Phenotype Phebruary + Workgroup 2025 OKRs

OHDSI Community Call
Feb. 4, 2025 • 11 am ET

Upcoming Community Calls

Date	Topic
Feb. 4	First Week of 2025 Workgroup OKRs/Phenotype Phebruary
Feb. 11	Second Week of 2025 Workgroup OKRs/Phenotype Phebruary
Feb. 18	Third Week of 2025 Workgroup OKRs/Phenotype Phebruary
Feb. 25	Fourth Week of 2025 Workgroup OKRs/Phenotype Phebruary
Mar. 4	Vocabulary Release Update, Winter 2025

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Gyubeom Hwang, So Hee Lee, Dong Yun Lee, ChulHyoung Park, Hyun Woong Roh, Sang Joon Son, and Rae Woong Park** on the publication of **Age-related eye diseases and subsequent risk of mental disorders in older adults: A real-world multicenter study** in the *Journal of Affective Disorders*.

Journal of Affective Disorders 375 (2025) 306–315

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Age-related eye diseases and subsequent risk of mental disorders in older adults: A real-world multicenter study

Gyubeom Hwang [a,1], So Hee Lee [b,1], Dong Yun Lee [a], ChulHyoung Park [a], Hyun Woong Roh [c], Sang Joon Son [c], Rae Woong Park [a,b,*,**]

[a] Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
[b] Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea
[c] Department of Psychiatry, Ajou University School of Medicine, Suwon, Republic of Korea

ARTICLE INFO

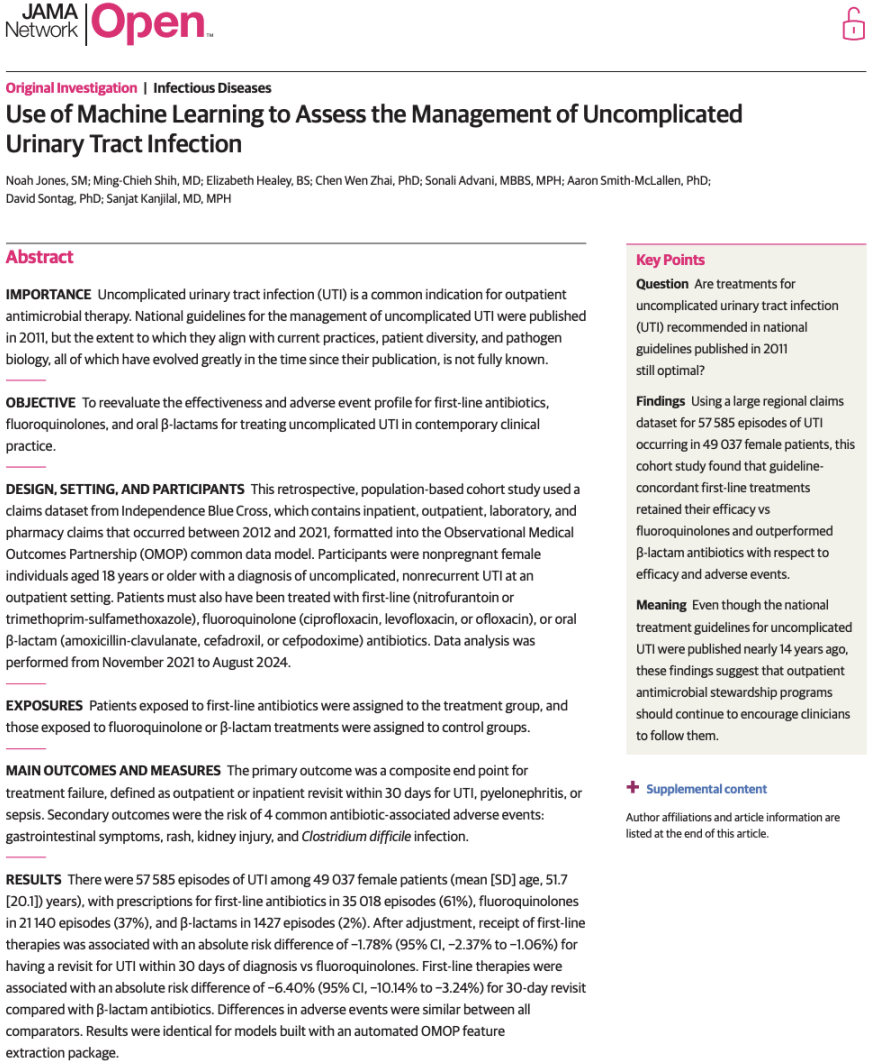
Keywords:
Age-related eye disease
Cataract
glaucoma
Age-related macular degeneration
Mental disorders
Older adults

ABSTRACT

Background: The relationship between age-related eye diseases and the subsequent risk of dementia and depressive disorders remains inconsistent. Furthermore, the effects on anxiety disorders and sleep disorders have been underexplored. This study aims to comprehensively examine the impact of age-related eye diseases on common mental disorders in older adults, thereby enhancing our understanding of the mental health implications in these conditions.
Methods: The electronic health records of 1,522,036 patients aged over 60 from ten institutions in South Korea were analyzed. Patients with and without age-related eye diseases were identified, and 1:4 propensity score matching (PSM) was implemented. A 10-year longitudinal analysis was conducted using the Cox proportional hazards model to calculate the hazard ratios (HR). A meta-analysis was performed to combine the results from different institutions. Subgroup analyses were conducted to explore the impact of specific age-related eye diseases (cataract, glaucoma, age-related macular degeneration) on mental disorders.
Results: A total of 41,637 patients with age-related eye disease were matched with 134,908 patients without such conditions. Patients with age-related eye disease showed a significantly higher risk of mental disorders (dementia, HR: 1.21 [95 % CI: 1.14–1.27]; depressive disorders, HR: 1.28 [95 % CI: 1.20–1.36]; anxiety disorders, HR: 1.31 [95 % CI: 1.22–1.41]; sleep disorders, HR: 1.29 [95 % CI: 1.22–1.37]). In subgroup analyses, each of the three age-related eye diseases was significantly associated with an increased risk of mental disorders. (cataract, HR: 1.25–1.33; glaucoma, HR: 1.15–1.49; age-related macular degeneration, HR: 1.18–1.37).
Conclusion: Age-related eye diseases increase the risk of developing mental disorders in older adults, highlighting the need for a multidisciplinary approach to patient care in these conditions.

OHDSI Shoutouts!

Congratulations to the team of **Noah Jones, Ming-Chieh Shih, Elizabeth Healey, Chen Wen Zhai, Sonali Advani, Aaron Smith-McLallen, David Sontag, and Sanjat Kanjilal** on the publication of **Use of Machine Learning to Assess the Management of Uncomplicated Urinary Tract Infection** in *JAMA Network Open.*



JAMA Network Open

Original Investigation | Infectious Diseases

Use of Machine Learning to Assess the Management of Uncomplicated Urinary Tract Infection

Noah Jones, SM; Ming-Chieh Shih, MD; Elizabeth Healey, BS; Chen Wen Zhai, PhD; Sonali Advani, MBBS, MPH; Aaron Smith-McLallen, PhD; David Sontag, PhD; Sanjat Kanjilal, MD, MPH

Abstract

IMPORTANCE Uncomplicated urinary tract infection (UTI) is a common indication for outpatient antimicrobial therapy. National guidelines for the management of uncomplicated UTI were published in 2011, but the extent to which they align with current practices, patient diversity, and pathogen biology, all of which have evolved greatly in the time since their publication, is not fully known.

OBJECTIVE To reevaluate the effectiveness and adverse event profile for first-line antibiotics, fluoroquinolones, and oral β-lactams for treating uncomplicated UTI in contemporary clinical practice.

DESIGN, SETTING, AND PARTICIPANTS This retrospective, population-based cohort study used a claims dataset from Independence Blue Cross, which contains inpatient, outpatient, laboratory, and pharmacy claims that occurred between 2012 and 2021, formatted into the Observational Medical Outcomes Partnership (OMOP) common data model. Participants were nonpregnant female individuals aged 18 years or older with a diagnosis of uncomplicated, nonrecurrent UTI at an outpatient setting. Patients must also have been treated with first-line (nitrofurantoin or trimethoprim-sulfamethoxazole), fluoroquinolone (ciprofloxacin, levofloxacin, or ofloxacin), or oral β-lactam (amoxicillin-clavulanate, cefadroxil, or cefpodoxime) antibiotics. Data analysis was performed from November 2021 to August 2024.

EXPOSURES Patients exposed to first-line antibiotics were assigned to the treatment group, and those exposed to fluoroquinolone or β-lactam treatments were assigned to control groups.

MAIN OUTCOMES AND MEASURES The primary outcome was a composite end point for treatment failure, defined as outpatient or inpatient revisit within 30 days for UTI, pyelonephritis, or sepsis. Secondary outcomes were the risk of 4 common antibiotic-associated adverse events: gastrointestinal symptoms, rash, kidney injury, and *Clostridium difficile* infection.

RESULTS There were 57 585 episodes of UTI among 49 037 female patients (mean [SD] age, 51.7 [20.1]) years), with prescriptions for first-line antibiotics in 35 018 episodes (61%), fluoroquinolones in 21 140 episodes (37%), and β-lactams in 1427 episodes (2%). After adjustment, receipt of first-line therapies was associated with an absolute risk difference of −1.78% (95% CI, −2.37% to −1.06%) for having a revisit for UTI within 30 days of diagnosis vs fluoroquinolones. First-line therapies were associated with an absolute risk difference of −6.40% (95% CI, −10.14% to −3.24%) for 30-day revisit compared with β-lactam antibiotics. Differences in adverse events were similar between all comparators. Results were identical for models built with an automated OMOP feature extraction package.

