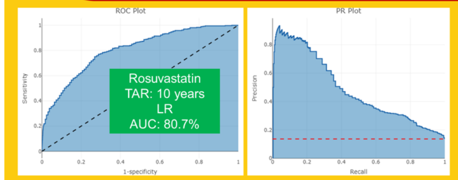


Figure 1. Area under curve (AUC) for predicting the risk of diabetes 10 years after rosuvastatin exposure, Testing AUC (left), Precision curve (right)

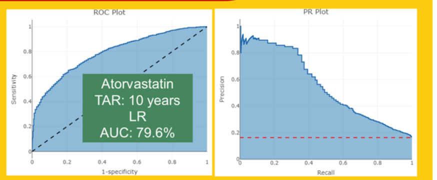


Figure 2. Area under curve (AUC) for predicting the risk of diabetes 10 years after atorvastatin exposure, Testing AUC (left), Precision curve (right)

Table 1. The model performance for predicting new onset of type 2 diabetes

Target	TAR	Model	Incidence Rate	Training AUC	Testing AUC	Sensitivity	Specificity	PPV	NPV
Rosuvastatin	5 years	LR	11.8%	81.1%	76.0%	68.7%	66.9%	21.7%	94.1%
		XGBoost		87.4%	75.7%	69%	68.1%	22.7%	94.2%
	10 years	LR	13.618%	81.4%	80.7%	75.1%	73.2%	30.6%	94.9%
		XGBoost		81.8%	80.0%	72.6%	71%	28.3%	94.3%
Atorvastatin	5 years	LR	13.399%	79%	77.6%	69%	68.1%	25%	93.4%
		XGBoost		86.2%	78.8%	70.5%	69.5%	26.3%	93.8%
	10 years	LR	16.274%	81.6%	79.6%	71.6%	71%	31.5%	92.8%
		XGBoost		79.9%	78.8%	70.2%	69.3%	30.7%	92.3%

RESULT:

Patient Characteristics:

- Rosuvastatin group (age 60-64, 16.53%).
- Atorvastatin group (age 55-59, 15.24%).
- Majority male in both cohorts.

Incidence of New-Onset Diabetes:

- Rosuvastatin: 13.6% (1,474/11,084)
- Atorvastatin: 16.2% (3,438/20,573)

Model Performance:

Best AUC: 80.7% with logistic regression for the rosuvastatin group, and 79.6% with logistic regression for the atorvastatin group, in predicting new-onset type 2 diabetes after 10 years of statin exposure.

Features:

- Demographic
- Chad2
- Chads2Vasc
- Charlson Comorbidity Index
- Drug group era
- Condition group era

CONCLUSION:

- We developed a prediction model using a Logistic regression algorithm to predict the new onset of T2DM among patients using statins with a history of coronary artery disease.
- This model can be applied in clinical practice to stratify patients by their risk of developing type 2 diabetes, facilitating early prevention and enabling personalized patient care.
- Through this collaboration showcase, we invite data partners within the OHDSI community to join us in validating these findings and strengthening the robustness of our study.

Septi Melisa, Christianus Heru Setiawan, Muhammad Solihuddin Muhtar, Phan Thanh-Phuc, Nguyen Phung-Anh, Jason C. Hsu

#OHDSISocialShowcase This Week

Friday

Aggregating and harmonizing registry databases for comparative analyses – lessons learnt

(Eva-Maria Didden, James Weaver, Dmytro Dymshyts, Amelie Beaudet, Audrey Muller, Andrius Kavaliunas)

Harmonizing and Aggregating Registry Databases for Comparative Analyses: Lessons Learnt

CO-AUTHORS: Eva-Maria Didden, James Weaver, Dmytro Dymshyts, Amelie Beaudet, Audrey Muller, Andrius Kavaliunas

PRESENTER: James Weaver

INTRO

- Pulmonary Arterial Hypertension (PAH) is a rare subgroup of Pulmonary Hypertension (PH).
- Real-world evidence (RWE) in PAH is limited by small, geographically dispersed populations and data access.
- Data harmonization by combining multiple fit-for-purpose data sources into one database has potential to mitigate this.
- This can enable comparative effect analyses that is otherwise infeasible.

Objective: Assess effectiveness of PAH triple combination relative to double combination therapy.
Effectiveness outcomes: Time to hospitalization, death, parenteral therapy, disease worsening

METHODS

Data pre-processing:

- 4 PH databases mapped to OMOP CDM [1,2].
- Data structure and content evaluation results sufficient to combined into 1 database (i.e., harmonize) [Table 1].

Analysis specifications:

Eligible patients: adult PAH patients.

Index date – Target cohort: add-on date of 3rd drug.
Index date – Comparator cohort: date of screening visit that would qualify patient for triple combination therapy, based on guidelines. (Ref. 3)

Statistical methods workflow:

1. Cohort characterization. inform
2. 1:1 PS matching: optimal matching & two greedy nearest neighbor matching approaches. prior to
3. Comparative effectiveness analyses.

ID	Design	Aim	Patient count	Study period	Region
1.	Prospective observational chohort study	Patients' characteristics, outcomes and safety	2674	April 2014 – June 2020	North America
2.	Retrospective chart review	Patients' characteristics and safety	3031	October 2013 – March 2017	North America
3.	Prospective observational chohort study	Patients' characteristics and outcomes	829	November 2016 - September 2021	North America
4.	Prospective observational chohort study	Patients' characteristics, outcomes and safety	2354	September 2017 - November 2021 (last available data cut)	North America and Europe

Table 1. Main characteristics of the data sources.

