OHDSI/OMOP Research Spotlight

OHDSI Community Call
March 24, 2026 • 11 am ET

Upcoming Community Calls

Date	Topic
Mar. 24	OHDSI/OMOP Research Spotlight
Mar. 31	Kickoff to Phenotype April
Apr. 7	Phenotype April, Week 1
Apr. 14	Phenotype April, Week 2
Apr. 21	NO MEETING / EUROPE SYMPOSIUM
Apr. 28	Pheotype April, Week 4
May 5	Europe Symposium Review/Phenotype April Finale

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Man Young Park, Jaeuk U. Kim, SunMi Choi, Youngheum Yoon, Byung-Kwan Seo, and Sangkwan Lee** on the recent publication of **Transforming Traditional Korean Medicine hospital EHRs into the OMOP common Data Model: methodology and implications** in *BMC Medical Informatics and Decision Making*.

Home > BMC Medical Informatics and Decision Making > Article

Transforming Traditional Korean Medicine hospital EHRs into the OMOP common Data Model: methodology and implications

Research | Open access | Published: 16 March 2026

article number , (2026) Cite this article

You have full access to this open access article

Download PDF | Save article

Man Young Park, Jaeuk U. Kim, SunMi Choi, Youngheum Yoon, Byung-Kwan Seo & Sangkwan Lee

Abstract

Background

Standardizing data from Traditional Korean Medicine (TKM) is essential for enhancing interoperability with international real-world data infrastructures, such as multi-institutional OMOP-CDM databases within the OHDSI network, multinational claims databases, and large-scale clinical data repositories, thereby enabling evidence-based research. This study aimed to convert electronic health records (EHRs) from a TKM hospital into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

OHDSI Shoutouts!

Congratulations to the team of **Khaoula El Mekkaoui, Chitra Mukherjee, Clodomir Santana, Hitalo Silva, Arnib Quazi, Meghana Balabhadruni, Aditya Ballal, Shamika Gavaskar, Kea Turqueza, Leslie Molina, Pablo Acevedo, Kay Moua, Omar Dawar, Federico Garcia, Mohammad S Soroya, XiaoDong Zhang, Ezra Amsterdam, Chao-Yin Chen, Padmini Sirish, Leighton Izu, Ronaldo Menezes, Nipavan Chiamvimonvat, Vladimir Filkov, Martin Cadeiras, and David A Liem** on the recent publication of **The impact of the Social Exposome in Cardiovascular Health and Disease** in *The Journal of Precision Health: Health and Disease*.

The Journal of Precision Medicine: Health and Disease 3 (2025) 100015

Contents lists available at ScienceDirect

The Journal of Precision Medicine: Health and Disease

journal homepage: www.sciencedirect.com/journal/the-journal-of-precision-medicine-health-and-disease

The impact of the social exposome in cardiovascular health and disease

Khaoula El Mekkaoui [a,b,c], Chitra Mukherjee [a], Clodomir Santana [a], Hitalo Silva [a], Arnib Quazi [a], Meghana Balabhadruni [a], Aditya Ballal [d], Shamika V. Gavaskar [a], Kea T. Turqueza [a], Leslie Molina [a], Pablo E. Acevedo [a], Kay Moua [a], Omar Dawar [a], Federico Garcia [a], Mohammad S. Soroya [a], Xiao-Dong Zhang [a], Ezra A. Amsterdam [a], Chao-Yin Chen [d], Padmini Sirish [a], Leighton T. Izu [d], Ronaldo Menezes [e], Nipavan Chiamvimonvat [a,f], Vladimir Filkov [g], Martin Cadeiras [a], David A. Liem [a,*]

[a] Department of Medicine, Division of Cardiovascular Disease, University of California, Davis, USA
[b] VTT Technical Research Centre of Finland, Espoo, Finland
[c] Department of Computer Science, Aalto University, Espoo, Finland
[d] Department of Pharmacology, School of Medicine, University of California, Davis, USA
[e] Department of Computer Science, University of Exeter, Exeter, UK
[f] Department of Basic Medical Sciences and Translational Cardiovascular Research Center, University of Arizona, College of Medicine, Phoenix, USA
[g] Department of Computer Science, University of California, Davis, USA

ARTICLE INFO

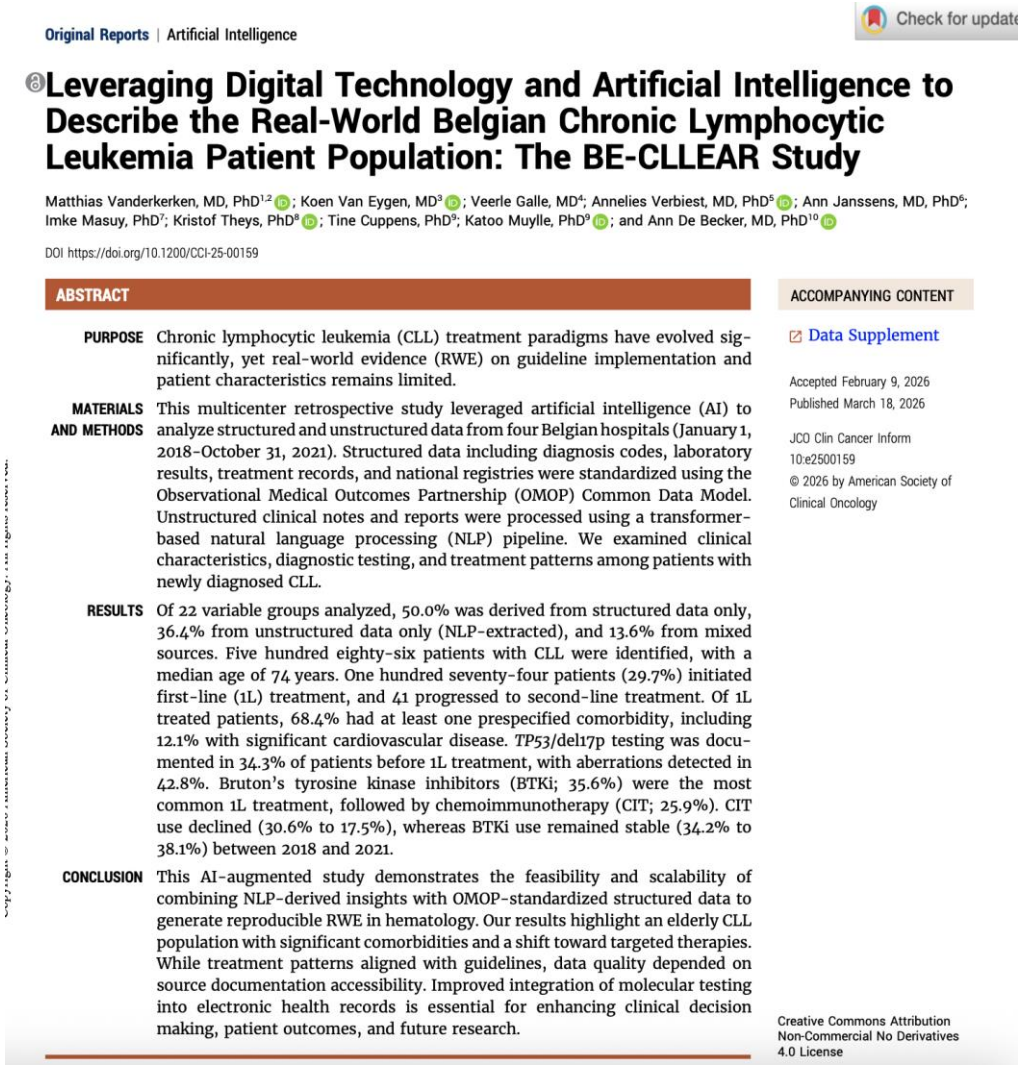
Keywords:
Heart disease
Exposome
Social determinants of health
Multimodal data