Key Points

Question Are treatments for uncomplicated urinary tract infection (UTI) recommended in national guidelines published in 2011 still optimal?

Findings Using a large regional claims dataset for 57 585 episodes of UTI occurring in 49 037 female patients, this cohort study found that guideline-concordant first-line treatments retained their efficacy vs fluoroquinolones and outperformed β-lactam antibiotics with respect to efficacy and adverse events.

Meaning Even though the national treatment guidelines for uncomplicated UTI were published nearly 14 years ago, these findings suggest that outpatient antimicrobial stewardship programs should continue to encourage clinicians to follow them.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

OHDSI Shoutouts!

Congratulations to the team of **Seok Jun Park, Seungwon Yang, Suhyun Lee, Sung Hwan Joo, Taemin Park, Dong Hyun Kim, Hyeonji Kim, Soyun Park, Jung-Tae Kim, Won Gun Kwack, Sung Wook Kang, Yun-Kyoung Song, Jae Myung Cha, Sang Youl Rhee, and Eun Kyoung Chung** on the publication of **Machine-Learning Parsimonious Prediction Model for Diagnostic Screening of Severe Hematological Adverse Events in Cancer Patients Treated with PD-1/PD-L1 Inhibitors: Retrospective Observational Study by Using the Common Data Model** in *Diagnostics.*

diagnostics

MDPI

Article

Machine-Learning Parsimonious Prediction Model for Diagnostic Screening of Severe Hematological Adverse Events in Cancer Patients Treated with PD-1/PD-L1 Inhibitors: Retrospective Observational Study by Using the Common Data Model

Seok Jun Park [1,2,†], Seungwon Yang [1,2,3,†], Suhyun Lee [3,†], Sung Hwan Joo [1,2], Taemin Park [1,2,3], Dong Hyun Kim [1,2,3], Hyeonji Kim [1,2,3], Soyun Park [1,2,3], Jung-Tae Kim [4], Won Gun Kwack [5], Sung Wook Kang [6], Yun-Kyoung Song [7], Jae Myung Cha [8,*], Sang Youl Rhee [9,10,*] and Eun Kyoung Chung [1,2,3,4,*]

1 Department of Regulatory Science, College of Pharmacy, Graduate School, Kyung Hee University, Seoul 02447, Republic of Korea; psjqkr0824@naver.com (S.J.P.); syang345@khu.ac.kr (S.Y.); shsonic95@khu.ac.kr (S.H.J.); ptyoon1@gmail.com (T.P.); waterlion3@khu.ac.kr (D.H.K.); u00u96@khu.ac.kr (H.K.); 1223psy@khu.ac.kr (S.P.)
2 Institute of Regulatory Innovation Through Science (IRIS), Kyung Hee University, Seoul 02447, Republic of Korea
3 Department of Pharmacy, College of Pharmacy, Kyung Hee University, Seoul 02447, Republic of Korea; sh198410@khu.ac.kr
4 Department of Pharmacy, Kyung Hee University Hospital at Gangdong, Seoul 05278, Republic of Korea; jtkim@khnmc.or.kr
5 Division of Pulmonary, Allergy and Critical Care Medicine, Kyung Hee University Hospital, Seoul 02447, Republic of Korea; wongunnim@naver.com
6 Department of Pulmonary and Critical Care Medicine, Kyung Hee University Hospital at Gangdong, Seoul 05278, Republic of Korea; aikra@khnmc.or.kr
7 College of Pharmacy, The Catholic University of Korea-Sungsim Campus, Bucheon 14662, Gyeonggi-do, Republic of Korea; yksong@catholic.ac.kr
8 Division of Gastroenterology, Department of Internal Medicine, Kyung Hee University Hospital at Gangdong, Kyung Hee University School of Medicine, Seoul 05278, Republic of Korea
9 Center for Digital Health, Medical Science Research Institute, College of Medicine, Kyung Hee University, Seoul 02447, Republic of Korea
10 Department of Endocrinology and Metabolism, Kyung Hee University School of Medicine, Seoul 02447, Republic of Korea
* Correspondence: drcha@khu.ac.kr (J.M.C.); rheesy@khu.ac.kr (S.Y.R.); cekchung@khu.ac.kr (E.K.C.)
† These authors contributed equally to this work.

check for updates

Received: 4 October 2024
Revised: 8 January 2025
Accepted: 17 January 2025
Published: 20 January 2025

Citation: Park, S.J.; Yang, S.; Lee, S.; Joo, S.H.; Park, T.; Kim, D.H.; Kim, H.; Park, S.; Kim, J.-T.; Kwack, W.G.; et al. Machine-Learning Parsimonious Prediction Model for Diagnostic Screening of Severe Hematological Adverse Events in Cancer Patients

Abstract: Background/Objectives: Earlier detection of severe immune-related hematological adverse events (irHAEs) in cancer patients treated with a PD-1 or PD-L1 inhibitor is critical to improving treatment outcomes. The study aimed to develop a simple machine learning (ML) model for predicting irHAEs associated with PD-1/PD-L1 inhibitors.

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	Atlas
Wednesday	8 am	Psychiatry
Wednesday	7 pm	Medical Imaging
Thursday	10 am	Themis
Thursday	11 am	Industry
Thursday	12 pm	Methods Research
Thursday	1 pm	OMOP CDM Oncology Vocabulary/Development Subgroup
Thursday	7 pm	Dentistry
Friday	10 am	GIS - Geographic Information System
Friday	10 am	Transplant
Friday	11:30 am	Steering
Monday	9 am	Vaccine Vocabulary
Monday	10 am	Healthcare Systems Interest Group
Tuesday	9 am	OMOP CDM Oncology Genomic Subgroup

Global Symposium: Oct. 7-9

The 2025 OHDSI Global Symposium will return to the Hyatt Regency Hotel in New Brunswick, N.J., on Oct. 7-9, 2025.

More details will be shared when available.

Save The Dates!

February Newsletter

OHDSI

The Journey Newsletter (February 2025)

OHDSI opened 2025 by looking at four potential community focuses, including a look at guideline-driven evidence generation opportunities. Collaborators around the world identified numerous examples where OHDSI could inform key healthcare questions. We invite you to learn about them and share what studies you might be interested in joining. This newsletter also includes community updates, recent studies, a new collaborator spotlight and more!
#JoinTheJourney

Podcast: Clinical Guideline Opportunities

Clinical guidelines can identify key evidence gaps in healthcare. In the latest On The Journey podcast, Patrick Ryan and Craig Sachson discuss why these guidelines are an opportunity for OHDSI to provide impact, the amazing global community response, and what are the next steps for these potential studies during Phenotype Phebruary. *(if video does not appear, please click view this email in your browser)*

Community Updates

Where Have We Been?

- Fourteen OHDSI collaborators provided presentations about guideline-driven evidence opportunities for community network studies. You can see those talks here, and share what studies you would like to join.
- The Bridge Network Training Program was introduced by Marc Twagirumukiza, a member of the OHDSI Africa Chapter. This five-year program is centered on infectious disease data and health informatics in Africa, and it seeks OHDSI input in a variety of ways. Please check out the video or slides, and share with anybody who may be interested in joining this initiative.
- OHDSI kicked off its 2025 community calls with a session focused on where we can go together over the next 12 months. Patrick Ryan highlighted four focus areas for the community: guideline-driven evidence generation, evidence-driven data standardization, evidence-driven open-source development, and evidence-driven collaborative education. There were details about monthly events, and upcoming clinical/scientific conferences over the next 18 months that can be end goals for dissemination.

Where Are We Now?

- OHDSI has more than 30 workgroups, which present opportunities for all community members to find a home for their talents and passions, and make meaningful contributions. Throughout February community calls, workgroup representatives will discuss their mission, objectives & key results, and share how they can impact OHDSI's global focuses in 2025.
- Phenotype Phebruary has begun, and this year it will focus on building the phenotypes needed for the guideline-driven studies introduced last month. Patrick Ryan provides more details about Phenotype Phebruary in the latest On The Journey podcast.
- The #OHDSISocialShowcase features posters, software demos and lightning talks from the 2024 Global Symposiums this month. Please make sure you are following OHDSI's LinkedIn, Twitter/X and Instagram feeds to receive daily updates on the research presented by our community.

Community Presents Guideline-Driven Evidence Opportunities; Learn More & Share Your Potential Interest In Collaboration

Clinical guidelines not only offer treatment recommendations for healthcare providers but also highlight evidence gaps that could shape critical questions for both clinicians and patients. The OHDSI community aimed to identify these gaps and explore how they could be addressed through network studies across the OHDSI Evidence Network.

Throughout January, collaborators around the world highlighted such gaps in a forum thread and joined a community call to provide a brief description of the gap ad why OHDSI is positioned to generate reliable and informative real-world evidence. Please check out the videos below or read about these evidence opportunities, and then fill out the brief form below to share your interest in joining one or multiple studies.