Harmonizing disparate data sources for adequately powered analysis creates opportunity but with limitations.

Reliable comparative effectiveness evidence requires meeting the exchangeability assumption after propensity-score adjustment, which our study did not achieve.

Let's collectively share our lessons-learnt!

RESULTS

1. Cohort characterization:
   - Target and comparator cohorts had similar demographics and clinical characteristics, but different geographic distribution.
   - Target cohort had substantially greater time from PAH diagnosis to index.
2. PS matching:
   - Covariates recorded across all databases were included in the PS model [Table 2a].
   - After PS matching 3 strategies, residual covariate imbalance persisted [Table 2b, Figure 1], including
     - ➢ time from PAH diagnosis to index, which is likely associated with study outcomes, making it a plausible confounder
     - ➢ several baseline conditions, making confounding by initial health status a plausible threat to validity

a) Covariates included in the propensity score models
Age, sex, index year, location, time between PAH Dx and index; mortality risk category, PAH etiology, # comorbidities at index, diabetes, hypertension, other cardiovascular disorders, cerebrovascular disorders, liver disorders, renal disorders, metabolic disorders, connective tissue disorders, autoimmune conditions
b) Post matching: imbalanced covariates with an absolute standardized mean difference of >0.1
Time between PAH Dx and index; cardiovascular disorders, cerebrovascular disorders, metabolic disorders, liver disease, renal disease, connective tissue disease, autoimmune conditions

Table 2. a) Covariates used for propensity score modelling b) Imbalanced covariates

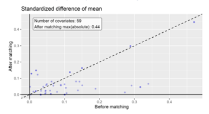


Figure 1. Standardized mean difference before and after matching

3. Comparative effectiveness analysis: not conducted because exchangeability diagnostic indicated confounding risk

CONCLUSION

- Clinically rich, harmonized PH database provided characterization evidence on patients exposed to different PAH treatment regimens
- PS matching strategies did not create adequately exchangeable exposure cohorts for valid comparative analyses
- Several limitations include:
  - ➢ Comparator cohort index date specification difficulty
  - ➢ Imbalanced time from PAH Dx to index date likely a strong confounder

This study demonstrated the feasibility of data harmonization across diverse sources in a rare disease area. However, we halted the comparative effectiveness analysis due to confounding risks identified in our exchangeability assessment. Diagnostic evaluations help prevent unreliable evidence dissemination, to the benefit of PAH patients and the RWE community.

REFERENCES:

1. Biedermann, P. et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol. 2021;21(1):238. doi: 10.1186/s12874-021-01434-3.
2. Handbook for PH registries to OMOP CDM conversion: https://github.com/OHDSI/ETL--PulmonaryHypertensionRegistries
3. Humbert M, et al., 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. European Respiratory Journal. 2023. DOI: 10.1183/13993003.00879-2022
4. Schuemie M, et al (2024). CohortMethod: New-User Cohort Method with Large Scale Propensity and Outcome Models. https://ohdsi.github.io/CohortMethod, https://github.com/OHDSI/CohortMethod.
5. Schuemie M, et al. Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research. Stud Health Technol Inform. 2024 Jan 25;310:966-970. doi: 10.3233/SHTI231108. PMID: 38269952; PMCID: PMC10868467.

Job Opening

Senior Program Officer, Clinical AI Innovation, Gates Foundation

Senior Program Officer, Clinical AI Innovation

Apply

Seattle, WA

Full time

Posted 6 Days Ago

B020184

The Foundation

We are the largest nonprofit fighting poverty, disease, and inequity around the world. Founded on a simple premise: people everywhere, regardless of identity or circumstances, should have the chance to live healthy, productive lives. We believe our employees should reflect the rich diversity of the global populations we aim to serve. We provide an exceptional benefits package to employees and their families which include comprehensive medical, dental, and vision coverage with no premiums, generous paid time off, paid family leave, foundation-paid retirement contribution, regional holidays, and opportunities to engage in several employee communities. As a workplace, we're committed to creating an environment for you to thrive both personally and professionally.

Your Role

Are you passionate about using the power of AI to reduce inequality in low- and middle-income countries? Do you have experience working in developing countries on AI and digital health initiatives? If so, we want you to join our team at the largest nonprofit fighting poverty, disease, and inequity around the world.

The Senior Program Officer, Clinical AI Innovation is a key member of the AI team. This role will support several teams at the Foundation who are considering and investing in multiple applications of AI in Health, which is a high priority area for the Foundation. As such, this individual will be responsible for developing our overarching strategy to healthcare applications in AI; conceptualising, investing and managing investments in health applications of AI; providing advice and technical assistance to other program teams considering investment in this area; advocate for the safe, responsible use of AI as force multiplier to reducing inequality in health in LMICs.

What You'll Do

Develop the foundations' approach to AI and health

- Ensure we have an approach to evaluation of clinical AI applications/ use cases
- This would include existing and planned investment in multiple applications of AI in health across diagnostics, end user engagement, decision support and decision sciences for health
- Develop a clear understanding of specific ecosystem constraints and opportunities related to AI in health
- Identify a key set of partners and stakeholders in order to be successful in this focus area across the technical, advocacy, government, academic and funding spheres

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?

The weekly OHDSI community call is held every Tuesday at 11 am ET.

Everybody is invited!

Links are sent out weekly and available at:
ohdsi.org/community-calls