ABSTRACT

Cardiovascular disease (CVD) remains a leading cause of mortality and morbidity worldwide, influenced by complex interactions between biological, environmental, and social factors. Studies employing traditional genetic approaches are often insufficient to capture the broader context of these heterogeneous factors. We highlight an integrative framework that combines electronic health records (EHRs) with public (census) data on socio-environmental exposures, leveraging the Observational Medical Outcomes Partnership (OMOP) Common Data Model and individual's geolocations linked through ZIP Code Tabulation Area (ZCTA) codes. By incorporating the exposome, including social determinants of health (SDoH), this framework enhances privacy-preserving analysis and improves personalized CVD risk stratification. Furthermore, Precision Medicine aims to identify tailored prevention and therapeutic strategies by considering an individual's unique genetic, environmental, and lifestyle factors in which the exposome plays a pivotal role. Accordingly, in this review we will discuss a holistic precision medicine approach centered around the exposome to address health disparities, to better identify communities at high-risk for CVD, and to advance tailored prevention and intervention strategies.

OHDSI Shoutouts!

Congratulations to the team of **Matthias Vanderkerken, Koen Van Eygen, Veerle Galle, Annelies Verbiest, Ann Janssens, Imke Masuy, Kristof Theys, Tine Cuppens, Katoo Muylle, and Ann De Becker** on the recent publication of **Leveraging Digital Technology and Artificial Intelligence to Describe the Real-World Belgian Chronic Lymphocytic Leukemia Patient Population: The BE-CLLEAR Study** in *JCO Clinical Cancer Informatics*.



Original Reports | Artificial Intelligence

Check for updates

Leveraging Digital Technology and Artificial Intelligence to Describe the Real-World Belgian Chronic Lymphocytic Leukemia Patient Population: The BE-CLLEAR Study

Matthias Vanderkerken, MD, PhD[1,2]; Koen Van Eygen, MD[3]; Veerle Galle, MD[4]; Annelies Verbiest, MD, PhD[5]; Ann Janssens, MD, PhD[6]; Imke Masuy, PhD[7]; Kristof Theys, PhD[8]; Tine Cuppens, PhD[9]; Katoo Muylle, PhD[9]; and Ann De Becker, MD, PhD[10]

DOI https://doi.org/10.1200/CCI-25-00159

ABSTRACT

PURPOSE Chronic lymphocytic leukemia (CLL) treatment paradigms have evolved significantly, yet real-world evidence (RWE) on guideline implementation and patient characteristics remains limited.

MATERIALS AND METHODS This multicenter retrospective study leveraged artificial intelligence (AI) to analyze structured and unstructured data from four Belgian hospitals (January 1, 2018–October 31, 2021). Structured data including diagnosis codes, laboratory results, treatment records, and national registries were standardized using the Observational Medical Outcomes Partnership (OMOP) Common Data Model. Unstructured clinical notes and reports were processed using a transformer-based natural language processing (NLP) pipeline. We examined clinical characteristics, diagnostic testing, and treatment patterns among patients with newly diagnosed CLL.

RESULTS Of 22 variable groups analyzed, 50.0% was derived from structured data only, 36.4% from unstructured data only (NLP-extracted), and 13.6% from mixed sources. Five hundred eighty-six patients with CLL were identified, with a median age of 74 years. One hundred seventy-four patients (29.7%) initiated first-line (1L) treatment, and 41 progressed to second-line treatment. Of 1L treated patients, 68.4% had at least one prespecified comorbidity, including 12.1% with significant cardiovascular disease. *TP53*/del17p testing was documented in 34.3% of patients before 1L treatment, with aberrations detected in 42.8%. Bruton's tyrosine kinase inhibitors (BTKi; 35.6%) were the most common 1L treatment, followed by chemoimmunotherapy (CIT; 25.9%). CIT use declined (30.6% to 17.5%), whereas BTKi use remained stable (34.2% to 38.1%) between 2018 and 2021.

CONCLUSION This AI-augmented study demonstrates the feasibility and scalability of combining NLP-derived insights with OMOP-standardized structured data to generate reproducible RWE in hematology. Our results highlight an elderly CLL population with significant comorbidities and a shift toward targeted therapies. While treatment patterns aligned with guidelines, data quality depended on source documentation accessibility. Improved integration of molecular testing into electronic health records is essential for enhancing clinical decision making, patient outcomes, and future research.

ACCOMPANYING CONTENT

Data Supplement

Accepted February 9, 2026
Published March 18, 2026

JCO Clin Cancer Inform
10:e2500159

OHDSI Shoutouts!

Congratulations to the team of **Francesco Pignatti, Tarec Christoffer El-Galaly, Martin Kaiser, Kimmo Porkka, Robin Doeswijk, Peter Mol, Donna R. Rivera, Catherine C. Lerro, Ulrich-Peter Rohr, Patrice Verpillat, Antonios Valachis, Dario Trapani, Massimo Di Maio, Nicola Latino, Raul Cordoba, Nathan Cherny, Miriam Koopman, Diogo Martins-Branco, George Pentheroudakis, and Douwe Postmus** on the recent publication of **Assessing Overall Survival Benefits in Advanced Cancers: The Role of External Comparator Cohort Studies with Real-World Data** in *Clinical Pharmacology & Therapeutics*.