Join a Guideline-Driven Evidence Network Study

Watch The Presentations

Forum Thread on Guideline Opportunities

January Publications

Basile AO, Verma A, Tang LA, Serper M, Scanga A, Farrell A, Destin B, Carr RM, Anyanwu-Ofili A, Rajagopal G, Krikhely A, Bessler M, Reilly MP, Ritchie MD, Tatonetti NP, Wattacheril J. Rapid identification and phenotyping of nonalcoholic fatty liver disease patients using a machine-based approach in diverse healthcare systems. Clin Transl Sci. 2025 Jan;18(1):e70105. doi: 10.1111/cts.70105. PMID: 39739635; PMCID: PMC11686338.

de Groot R, Glaser S, Kogan A, Medlock S, Alloni A, Gabetta M, Wilk S, de Keizer N, Cornet R. ATC-to-RxNorm mappings - A comparison between OHDSI Standardized Vocabularies and UMLS Metathesaurus. Int J Med Inform. 2024 Dec 31;195:105777. doi: 10.1016/j.ijmedinf.2024.105777. Epub ahead of print. PMID: 39753061.

Finster M, Moinat M, Taghizadeh E. ETL: From the German Health Data Lab data formats to the OMOP Common Data Model. PLoS One. 2025 Jan 6;20(1):e0311511. doi: 10.1371/journal.pone.0311511. PMID: 39761272; PMCID: PMC11703056.

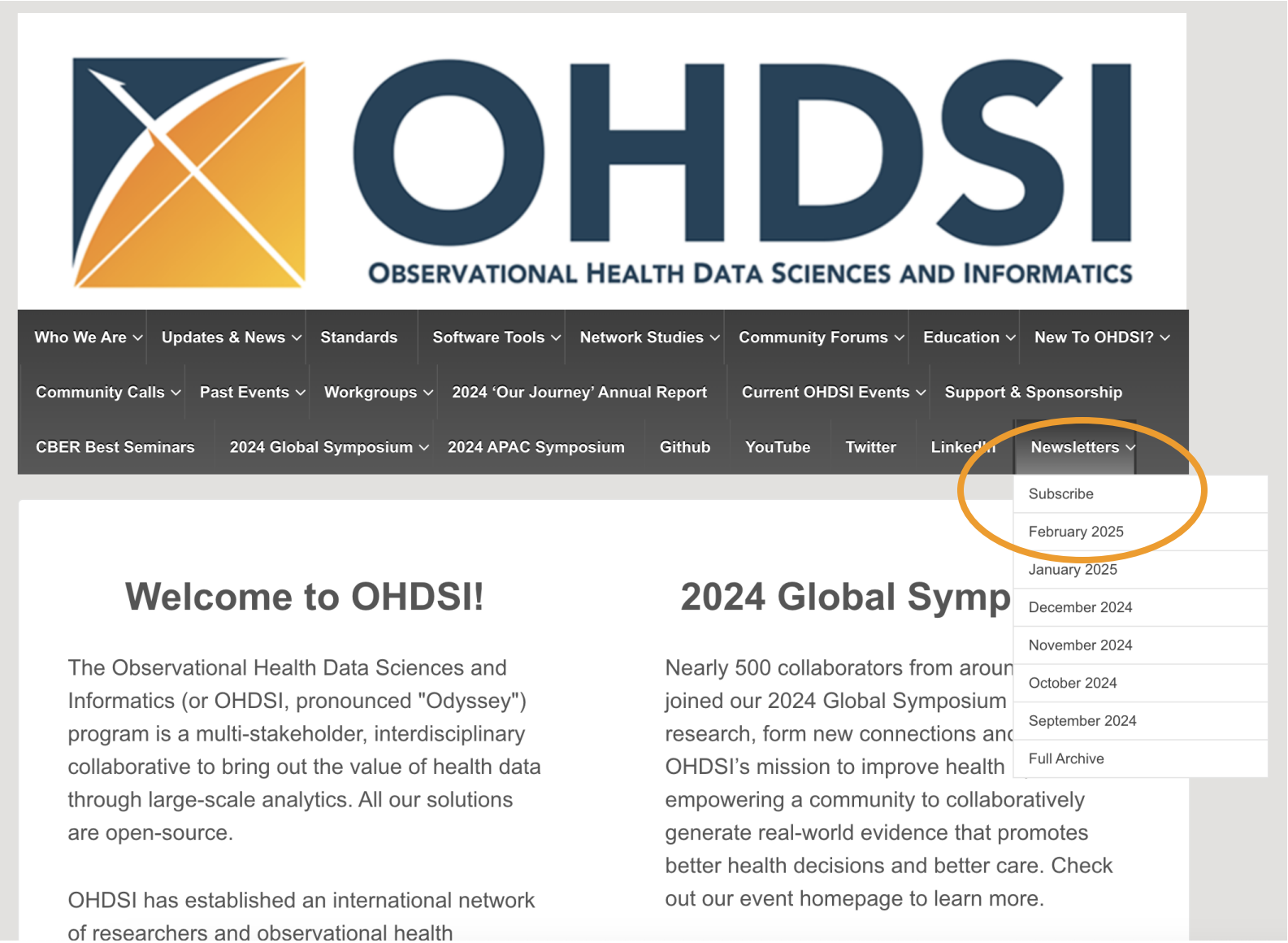
Hong N, Ko YH, Park JH, Ha EJ, Lee SH, Kim KM, Kang HS, Kim JE, Kim K, Cho WS. A common data model for oral anticoagulants-related risk of spontaneous intracranial hemorrhage. J Clin Neurosci. 2025 Jan 8;133:111039. doi: 10.1016/j.jocn.2025.111039. Epub ahead of print. PMID: 39787902.

Schuemie MJ, Ostropolets A, Zhuk A, Korsik U, Seo SI, Suchard MA, Hripcsak G, Ryan PB. Standardized patient profile review using large language models for case adjudication in observational research. NPJ Digit Med. 2025 Jan 9;8(1):18. doi: 10.1038/s41746-025-01433-4. PMID: 39789235; PMCID: PMC11718233.

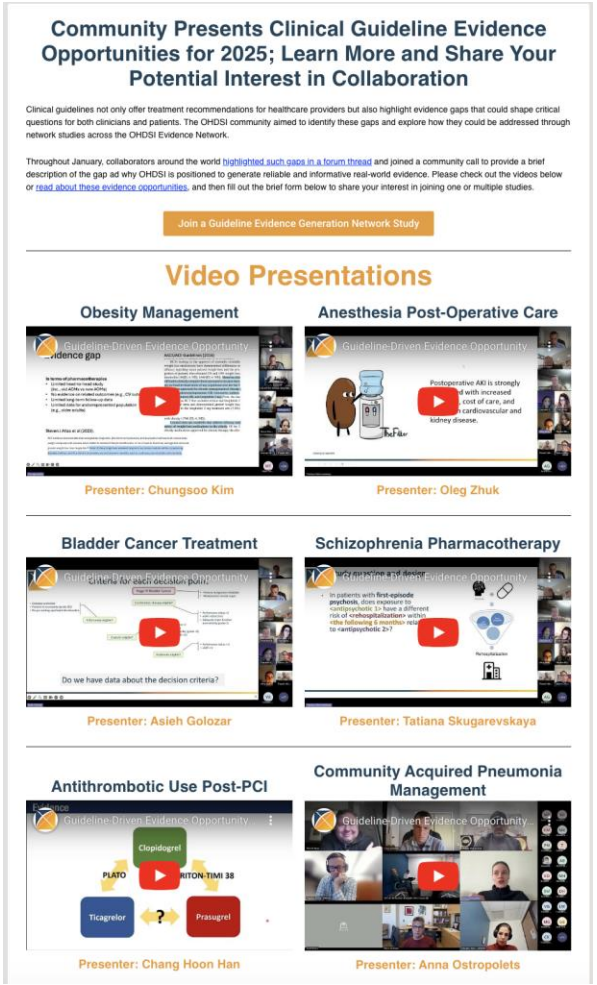
Kwon YE, Ahn SY, Ko GJ, Kwon YJ, Kim JE. Impact of Uric Acid Levels on Mortality and Cardiovascular Outcomes in Relation to Kidney Function. J Clin Med. 2024 Dec 24;14(1):20. doi: 10.3390/jcm14010020. PMID: 39797103; PMCID: PMC11721403.

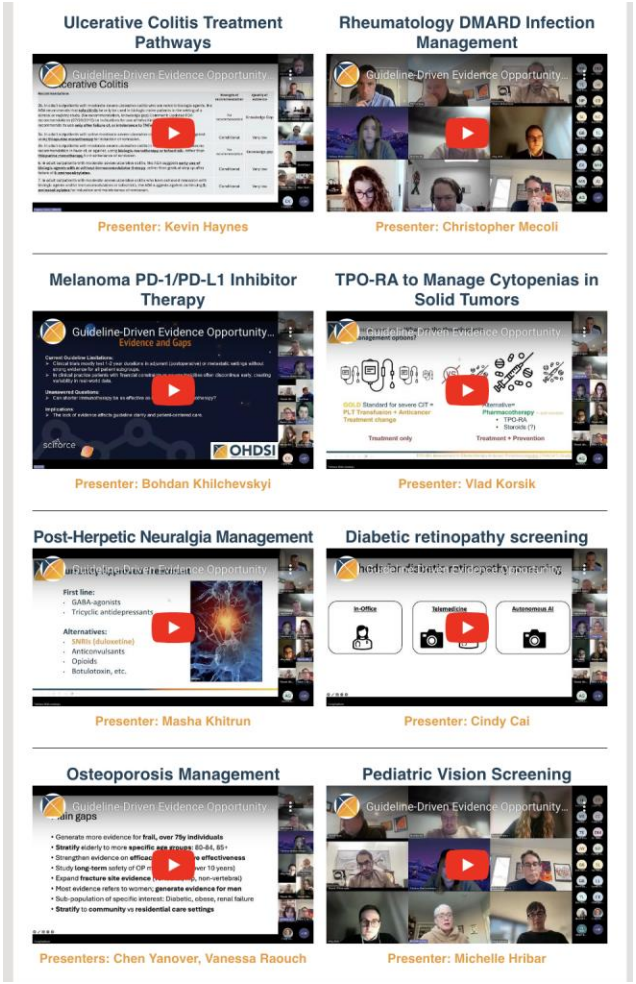
Conover MM, Albogami Y, Hardin J, Reich CG, Ostropolets A, Ryan PB; Observational Health Data Sciences and Informatics (OHDSI) Research Network. Glucagon-Like Peptide 1 Receptor Agonists and Chronic Lower Respiratory Disease Among Type 2 Diabetes Patients: Replication and Reliability Assessment Across a Research Network. Pharmacoepidemiol Drug Saf. 2025 Jan;34(1):e70087. doi: 10.1002/pds.70087. PMID: 39805811; PMCID: PMC11730806.