ARTICLE

Assessing Overall Survival Benefits in Advanced Cancers: The Role of External Comparator Cohort Studies with Real-World Data

Francesco Pignatti[1,*], Tarec Christoffer El-Galaly[2], Martin Kaiser[3], Kimmo Porkka[4], Robin Doeswijk[5], Peter Mol[6,7], Donna R. Rivera[8], Catherine C. Lerro[8], Ulrich-Peter Rohr[9], Patrice Verpillat[1], Antonios Valachis[10], Dario Trapani[11,12], Massimo Di Maio[13], Nicola Latino[14], Raul Cordoba[15], Nathan Cherny[16], Miriam Koopman[17], Diogo Martins-Branco[14], George Pentheroudakis[14] and Douwe Postmus[6]

External comparator cohort (ECC) studies with real-world data (RWD) may provide more reliable estimates of treatment differences compared to single-arm trials (SAT), yet they face limitations such as selection bias and data heterogeneity. This study assessed the perceived strength of evidence of ECC studies compared to SAT and randomized controlled studies (RCT). The study included healthcare professionals (HCP) from the European Hematology Association (EHA), the European Society for Medical Oncology (ESMO), and assessors from international regulatory agencies (RA). A conjoint analysis evaluated strength of evidence ratings for establishing an effect on OS for different hypothetical scenarios, based on different designs, RWD quality, and observed OS improvement, for a new cancer treatment for advanced disease and no effective treatments. Participants from HCP organizations rated RWD studies favorably (advantages outweigh disadvantages) more frequently (47.6%; $n=103$) compared to RA participants (12.9%; $n=116$). Compared to a SAT, a high-quality RWD ECC study showing a 1.5-month and 3-month OS improvement had 2.7 (95% CI: 1.9–3.8) and 14.7 (95% CI: 10.0–21.5) times higher odds of receiving a higher strength of evidence rating, respectively. The OR for RCT v. SAT was 36.4 (95% CI: 24.0–55.2) and 358.4 (95% CI: 217.3–591.3), respectively. Strength of evidence ratings were associated with maximum acceptable risk of severe or symptomatic toxicity. In conclusion, when evaluating the OS of new therapies, ECC studies with RWD, especially when based on high-quality RWD or demonstrating a larger OS benefit, were rated as more convincing than SAT without a formal control.

OHDSI Shoutouts!

Congratulations to the team of **Montserrat León-García, Sergio Álvarez-Pérez, Janire Gesto-Gómez, Clara Urbano-Molina, Sonia Soto-Díaz, Juan Cárdenas-Valladolid, Luis Rodríguez-Rodríguez, Antonio Díaz-Holgado, Isabel Del Cura-González, Javier De La Cruz-Bertolo, Noelia García-Barrio, Juan Luis Cruz-Bermúdez, Cristina García-Fernández, Carlos Rodríguez-Antolín, Laila García-Aldars, Elsa María Moreda-Sánchez, Álvaro Roldán López, José María Veganzones Alonso-Cortés, Ana Isabel Gonzalez Gonzalez, Miguel A Salinero-Fort, and the HealthData@MAD-R&I Working Group** on the recent publication of **HealthData@MAD-R&I: Protocol for Design and Development of a Regional Health Data Infrastructure to Enable Secondary Use of Health Data in Research and Innovation** in *JMIR Research Protocols*.

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

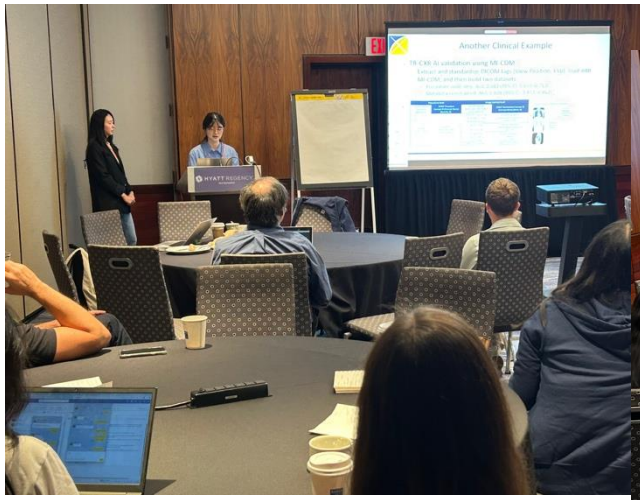
Where Are We Going?

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Wednesday	10 am	Surgery and Perioperative Medicine
Wednesday	10 am	Women of OHDSI
Wednesday	10 am	Common Data Model
Wednesday	1 pm	Latin America Chapter
Thursday	6:30 am	India Community Call
Thursday	9 am	Phenotype Development and Evaluation
Thursday	10 am	Africa Chapter (ZOOM)
Thursday	10 am	GIS – Geographic Information System
Thursday	11 am	Perinatal & Reproductive Health
Friday	9 am	Network Data Quality
Friday	11 am	Clinical Trials
Monday	9 am	Vaccine Vocabulary
Tuesday	9 am	Open EHR and OMOP
Tuesday	10 am	CDM Survey

Workgroups Leadership Retreat

Attention, workgroup leads: There will be a workgroup leadership retreat this Friday, March 27, from 10-11:30 am ET in our MS Teams environment. If you didn't receive the call invite, please reach out to Paul Nagy or Craig Sachson.



ohdsi.org/workgroups

UK Symposium Call for Abstracts Opens

HDR UK Event

OHDSI UK 2026

We're delighted to announce that OHDSI UK 2026 will be held on the 18th of September at the University of Nottingham. For the first time, there will also be an OMOP training day on the 17th of September.

Share this page

OHDSI (Observational Health Data Sciences and Informatics, pronounced "Odyssey") is an international community of stakeholders dedicated to unlocking the value of health data through large-scale analytics. OHDSI promotes open science and collaboration in health data research with a key focus on adoption of the OMOP Common Data Model, a global standard for harmonising data and facilitating federated analytics across institutions. **Find out more about OHDSI.**

Call for Abstracts

We invite you to submit an abstract for consideration at OHDSI UK 2026. Whether you wish to present a poster, software demo, or lightning talk, we welcome contributions from across the community. Abstract submission is available via this form, and the deadline is 1st May 2026. Please use this template to prepare your abstract and save it as a PDF, and start your file name with the surname (family name) of the presenting author.

Key dates:

Registration Opens: 20th April 2026

Registration Closes: 4th September 2026

Abstract Submission Opens: 20th March 2026

Abstract Submission Deadline: 1st May 2026

Training day: 17th September 2026

Symposium: 18th September 2026

OMOP School in Stockholm, Sweden

passion2improve edence Health

The OMOP School

3+1-day OMOP CDM Bootcamp

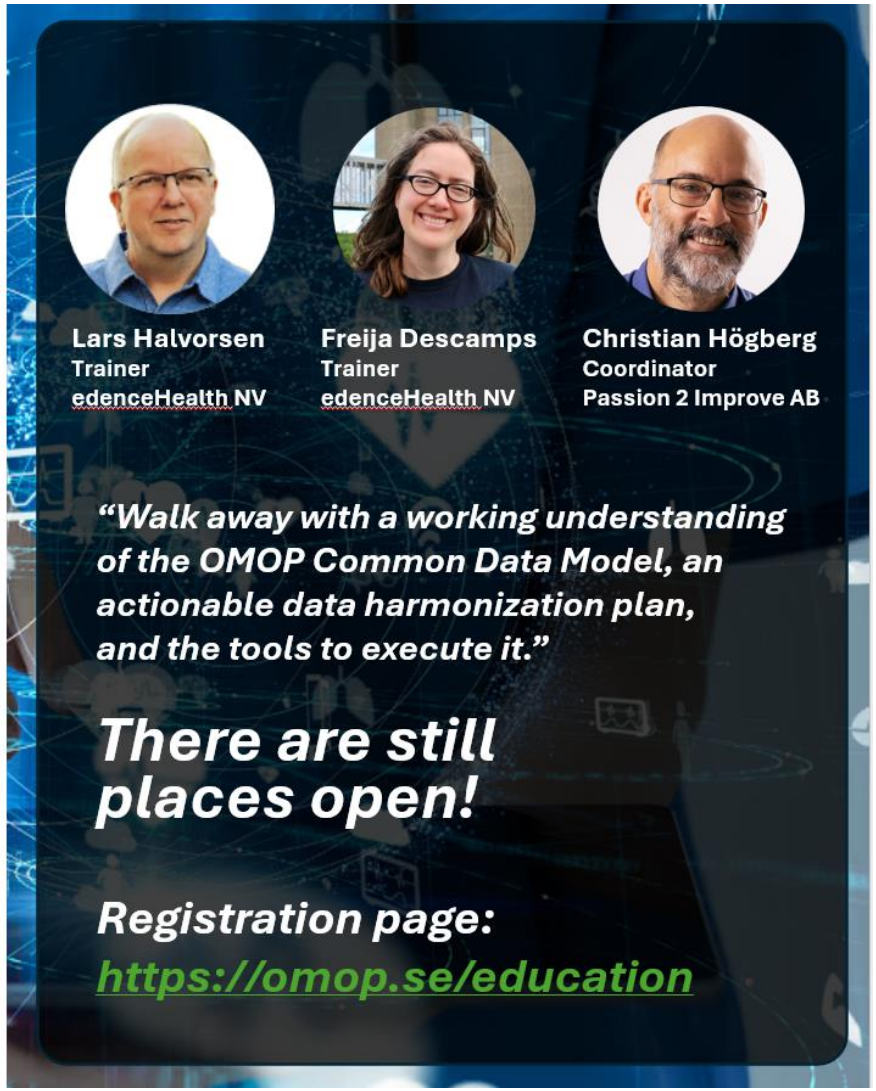
A hands-on training and workshop that turns your data harmonization vision into reality.

May 26th – 28th
+ May 29th (optional extra day for Use Cases Deep Dive)
09:00-17:00 Tue-Thu, 09:00-16:00 Fri
Stockholm, Sweden (venue TBD)

Learning objectives:
- Explain the role of standardization in federated research
- Understand the OMOP Common Data Model and how it can be applied
- Perform semantic mapping using OMOP vocabularies
- Design and implement an OMOP ETL process
- Evaluate and improve data quality
- Conduct standardized observational analyses
- Execute an end-to-end mini-harmonization project
- Use selective OHDSI methods and tools

Who to attend?
- Health Data Owners/Data Custodians
- Data Scientists, Clinical Researchers & Epidemiologists
- IT/Data Architects/ETL Developers/Health Informatics Specialists
- Digital Transformation Leads, Registry Directors, Healthcare Strategists
- Healthcare Policy Stakeholders

OMOP4SWEDEN!
EVIDENCE GENERATION FOR BETTER HEALTHCARE



First Latin America Symposium – July 30-31

2026 Europe Symposium

The 2026 OHDSI Europe Symposium returns to Rotterdam next year and will be held **April 18-20.**

Registration is open on the **OHDSI & OHDSI Europe** web sites.

Time	Symposium Agenda - Monday April 20, 2026	Location
8:00	***Registration and Coffee***	Queen's Lounge
9:00	**Welcome to OHDSI Europe** Dr. Renske Los, Department of Medical Informatics, *Erasmus MC* Dr. Aniek Markus, Department of Medical Informatics, *Erasmus MC*	Theatre
9:05	**Journey of OHDSI** Prof. Peter Rijnbeek, Chair Department of Medical Informatics, *Erasmus MC*	Theatre
9:30	**Collaborator Showcase - part 1** Moderated by Dr. Egill Fridgeirsson, Department of Medical Informatics, *Erasmus MC*	Theatre
10:00	**Speed networking**	Theatre
10:15	***Coffee Break & posters National Nodes***	Queen's Lounge
11:15	**Collaborator Showcase - part 2** Moderated by Dr. Egill Fridgeirsson, Department of Medical Informatics, *Erasmus MC*	Theatre
11:45	**Dreaming about the OHDSI journey ahead** Dr. Patrick Ryan, Vice President, Observational Health Data Analytics, *Johnson & Johnson* Dr. Renske Los, Department of Medical Informatics, *Erasmus MC*	Theatre
12:15	***Lunch break & networking & posters/demo's*** *(Early investigator meeting - 13:00-13:45 Queen's Lounge)*	La Fontaine & Odyssee Room
13:45	**From dreams to reality** OHDSI Titan Award winners	Theatre
14:30	**Propositions for collaboration from the National Nodes** National Node leads	Theatre
14:45	***Coffee break & posters/demo's***	La Fontaine & Odyssee Room
16:15	**The OH Factor** To be announced	Theatre
17:00	**Closing**	Theatre
17:15	***Networking reception***	Queen's Lounge