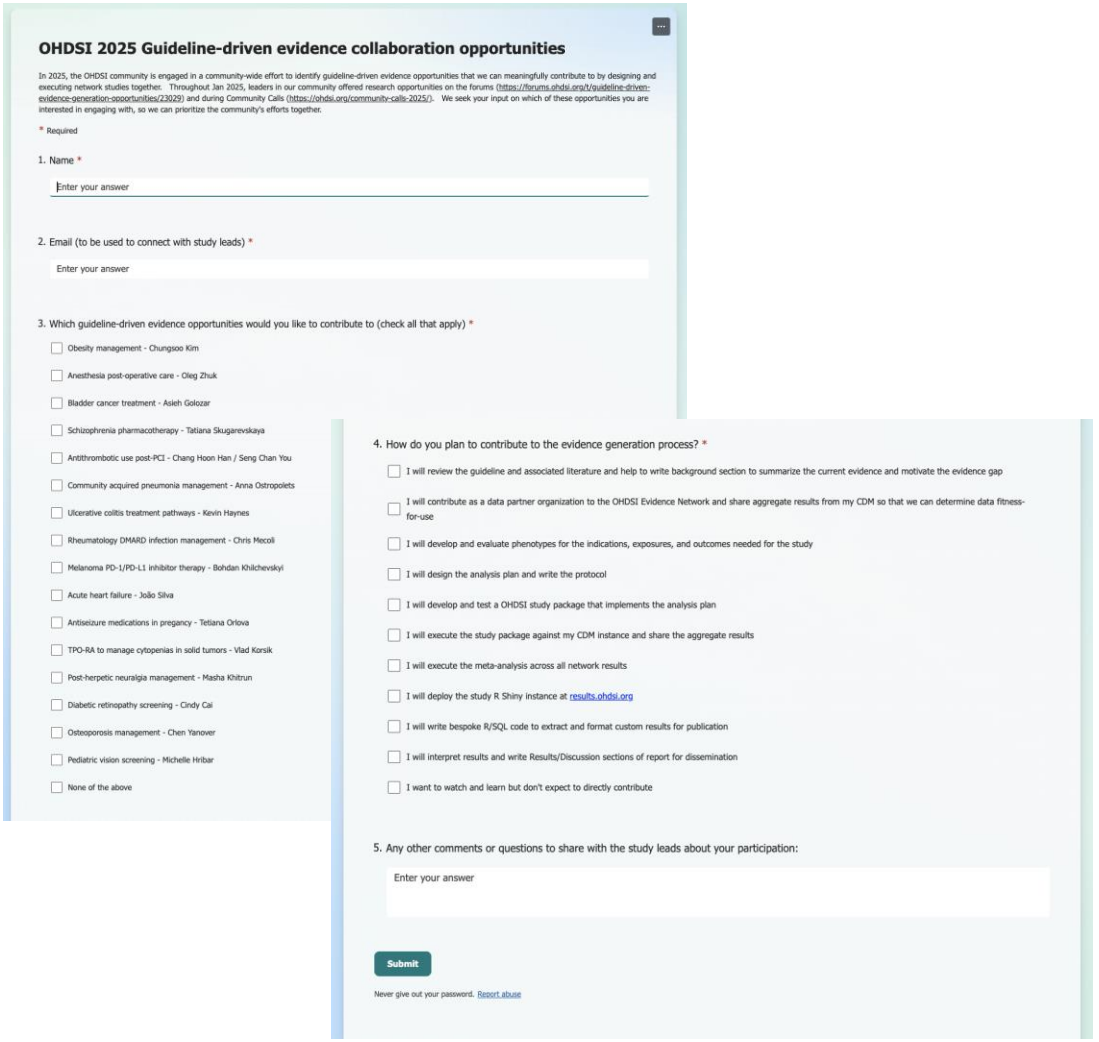
February Newsletter



Guideline-Driven Evidence Opportunities







ohdsi.org/clinical-guideline-evidence-opportunities-2025

Collaborator Spotlight: Cynthia Sung

Dr. Cynthia Sung is an Adjunct Associate Professor for the Centre of Regulatory Excellence at Duke-National University of Singapore Medical School. She earned the 2023 Titan Award for Community Collaboration.

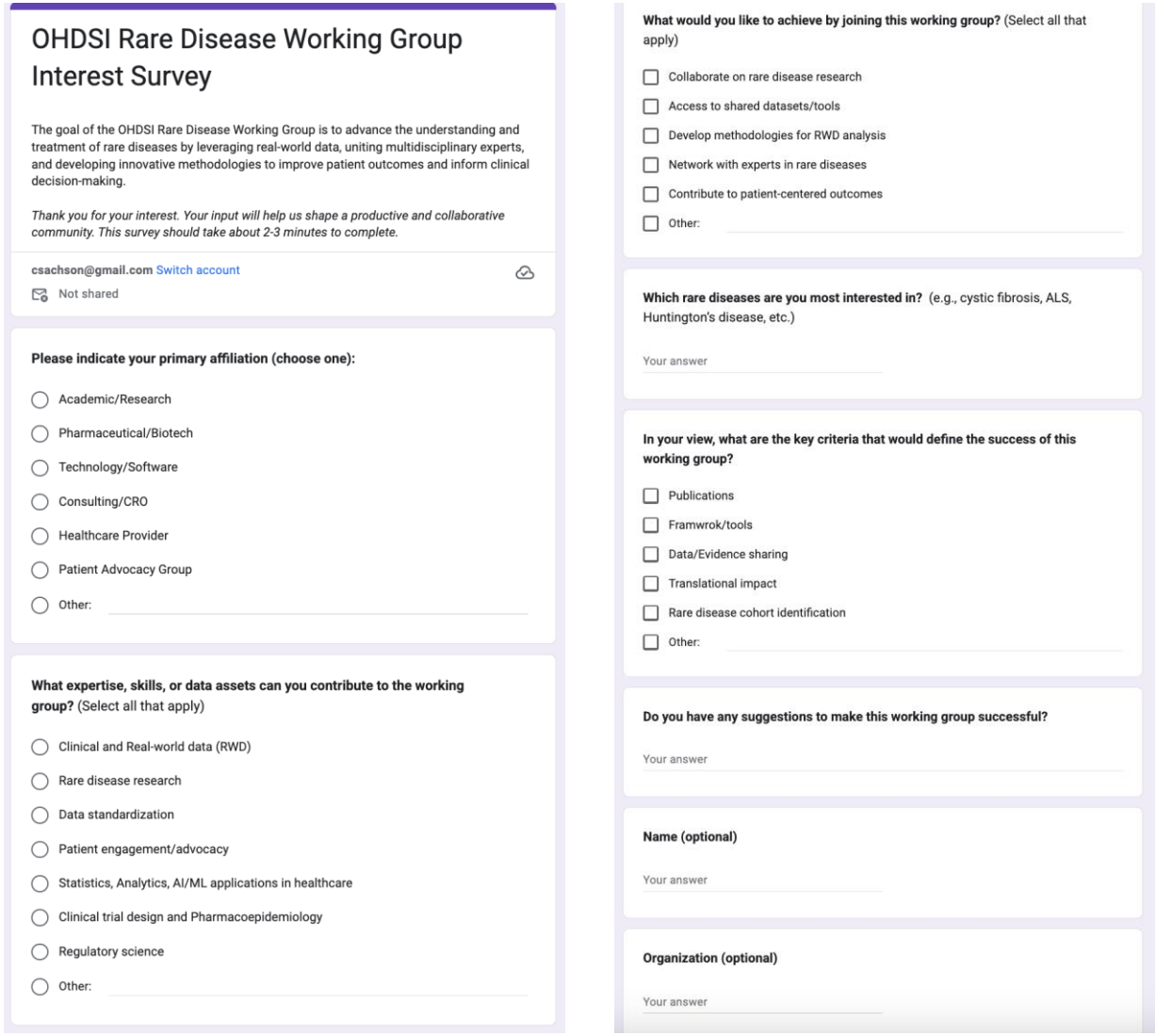
In the latest collaborator spotlight, Cynthia discusses a career journey that has taken her around the world, the need for FAIR data in less-represented populations, exciting developments within Africa, and more.

ohdsi.org/spotlight-Cynthia-Sung

Rare Disease WG Survey

The goal of the OHDSI Rare Disease Working Group is to advance the understanding and treatment of rare diseases by leveraging real-world data, uniting multidisciplinary experts, and developing innovative methodologies to improve patient outcomes and inform clinical decision-making.

The workgroup has posted a brief interest survey to help shape a productive and collaborative community. Please fill out this survey by Tuesday, Feb. 18.