2026 Global Symposium

The 2026 OHDSI Global Symposium will return to the Hyatt Regency Hotel in New Brunswick, N.J., on **Oct. 20-22.**

#OHDSISocialShowcase This Week

Monday

Building the OHDSI Evidence Network – A Global, Open, Federated Collaboration

(Clair Blacketer, Haeun Lee, Benjamin Martijn, Evanette Burrows, Patricia Mabry, Deran McKeen, Sam Patnoe, Ben Gerber, Pantelis Natsiavas, Aamirah Vadsariya, Hanieh Razzaghi, Paul Nagy)

Building the OHDSI Evidence Network: *A Global, Open, Federated Collaboration*

PRESENTER: **Clair Blacketer**

INTRODUCTION

- **Real-world data is plentiful** and reflects natural conditions, **but is siloed, due to privacy concerns**, preventing the benefits of dataset integration from being realized, e.g. study of rare events, generalizability, use of data-hungry AI tools to reveal new insights
- **Federated networks address this problem** by sharing only aggregated results (not record level data) to preserve data privacy
- **The *OHDSI Evidence Network* was launched in 2024** inspired by the success of other federated networks, e.g., European Health Data and Evidence Network (EHDEN) and the Data Analysis and Real World Interrogation Network (DARWIN EU).

METHODS

- **The Evidence Network (EN) is composed of "Data Partner Organizations" (DPOs)** who volunteer to run analytic code on their organization's data.
- **Membership in EN is voluntary** - no contracts or centralized data sharing!
- **Governance is decentralized**; each DPO adheres to its local IRB requirements.
- **To catalog data available** in the EN, each DPO is sent a **Database Diagnostics software package** which they run **locally to produce a standardized DbProfiles** - aggregated metadata describing the DPO's database(s)
- **All EN activities are opt-in** and include EN workgroup meetings, steering committee representation, monthly data partner calls, and EN study co-development
- **A pilot study**, "Save Our Sisyphus", measured partner engagement. Results **led to the adoption of best practices by the EN** (learning, clear protocols, transparent communication).

RESULTS

- **28 DPOs onboarded since inception, contributing access to 48 databases** across 4 continents (see map)
- The EN supported **20 rapid fit-for-purpose assessments** and study co-developments in 2025

Figure 1: *Global map of current OHDSI Evidence Network data partner organizations.*

KEY LESSONS:

- **Decentralized**, federated, community-led **governance is feasible** and effective at a global scale
- **Trust and transparency** drive collaboration
- Low-burden participation lowers barriers
- Shared tools enable shared learning

FUTURE GOALS:

- Address funding/sustainability challenges
- Develop and test a process for study development and support
- Refine DPO–study matching
- Expand DPO membership

Clair Blacketer[1,2,4], Haeun Lee[1,8], Benjamin Martijn[1,8], Evanette Burrows[1,4], Patricia Mabry[1,10], Deran McKeen[1,10], Sam Patnoe[1,10], Elizabeth Grossman[1,10], Ben Gerber[1,5], Pantelis Natsiavas[1,6], Aamirah Vadsariya[1,7], Hanieh Razzaghi[1,9], Paul Nagy[1,8]

1. OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA 2. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, NL 3. Department of Biomedical Informatics, Columbia University, New York, NY, USA 4. Johnson & Johnson, Raritan, NJ, USA 5. Department of Population and Quantitative Health Sciences, UMass Chan Medical School, Boston, MA, USA 6. Institute of Applied Biosciences, Centre for Research & Technology Hellas, Thessaloniki, GR 7. Clinical Informatics Center, University of Texas Southwestern Medical Center, Dallas, TX, USA 8. Johns Hopkins University, Baltimore, MD, USA 9. Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, PA 10.Health Partners Institute, Bloomington, MN, USA

OHDSI

#OHDSISocialShowcase This Week

Tuesday

Bridging FHIR and OMOP: Data Lineage for Observational Data Conversion

(Benjamin Berk, Melissa Benzie, Bindu Bolisetty, Scott Favre, Jeremy Fortune, Jeremy Goslin, Vik Kheterpal, Kathleen Marinan, Anne Marsan, Ed Ramos, Sunanda Venumuddula, John Wyderko)

Title: *Bridging FHIR & OMOP: Data Lineage for Observational Data Conversion*

PRESENTER: **Benjamin Berk MD**

INTRO

While FHIR excels as a clinical data exchange standard, OMOP provides a standardized format optimized for observational research. Converting data between these models presents significant challenges in maintaining data integrity and traceability. In converting from FHIR to OMOP, we would like to capture:

- **Data Source Provenance:** Tracking source system provenance of each element.
- **Entity Lineage:** Maintaining traceability from FHIR resources to OMOP table rows.
- **Concept Standardization:** Recording source codes to OMOP standard concepts.
- **Processing Events:** Logs of transformation decisions, including warnings, informational messages, errors, and documentation of unmapped FHIR resources

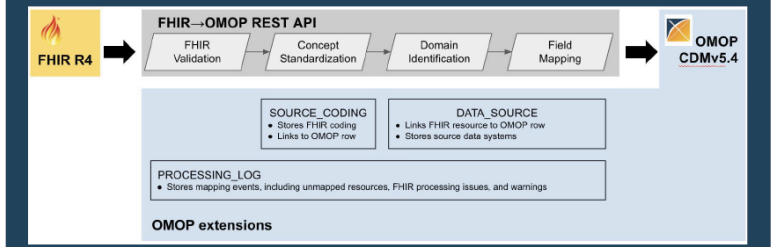
METHODS

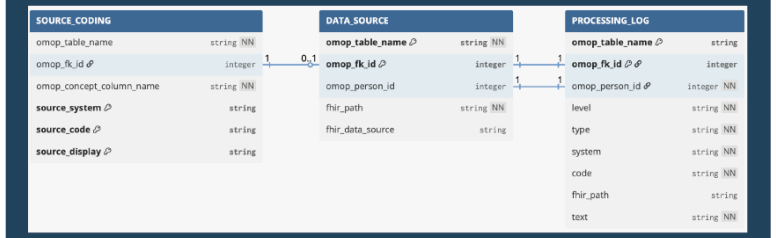
1. Developed REST-based FHIR→OMOP conversion API that implements a data transformation pipeline
2. Source codes mapped to OMOP standard concepts via CareEvolution Orchestrate Terminology API (using exact code and display matching & natural language processing)
3. Worked with HL7 Vulcan FHIR→OMOP Working Group to develop a standard implementation guide for this transformation (now in HL7 Balloting)
4. Captured data lineage in 3 extension tables
5. Tested with data collected from PRediction Of Glycemic RESponse Study (PROGRESS) study conducted by Scripps Digital Clinical Trials Center using CareEvolution's MyDataHelps research platform