OHDSI Rare Disease Working Group Interest Survey

The goal of the OHDSI Rare Disease Working Group is to advance the understanding and treatment of rare diseases by leveraging real-world data, uniting multidisciplinary experts, and developing innovative methodologies to improve patient outcomes and inform clinical decision-making.

Thank you for your interest. Your input will help us shape a productive and collaborative community. This survey should take about 2-3 minutes to complete.

csachson@gmail.com Switch account

Not shared

Please indicate your primary affiliation (choose one):

- Academic/Research
- Pharmaceutical/Biotech
- Technology/Software
- Consulting/CRO
- Healthcare Provider
- Patient Advocacy Group
- Other:

What expertise, skills, or data assets can you contribute to the working group? (Select all that apply)

- Clinical and Real-world data (RWD)
- Rare disease research
- Data standardization
- Patient engagement/advocacy
- Statistics, Analytics, AI/ML applications in healthcare
- Clinical trial design and Pharmacoepidemiology
- Regulatory science
- Other:

What would you like to achieve by joining this working group? (Select all that apply)

- Collaborate on rare disease research
- Access to shared datasets/tools
- Develop methodologies for RWD analysis
- Network with experts in rare diseases
- Contribute to patient-centered outcomes
- Other:

Which rare diseases are you most interested in? (e.g., cystic fibrosis, ALS, Huntington's disease, etc.)

Your answer

In your view, what are the key criteria that would define the success of this working group?

- Publications
- Framwrok/tools
- Data/Evidence sharing
- Translational impact
- Rare disease cohort identification
- Other:

Do you have any suggestions to make this working group successful?

Your answer

Name (optional)

Your answer

Organization (optional)

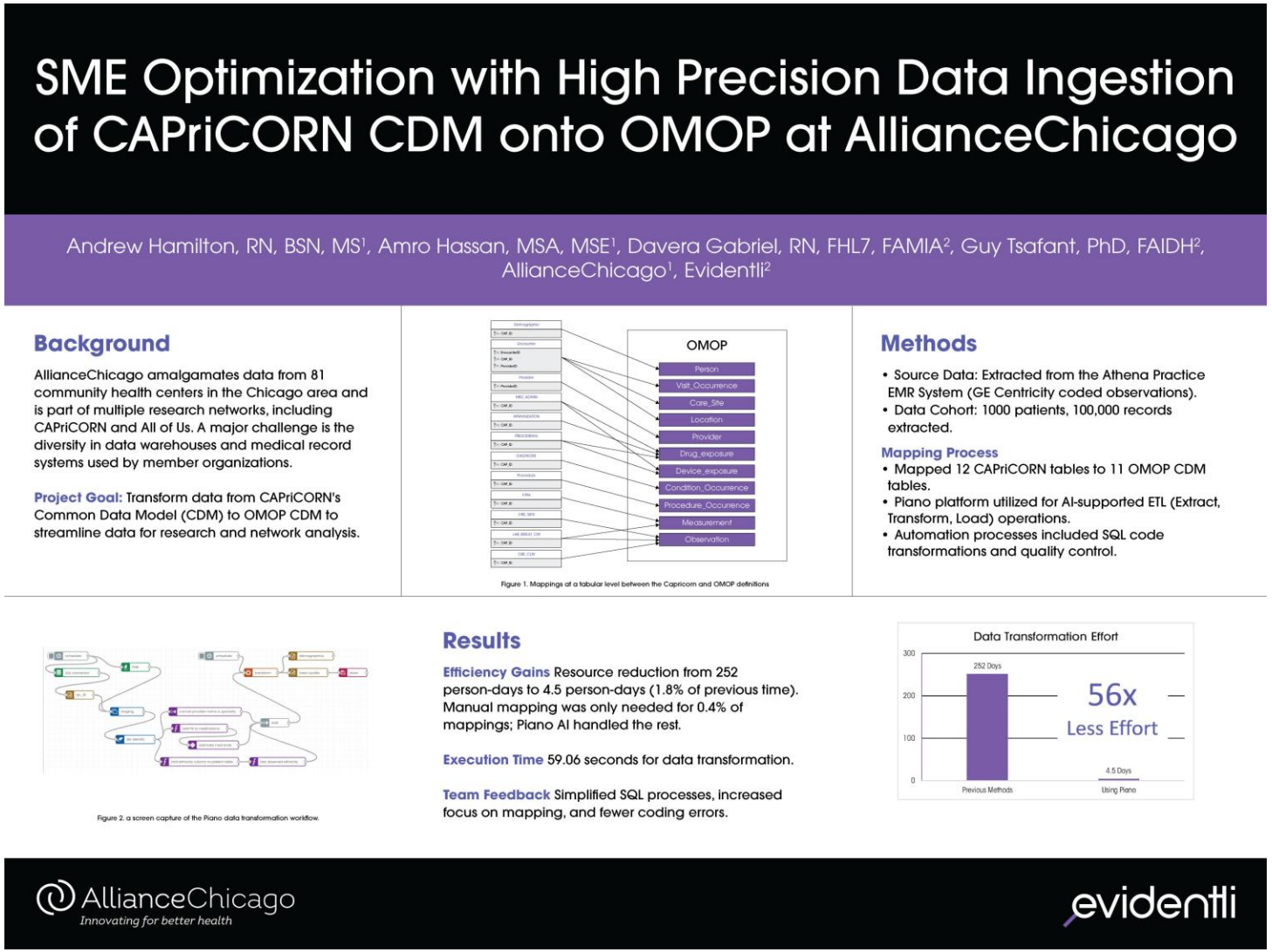
Your answer

#OHDSISocialShowcase This Week

Monday

SMEs optimization with high precision data ingestion of CAPriCORN CDM onto OMOP at AllianceChicago

(**Andrew Hamilton**, Amro Hassan, Davera Gabriel, Guy Tsafnat)



SME Optimization with High Precision Data Ingestion of CAPriCORN CDM onto OMOP at AllianceChicago

Andrew Hamilton, RN, BSN, MS[1], Amro Hassan, MSA, MSE[1], Davera Gabriel, RN, FHL7, FAMIA[2], Guy Tsafant, PhD, FAIDH[2], AllianceChicago[1], Evidentli[2]

Background

AllianceChicago amalgamates data from 81 community health centers in the Chicago area and is part of multiple research networks, including CAPriCORN and All of Us. A major challenge is the diversity in data warehouses and medical record systems used by member organizations.

Project Goal: Transform data from CAPriCORN's Common Data Model (CDM) to OMOP CDM to streamline data for research and network analysis.

OMOP

Person
Visit_Occurrence
Care_Site
Location
Provider
Drug_exposure
Device_exposure
Condition_Occurrence
Procedure_Occurrence
Measurement
Observation

Figure 1. Mappings at a tabular level between the Capricorn and OMOP definitions

Methods

- Source Data: Extracted from the Athena Practice EMR System (GE Centricity coded observations).
- Data Cohort: 1000 patients, 100,000 records extracted.

Mapping Process

- Mapped 12 CAPriCORN tables to 11 OMOP CDM tables.
- Piano platform utilized for AI-supported ETL (Extract, Transform, Load) operations.
- Automation processes included SQL code transformations and quality control.

Figure 2. a screen capture of the Piano data transformation workflow.

Results

Efficiency Gains Resource reduction from 252 person-days to 4.5 person-days (1.8% of previous time). Manual mapping was only needed for 0.4% of mappings; Piano AI handled the rest.

Execution Time 59.06 seconds for data transformation.

Team Feedback Simplified SQL processes, increased focus on mapping, and fewer coding errors.

Data Transformation Effort

300
200
100
0
252 Days
56x
Less Effort
4.5 Days
Previous Methods
Using Piano

AllianceChicago
Innovating for better health

evidentli

#OHDSISocialShowcase This Week

Tuesday

Automating data standardization through ad hoc SNOMED modeling with LLM: proof of concept

(Eduard Korchmar, Vojtech Huser, Christian Reich, Alexander Davydov)

Automating data standardization through ad hoc SNOMED modeling with LLM: proof of concept

Eduard Korchmar[1], Vojtech Huser, MD PhD[1], Christian Reich, MD PhD[2], Alexander Davydov, MD[1]
[1]Odysseus, an EPAM Company, Cambridge, MA; [2]Northeastern University, Boston, MA

BACKGROUND

Standardizing source medical data and external ontologies to the OMOP Common Data Model (CDM) is a prerequisite to enable the powerful observational research toolset of the OHDSI ecosystem. The emerging Large Language Models (LLM) as the latest iteration of applied Machine Learning (ML) technology have created the opportunity to revise current approaches to standardization. Currently, the only widely practiced automation approach is utilizing the TF-IDF algorithm. [1] This approach has limited accuracy in edge cases where nuances of the term meaning depend on context. [2] At the same time, LLMs that are put to the task of medical data standardization often demonstrate accuracy that is on par, if not worse, compared to TF-IDF. [3] The reasons for this are: sparse specialized training datasets required for such tasks, and limited information held in each record that does not allow LLM to benefit from signature large context windows. [4] In addition, LLMs, often fail at tasks that require formal logical approach, [4]–[6] which semantic capture ultimately is. [2] This puts a limit on what it can achieve even with specialized pre-training. [4], [6]

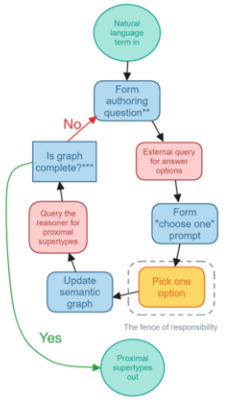


Figure 1: Generalized principle of work. The LLM is only ever prompted for data retrieval

```
# Source concept: Pyogenic abscess of liver
<<< 64572001 |Disease (disorder)| :
      116676008 |Associated morphology| = 418453007 |Pyogenic abscess (morphologic abnormality)| ,
      363698007 |Finding site| = 10200004 |Liver structure (body structure)| ,
      47429007 |Associated with (attribute)| = 103424003 |Pyogenic bacterium (organism)| ,
      246075003 |Causative agent (attribute)| = 409822003 |Domain Bacteria (organism)| ,
      370135005 |Pathological process (attribute)| = 441862004 |Infectious process (qualifier value)|
}
```

Figure 2: Post-coordinated expression (PCE) for 'Pyogenic abscess of liver' generated by our algorithm.