RESULTS

- Processed over 2.8 million FHIR resources, generating comprehensive lineage data through three extension tables.
- 1.1 million FHIR resource IDs linked to corresponding OMOP table rows
- 3.8 million source codings, including successful standardizations and unmappable source codes

Proposal: Integrate data lineage tables into CDM for FHIR→OMOP conversion

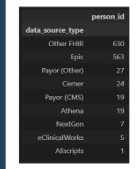




Take a picture to download the full paper

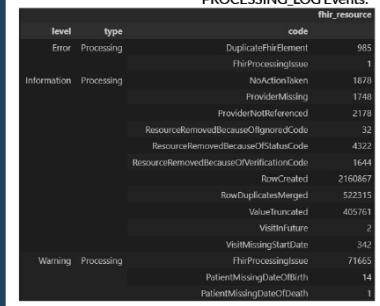
Now Under Ballot:
https://build.fhir.org/ig/HL7/fhir-omop-ig
https://www.hl7.org/ballotdesktop/

665 individuals sourced via participant-mediated exchange using MyDataHelps platform from diverse EMRs and other FHIR providers

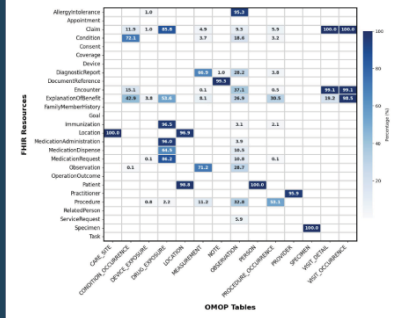


More than 99% of all non-duplicate resources generated rows in OMOP tables

PROCESSING_LOG Events:



DATA_SOURCE table enables tracing OMOP to FHIR:



Berk BB, Benzie M, Bolisetty B, Favre S, Fortune J, Goslin J, Kheterpal V, Marinan K, Marsan A, Ramos E, Venumuddula S, Wyderko J
[1]All authors: CareEvolution, Ann Arbor, MI, USA; [2]Also affiliated with: Digital Trials Center, Scripps Research Translational Institute, La Jolla, CA, USA)

Scan me!

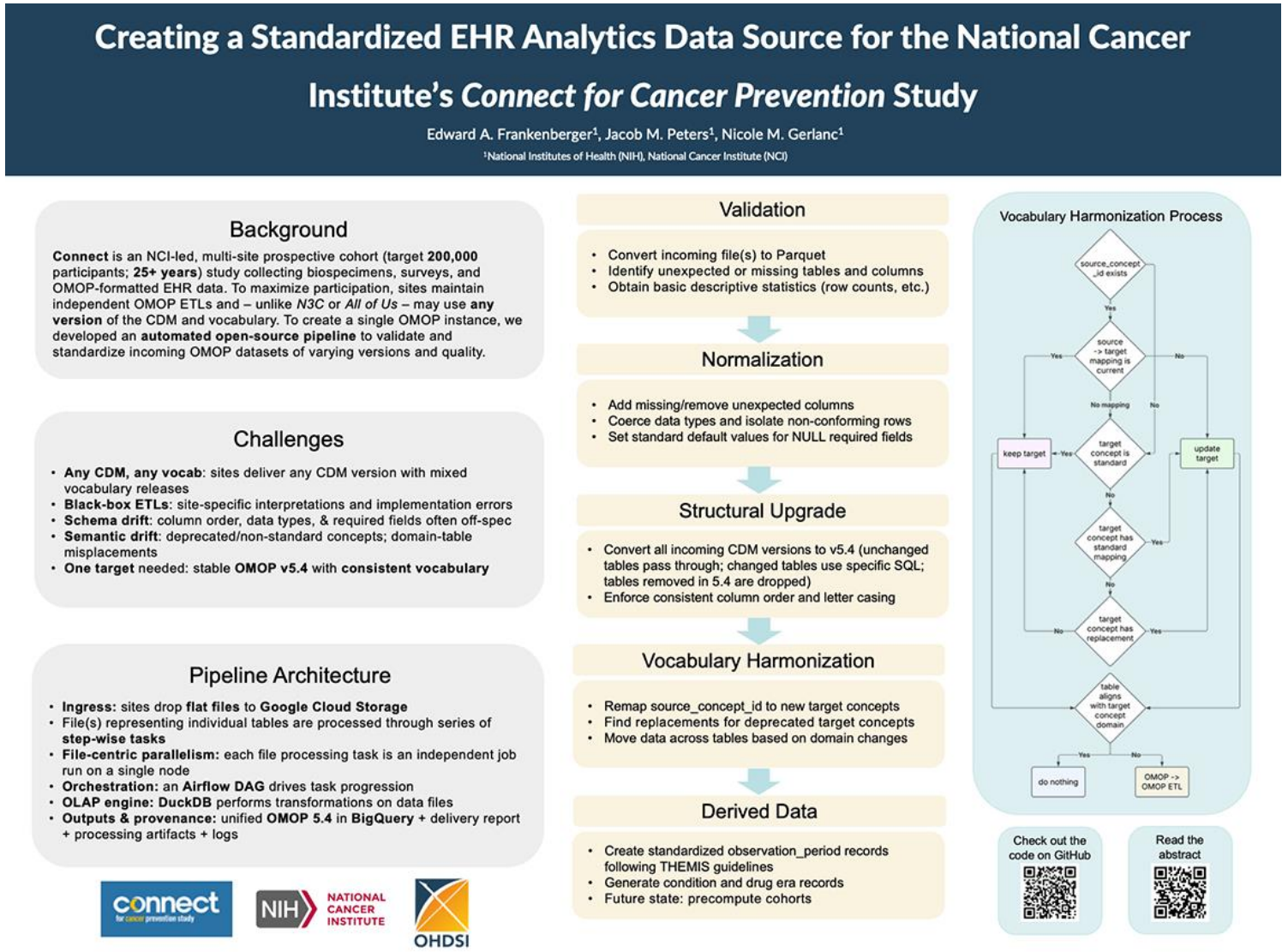
Care Evolution
Orchestrate
MyDataHelps
OHDSI