We believe by incorporating a rigid semantic data model, such as one provided by SNOMED Clinical Terms' (SNOMED CT) description logic, we can define a strict rule set for standardization, [7], [8] provide an ML agent with additional context and reduce the risk of hallucinations. [9] We created a methodology that uses a constrained LLM for the task of semantic mapping.

METHODS

ALGORITHM OVERVIEW

The principle of our approach is to limit responsibilities of an LLM agent by a framework of formally defined rules, and to use it only for knowledge retrieval. We utilize SNOMED CT Machine Readable Concept Model (MRCM) to iteratively populate a semantic graph representing the meaning of a medical term. [7], [8], [10] We then convert this graph to a post-coordinated expression (PCE) in SNOMED CT compositional grammar [11] to enable unambiguous placing of an expression into the standard SNOMED hierarchy through a reasoner. [8] In the final step, we derive relationships from the expression in the form of equivalence or subsumption relationships.

To start the process of building a graph, we query the MRCM for a set of entry points for a given term. They loosely correspond to SNOMED CT hierarchy tags. Hierarchy tags segregate concepts into categories like "Procedure", "Clinical Finding", "Body Structure", etc. [10] The set of tags is then presented to an LLM agent with a strict instruction to pick one closest matching tag for the source concept. After that, MRCM is queried for a set of required or optional attributes associated with this chosen tag. The LLM agent is presented with this list of attributes and must pick one. The process is iteratively repeated for subtags, attributes, attribute values, and primitive descendants (i.e., semantically unrepresentable by existing attribute-value model). Figure 1 shows a general overview of the process as a recursive interaction between a rule-based agent (our original algorithm) and a knowledge-based agent (LLM). Whenever a new node is added to the semantic graph, the entire graph is re-submitted for evaluation by the SNOMED reasoner to narrow down the hierarchical context, which increases the granularity of attributes and constraints retrieved from the MRCM. In addition, prompts to LLM agent can be extended with additional context via the Retrieval-Augmented Generation (RAG) mechanism, e.g. by pulling relevant free-text SNOMED authoring documentation. [12], [13] This process is repeated until all possible proximal primitive parents and attribute-value templates provided by MRCM are either filled out or rejected by the LLM agent. Once the rule-based agent determines options for graph extension to be exhausted, the graph is submitted for the final evaluation step in the form of a PCE (see Figure 2), which will define its place in the SNOMED hierarchy.

PILOT EVALUATION

We used a set of 45 medical terms (categorized by a human expert as simple medical terms from a mapping perspective) to evaluate the methodology. ML-generated mappings were classified as correct or incorrect by human manual evaluation.

RESULTS

We have defined a semantic capture automaton utilizing the MRCM for defining constraints and retrieving concepts, and a LLM for knowledge retrieval and semantic processing. We developed the algorithm and used it to analyze a pilot set of natural language terms and manually validated the output quality. Table 1 shows 3 mapping examples. Our project github repository has the full set and pilot evaluation results.

CONCLUSION

We have demonstrated the feasibility of using algorithmically guided LLMs using predefined rule set on top of a hierarchical knowledge graph for maping medical terms.

Our next steps are to iteratively improve the algorithm based on application to more medical terms. We also hope to evaluate the accuracy using a much larger set of real-world input medical terms. Finally, we want to compare the accuracy of our method against other automated methods. Depending on the benchmarked performance, we see its possible applications in standardization of source natural-language data to SNOMED CT concepts and in medical ontology authoring (for SNOMED CT ontology itself or other ontologies).

Source term	Suggested parents	Reviewer comment
QA48.0 Care or examination immediately after delivery	**236823008** Female genital and obstetric procedures **9632001** Nursing procedure **386637004** Obsteric procedure ...	Correct, but non-specific and contains redundant terms.
1F66.1 Mansonelliasis	**1300177006** Invasive infectious disease **84706005** Infection caused by Nematoda **86820007** Endemic disease ...	Same as above, but with a more specific parent.
MD41 Clinical findings on diagnostic imaging of lung	**129893005** Respiratory alteration **301230006** Lung finding **128973006** Finding related physiologic patient state **72670004** Sign ...	A tricky concept to represent, but the suggestion is correct.

Table 1: Example results for 3 ICD11 terms.

See our GitHub repository for additional results and bibliography:

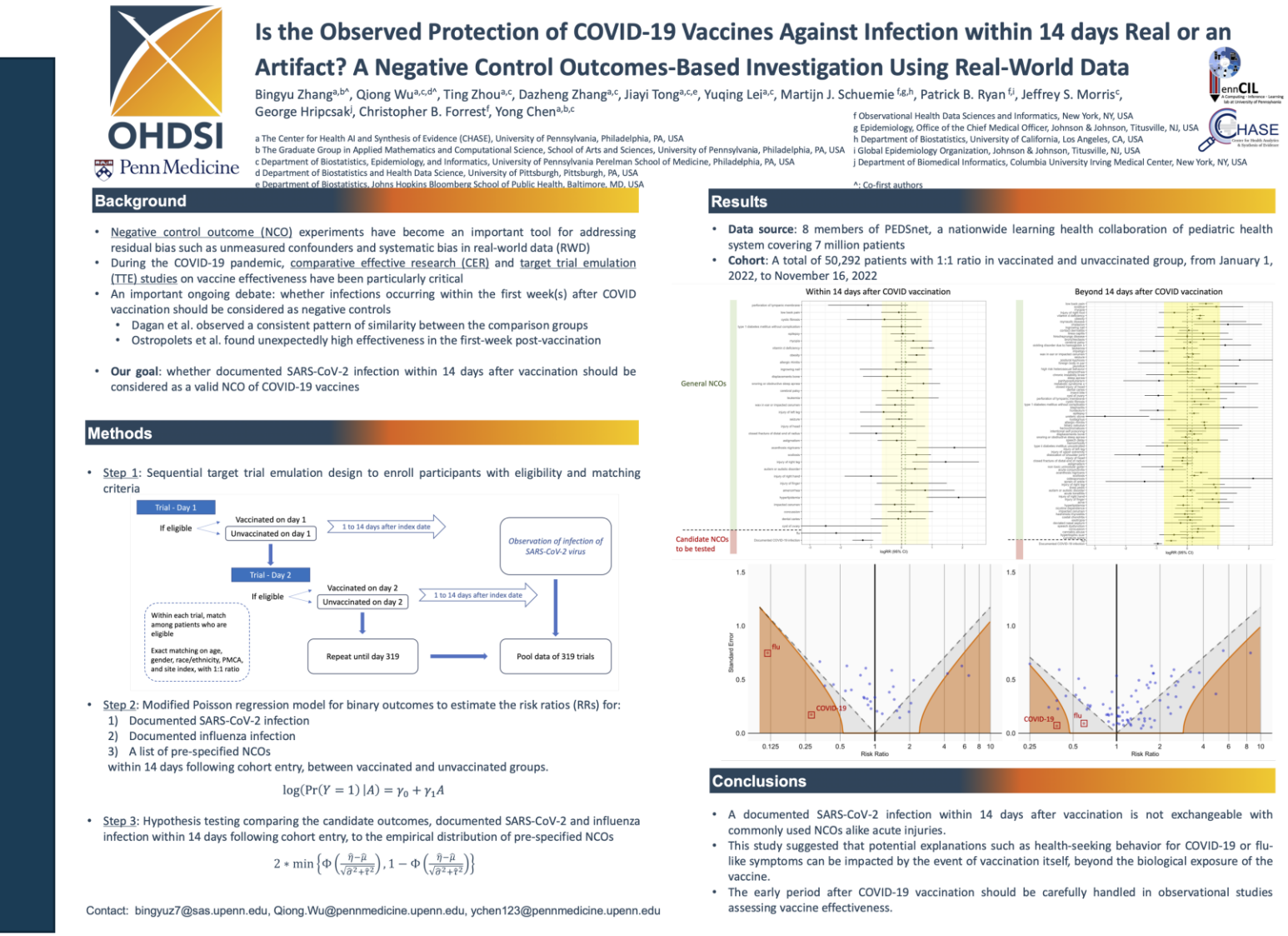
https://github.com/odysseusinc/guided-llm-modeling

#OHDSISocialShowcase This Week

Wednesday

Is the Observed Protection of COVID-19 Vaccines Against Infection within 14 days Real or an Artifact? A Negative Control Outcomes-Based Investigation Using Real-World Data

(Bingyu Zhang, Qiong Wu, Ting Zhou, Dazheng Zhang, Jiayi Tong, Yuqing Lei, Martijn Schuemie, Patrick B. Ryan, Jeffrey S. Morris, George Hripcsak, Christopher B. Forrest, Yong Chen)



Is the Observed Protection of COVID-19 Vaccines Against Infection within 14 days Real or an Artifact? A Negative Control Outcomes-Based Investigation Using Real-World Data