#OHDSISocialShowcase This Week

Wednesday

Creating a Standardized EHR Analytics Data Source for the National Cancer Institute's Connect for Cancer Prevention Study

(Edward A. Frankenberger, Jacob M. Peters, Nicole M. Gerlanc)

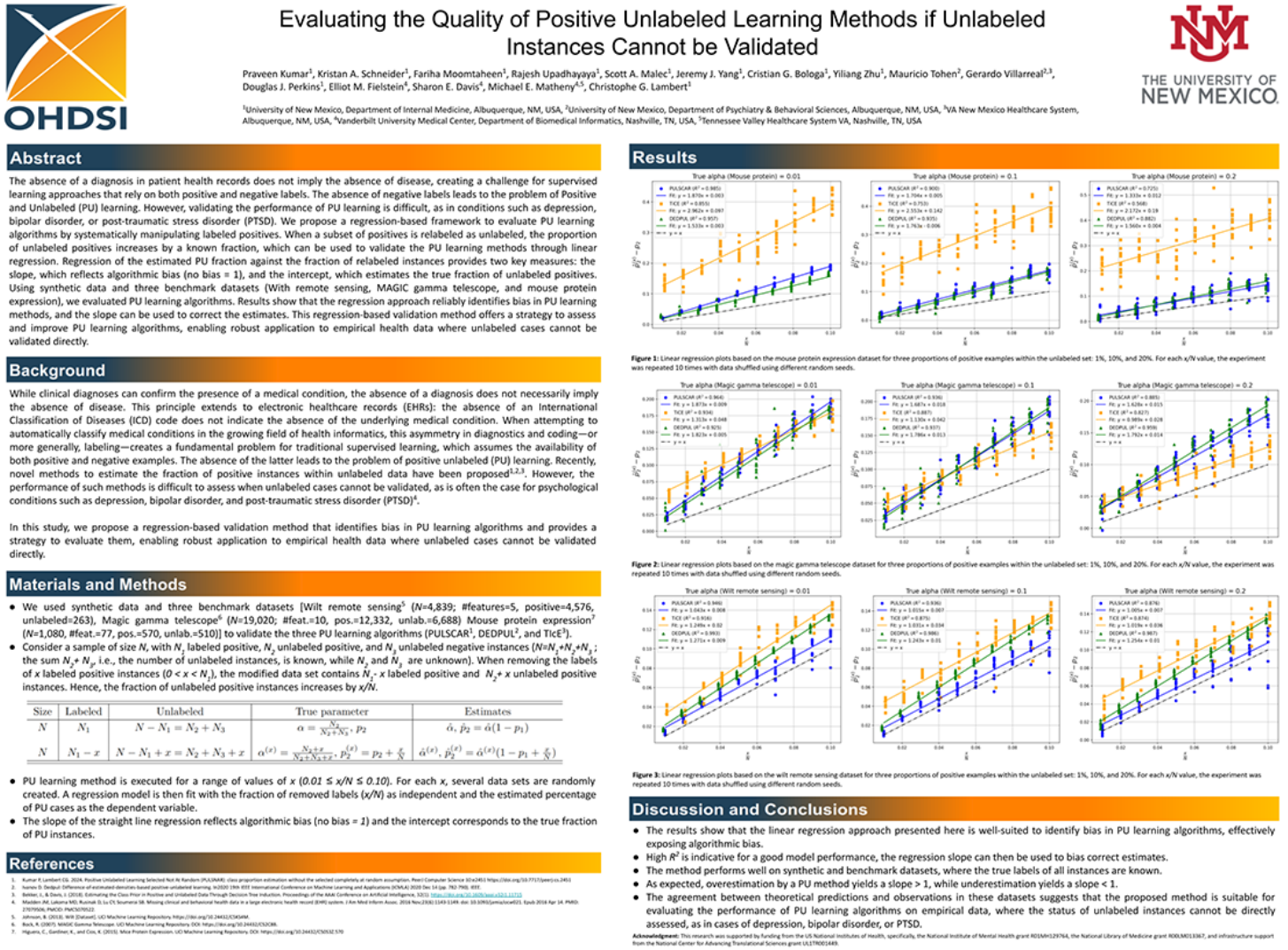


#OHDSISocialShowcase This Week

Thursday

Evaluating the Quality of Positive Unlabeled Learning Methods if Unlabeled Instances Cannot be Validated

(Praveen Kumar, Kristan A Schneider, Fariha Moomtaheen, Rajesh Upadhayaya, Scott A. Malec, Jeremy J. Yang, Cristian G. Bologa, Yiliang Zhu, Mauricio Tohen, Gerardo Villarreal, Douglas J. Perkins, Elliot M. Fielstein, Sharon E. Davis, Michael E. Matheny, Christophe G. Lambert)



Evaluating the Quality of Positive Unlabeled Learning Methods if Unlabeled Instances Cannot be Validated

Praveen Kumar[1], Kristan A. Schneider[1], Fariha Moomtaheen[1], Rajesh Upadhayaya[1], Scott A. Malec[1], Jeremy J. Yang[1], Cristian G. Bologa[1], Yiliang Zhu[1], Mauricio Tohen[2], Gerardo Villarreal[2,3], Douglas J. Perkins[1], Elliot M. Fielstein[4], Sharon E. Davis[4], Michael E. Matheny[4,5], Christophe G. Lambert[1]

[1]University of New Mexico, Department of Internal Medicine, Albuquerque, NM, USA, [2]University of New Mexico, Department of Psychiatry & Behavioral Sciences, Albuquerque, NM, USA, [3]VA New Mexico Healthcare System, Albuquerque, NM, USA, [4]Vanderbilt University Medical Center, Department of Biomedical Informatics, Nashville, TN, USA, [5]Tennessee Valley Healthcare System VA, Nashville, TN, USA