Bingyu Zhang[a,b^], Qiong Wu[a,c,d^], Ting Zhou[a,c], Dazheng Zhang[a,c], Jiayi Tong[a,c,e], Yuqing Lei[a,c], Martijn J. Schuemie[f,g,h], Patrick B. Ryan[f,i], Jeffrey S. Morris[c], George Hripcsak[j], Christopher B. Forrest[f], Yong Chen[a,b,c]

a The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA
b The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA
c Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
d Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, PA, USA
e Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
f Observational Health Data Sciences and Informatics, New York, NY, USA
g Epidemiology, Office of the Chief Medical Officer, Johnson & Johnson, Titusville, NJ, USA
h Department of Biostatistics, University of California, Los Angeles, CA, USA
i Global Epidemiology Organization, Johnson & Johnson, Titusville, NJ, USA
j Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA

^: Co-first authors

Background

- Negative control outcome (NCO) experiments have become an important tool for addressing residual bias such as unmeasured confounders and systematic bias in real-world data (RWD)
- During the COVID-19 pandemic, comparative effective research (CER) and target trial emulation (TTE) studies on vaccine effectiveness have been particularly critical
- An important ongoing debate: whether infections occurring within the first week(s) after COVID vaccination should be considered as negative controls
  - Dagan et al. observed a consistent pattern of similarity between the comparison groups
  - Ostropolets et al. found unexpectedly high effectiveness in the first-week post-vaccination
- **Our goal**: whether documented SARS-CoV-2 infection within 14 days after vaccination should be considered as a valid NCO of COVID-19 vaccines

Methods

- Step 1: Sequential target trial emulation design to enroll participants with eligibility and matching criteria

- Step 2: Modified Poisson regression model for binary outcomes to estimate the risk ratios (RRs) for:
  1) Documented SARS-CoV-2 infection
  2) Documented influenza infection
  3) A list of pre-specified NCOs

  within 14 days following cohort entry, between vaccinated and unvaccinated groups.

$$\log(\Pr(Y=1)\,|A) = \gamma_0 + \gamma_1 A$$

- Step 3: Hypothesis testing comparing the candidate outcomes, documented SARS-CoV-2 and influenza infection within 14 days following cohort entry, to the empirical distribution of pre-specified NCOs

$$2 * \min\left\{\Phi\left(\frac{\hat{\eta}-\hat{\mu}}{\sqrt{\hat{\sigma}^2+\hat{\tau}^2}}\right), 1-\Phi\left(\frac{\hat{\eta}-\hat{\mu}}{\sqrt{\hat{\sigma}^2+\hat{\tau}^2}}\right)\right\}$$

Contact: bingyuz7@sas.upenn.edu, Qiong.Wu@pennmedicine.upenn.edu, ychen123@pennmedicine.upenn.edu

Results

- **Data source**: 8 members of PEDSnet, a nationwide learning health collaboration of pediatric health system covering 7 million patients
- **Cohort**: A total of 50,292 patients with 1:1 ratio in vaccinated and unvaccinated group, from January 1, 2022, to November 16, 2022

Conclusions

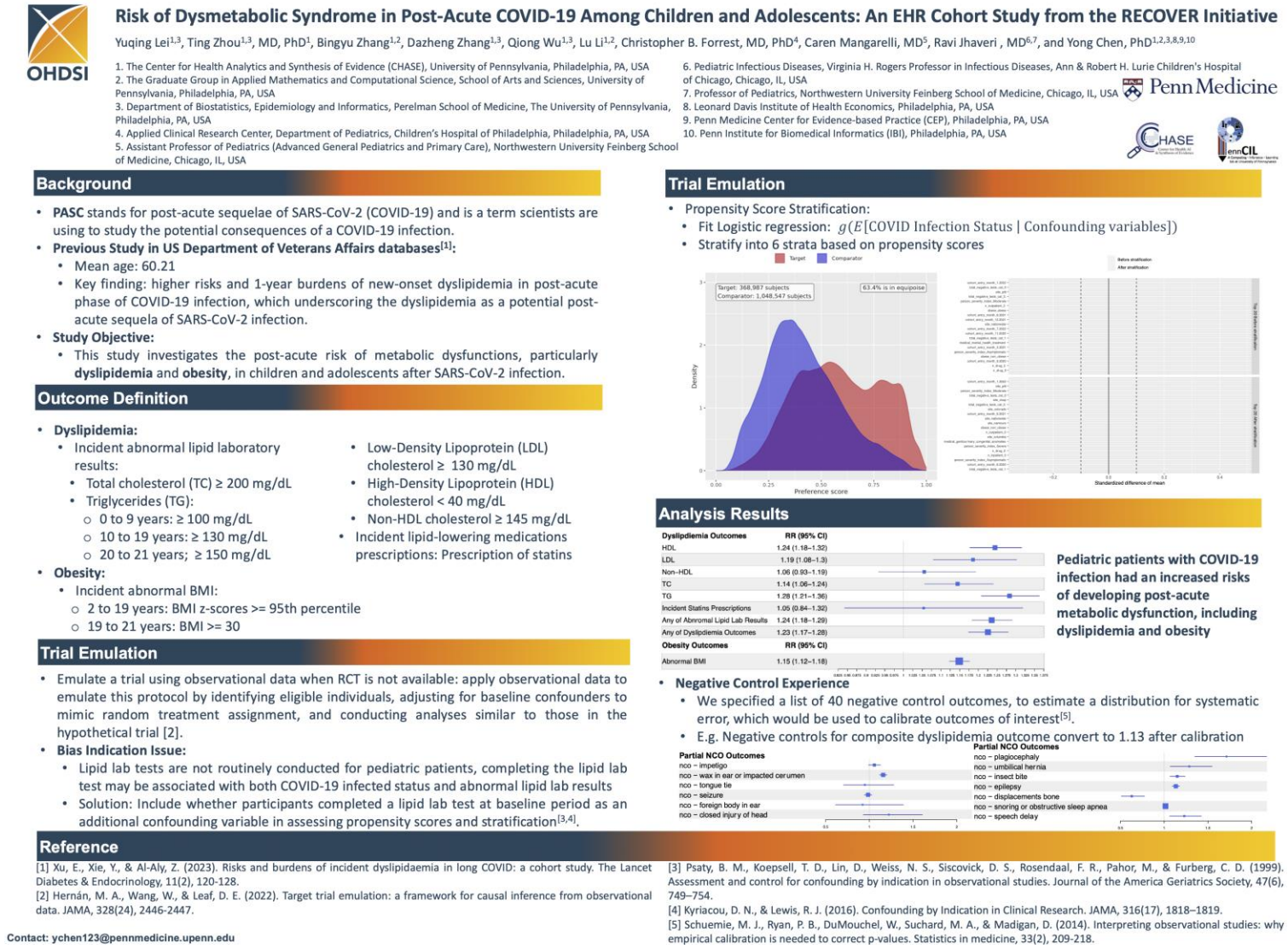
- A documented SARS-CoV-2 infection within 14 days after vaccination is not exchangeable with commonly used NCOs alike acute injuries.
- This study suggested that potential explanations such as health-seeking behavior for COVID-19 or flu-like symptoms can be impacted by the event of vaccination itself, beyond the biological exposure of the vaccine.
- The early period after COVID-19 vaccination should be carefully handled in observational studies assessing vaccine effectiveness.

#OHDSISocialShowcase This Week

Thursday

Risk of Dysmetabolic Syndrome in Post-Acute COVID-19 Among Children and Adolescents: An EHR Cohort Study from the RECOVER Initiative

(Yuqing Lei, Ting Zhuo, Bingyu Zhang, Dazheng Zhang, Qiong Wu, Lu Li, Christopher B. Forrest, Caren Mangarelli, Ravi Jhaveri and Yong Chen)



Risk of Dysmetabolic Syndrome in Post-Acute COVID-19 Among Children and Adolescents: An EHR Cohort Study from the RECOVER Initiative

Yuqing Lei[1,3], Ting Zhou[1,3], MD, PhD[1], Bingyu Zhang[1,2], Dazheng Zhang[1,3], Qiong Wu[1,3], Lu Li[1,2], Christopher B. Forrest, MD, PhD[4], Caren Mangarelli, MD[5], Ravi Jhaveri , MD[6,7], and Yong Chen, PhD[1,2,3,8,9,10]

1. The Center for Health Analytics and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA
2. The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA
3. Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, USA
4. Applied Clinical Research Center, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
5. Assistant Professor of Pediatrics (Advanced General Pediatrics and Primary Care), Northwestern University Feinberg School of Medicine, Chicago, IL, USA
6. Pediatric Infectious Diseases, Virginia H. Rogers Professor in Infectious Diseases, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL, USA
7. Professor of Pediatrics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
8. Leonard Davis Institute of Health Economics, Philadelphia, PA, USA
9. Penn Medicine Center for Evidence-based Practice (CEP), Philadelphia, PA, USA
10. Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, USA

Background

- **PASC** stands for post-acute sequelae of SARS-CoV-2 (COVID-19) and is a term scientists are using to study the potential consequences of a COVID-19 infection.
- **Previous Study in US Department of Veterans Affairs databases[1]:**
  - Mean age: 60.21
  - Key finding: higher risks and 1-year burdens of new-onset dyslipidemia in post-acute phase of COVID-19 infection, which underscoring the dyslipidemia as a potential post-acute sequela of SARS-CoV-2 infection.
- **Study Objective:**
  - This study investigates the post-acute risk of metabolic dysfunctions, particularly **dyslipidemia** and **obesity**, in children and adolescents after SARS-CoV-2 infection.

Outcome Definition

- **Dyslipidemia:**
  - Incident abnormal lipid laboratory results:
    - Total cholesterol (TC) ≥ 200 mg/dL
    - Triglycerides (TG):
      - 0 to 9 years: ≥ 100 mg/dL
      - 10 to 19 years: ≥ 130 mg/dL
      - 20 to 21 years; ≥ 150 mg/dL
    - Low-Density Lipoprotein (LDL) cholesterol ≥ 130 mg/dL
    - High-Density Lipoprotein (HDL) cholesterol < 40 mg/dL
    - Non-HDL cholesterol ≥ 145 mg/dL
  - Incident lipid-lowering medications prescriptions: Prescription of statins
- **Obesity:**
  - Incident abnormal BMI:
    - 2 to 19 years: BMI z-scores >= 95th percentile
    - 19 to 21 years: BMI >= 30

Trial Emulation

- Emulate a trial using observational data when RCT is not available: apply observational data to emulate this protocol by identifying eligible individuals, adjusting for baseline confounders to mimic random treatment assignment, and conducting analyses similar to those in the hypothetical trial [2].
- **Bias Indication Issue:**
  - Lipid lab tests are not routinely conducted for pediatric patients, completing the lipid lab test may be associated with both COVID-19 infected status and abnormal lipid lab results
  - Solution: Include whether participants completed a lipid lab test at baseline period as an additional confounding variable in assessing propensity scores and stratification[3,4].

Trial Emulation

- Propensity Score Stratification:
  - Fit Logistic regression: $g(E[\text{COVID Infection Status} \mid \text{Confounding variables}])$
  - Stratify into 6 strata based on propensity scores

Analysis Results

Pediatric patients with COVID-19 infection had an increased risks of developing post-acute metabolic dysfunction, including dyslipidemia and obesity

- **Negative Control Experience**
  - We specified a list of 40 negative control outcomes, to estimate a distribution for systematic error, which would be used to calibrate outcomes of interest[5].
  - E.g. Negative controls for composite dyslipidemia outcome convert to 1.13 after calibration

Reference

[1] Xu, E., Xie, Y., & Al-Aly, Z. (2023). Risks and burdens of incident dyslipidaemia in long COVID: a cohort study. The Lancet Diabetes & Endocrinology, 11(2), 120-128.
[2] Hernán, M. A., Wang, W., & Leaf, D. E. (2022). Target trial emulation: a framework for causal inference from observational data. JAMA, 328(24), 2446-2447.
[3] Psaty, B. M., Koepsell, T. D., Lin, D., Weiss, N. S., Siscovick, D. S., Rosendaal, F. R., Pahor, M., & Furberg, C. D. (1999). Assessment and control for confounding by indication in observational studies. Journal of the America Geriatrics Society, 47(6), 749–754.
[4] Kyriacou, D. N., & Lewis, R. J. (2016). Confounding by Indication in Clinical Research. JAMA, 316(17), 1818–1819.
[5] Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A., & Madigan, D. (2014). Interpreting observational studies: why empirical calibration is needed to correct p-values. Statistics in medicine, 33(2), 209-218.

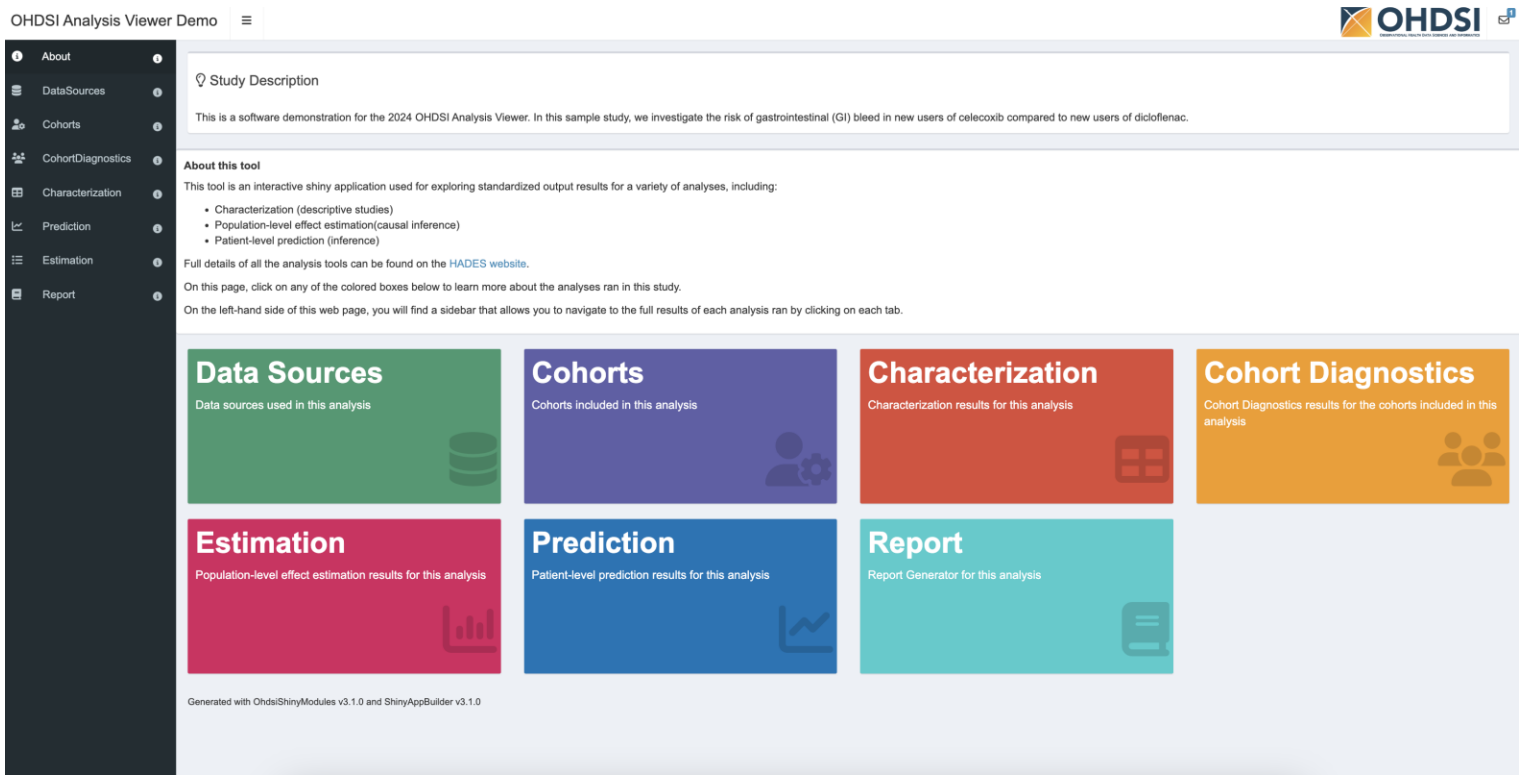
Contact: ychen123@pennmedicine.upenn.edu

#OHDSISocialShowcase This Week

Friday

Advancing the OHDSI Analysis Viewer: Enhanced Performance, Integration, and User Experience

(Nathan Hall, Frank DeFalco, Vishakha Gupta, Jenna Reps)



Job Opening

Senior Program Officer, Clinical AI Innovation, Gates Foundation

Fred Hutch Cancer Center

Analytics Engineer II

Job ID 29138 **Type** Regular Full-Time US-WA-Seattle **Category** Data Science

The **Analytics Engineer II** will join the Data Science team in OCDO to support data excellence as a core value and feature of data science work at Fred Hutch. As an Analytics Engineer, you will collaborate with our data science team to design, build, and maintain multi-modal data models that integrate various data sources. Your work will support AI-driven research and translational data science use cases, helping to advance our scientific and technological goals. You will play a key role in data preparation, model optimization, and ensuring data quality to enable effective data-driven decision-making across the organization. You will be a data librarian, an organizer of Fred Hutch's data sets coming from both internal (e.g., electronic health record) and external (e.g. genomic screenings) clinical applications. At Fred Hutch, Analytics Engineers are passionate about what the data means and work with researchers, clinical leaders, and other stakeholders to understand how data is generated and how it can be used to answer data-driven questions about Fred Hutch and its patients.

Responsibilities

- Work with data engineers and data scientists to develop Fred Hutch's OMOP common data model, bringing in data from multiple clinical domains
- Develop and optimize extensions for multi-modal data models that combine structured and unstructured data, enabling AI-driven translational research and data science use cases.
- Ensure data quality, consistency, and governance standards are met throughout the data lifecycle, to create clean, transformed, curated data sets for clinical researchers using Databricks
- Use software engineering best practices to test data quality and deploy data models and transformations efficiently and accurately.
- Democratize data access and knowledge by documenting data sets as well as the code/tools used to generate them, employing automation as much as possible. This includes documenting data workflows, data models, and pipeline designs.
- Help to define and improve standards for style, maintainability, and best practices for data transformations with our Engineering team to support a seamless transition from our infrastructure to Fred Hutch data science and analytics teams.
- Develop a deep understanding of the data needs of clinician-researchers and train them to use curated, self-service data sets as well as analysis and visualization tools
- Collaborate with data scientists, analytic engineers, and software developers to understand data requirements, integrations, and optimize workflows for performance, scalability and reliability.

Vocabulary Update

Presenter: Masha Khitrun

Book of OHDSI Update

Presenters: Christian Reich, Sarah Seager

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-2025