Abstract

The absence of a diagnosis in patient health records does not imply the absence of disease, creating a challenge for supervised learning approaches that rely on both positive and negative labels. The absence of negative labels leads to the problem of Positive and Unlabeled (PU) learning. However, validating the performance of PU learning is difficult, as in conditions such as depression, bipolar disorder, or post-traumatic stress disorder (PTSD). We propose a regression-based framework to evaluate PU learning algorithms by systematically manipulating labeled positives. When a subset of positives is relabeled as unlabeled, the proportion of unlabeled positives increases by a known fraction, which can be used to validate the PU learning methods through linear regression. Regression of the estimated PU fraction against the fraction of relabeled instances provides two key measures: the slope, which reflects algorithmic bias (no bias = 1), and the intercept, which estimates the true fraction of unlabeled positives. Using synthetic data and three benchmark datasets (With remote sensing, MAGIC gamma telescope, and mouse protein expression), we evaluated PU learning algorithms. Results show that the regression approach reliably identifies bias in PU learning methods, and the slope can be used to correct the estimates. This regression-based validation method offers a strategy to assess and improve PU learning algorithms, enabling robust application to empirical health data where unlabeled cases cannot be validated directly.

Background

While clinical diagnoses can confirm the presence of a medical condition, the absence of a diagnosis does not necessarily imply the absence of disease. This principle extends to electronic healthcare records (EHRs): the absence of an International Classification of Diseases (ICD) code does not indicate the absence of the underlying medical condition. When attempting to automatically classify medical conditions in the growing field of health informatics, this asymmetry in diagnostics and coding—or more generally, labeling—creates a fundamental problem for traditional supervised learning, which assumes the availability of both positive and negative examples. The absence of the latter leads to the problem of positive unlabeled (PU) learning. Recently, novel methods to estimate the fraction of positive instances within unlabeled data have been proposed[1,2,3]. However, the performance of such methods is difficult to assess when unlabeled cases cannot be validated, as is often the case for psychological conditions such as depression, bipolar disorder, and post-traumatic stress disorder (PTSD)[4].

In this study, we propose a regression-based validation method that identifies bias in PU learning algorithms and provides a strategy to evaluate them, enabling robust application to empirical health data where unlabeled cases cannot be validated directly.

Materials and Methods

- We used synthetic data and three benchmark datasets [Wilt remote sensing[5] (N=4,839; #features=5, positive=4,576, unlabeled=263), Magic gamma telescope[6] (N=19,020; #feat.=10, pos.=12,332, unlab.=6,688) Mouse protein expression[7] (N=1,080, #feat.=77, pos.=570, unlab.=510)] to validate the three PU learning algorithms (PULSCAR[1], DEDPUL[2], and TIcE[3]).
- Consider a sample of size N, with N_1 labeled positive, N_2 unlabeled positive, and N_3 unlabeled negative instances ($N=N_1+N_2+N_3$); the sum N_2+N_3, i.e., the number of unlabeled instances, is known, while N_2 and N_3 are unknown). When removing the labels of x labeled positive instances ($0 < x < N_1$), the modified data set contains $N_1 - x$ labeled positive and $N_2 + x$ unlabeled positive instances. Hence, the fraction of unlabeled positive instances increases by x/N.

Size	Labeled	Unlabeled	True parameter	Estimates
N	N_1	$N - N_1 = N_2 + N_3$	$\alpha = \frac{N_2}{N_2+N_3}$, p_2	$\hat{\alpha}$, $\hat{p}_2 = \hat{\alpha}(1-p_1)$
N	$N_1 - x$	$N - N_1 + x = N_2 + N_3 + x$	$\alpha^{(x)} = \frac{N_2+x}{N_2+N_3+x}$, $p_2^{(x)} = p_2 + \frac{x}{N}$	$\hat{\alpha}^{(x)}$, $\hat{p}_2^{(x)} = \hat{\alpha}^{(x)}(1 - p_1 + \frac{x}{N})$

- PU learning method is executed for a range of values of x ($0.01 \le x/N \le 0.10$). For each x, several data sets are randomly created. A regression model is then fit with the fraction of removed labels (x/N) as independent and the estimated percentage of PU cases as the dependent variable.
- The slope of the straight line regression reflects algorithmic bias (no bias = 1) and the intercept corresponds to the true fraction of PU instances.

References

1. Kumar P, Lambert CG. 2024. Positive Unlabeled Learning Selected Not At Random (PULSNAR): class proportion estimation without the selected completely at random assumption. PeerJ Computer Science 10:e2451 https://doi.org/10.7717/peerj-cs.2451
2. Ivanov D. DedpuI: Difference-of-estimated-densities-based positive-unlabeled learning. In2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) 2020 Dec 14 (pp. 782-790). IEEE.
3. Bekker, J., & Davis, J. (2018). Estimating the Class Prior in Positive and Unlabeled Data Through Decision Tree Induction. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11715
4. Madden JM, Lakoma MD, Rusinak D, Lu CY, Soumerai SB. Missing clinical and behavioral health data in a large electronic health record (EHR) system. J Am Med Inform Assoc. 2016 Nov;23(6):1143-1149. doi: 10.1093/jamia/ocw021. Epub 2016 Apr 14. PMID: 27079506; PMCID: PMC5070523
5. Johnson, B. (2013). Wilt [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MS4W.
6. Bock, R. (2007). MAGIC Gamma Telescope. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52C8M.
7. Higuera, C., Gardiner, K., and Cios, K. (2015). Mice Protein Expression. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50S3Z.570

Results

Figure 1: Linear regression plots based on the mouse protein expression dataset for three proportions of positive examples within the unlabeled set: 1%, 10%, and 20%. For each x/N value, the experiment was repeated 10 times with data shuffled using different random seeds.

Figure 2: Linear regression plots based on the magic gamma telescope dataset for three proportions of positive examples within the unlabeled set: 1%, 10%, and 20%. For each x/N value, the experiment was repeated 10 times with data shuffled using different random seeds.

Figure 3: Linear regression plots based on the wilt remote sensing dataset for three proportions of positive examples within the unlabeled set: 1%, 10%, and 20%. For each x/N value, the experiment was repeated 10 times with data shuffled using different random seeds.

Discussion and Conclusions

- The results show that the linear regression approach presented here is well-suited to identify bias in PU learning algorithms, effectively exposing algorithmic bias.
- High R^2 is indicative for a good model performance, the regression slope can then be used to bias correct estimates.
- The method performs well on synthetic and benchmark datasets, where the true labels of all instances are known.
- As expected, overestimation by a PU method yields a slope > 1, while underestimation yields a slope < 1.
- The agreement between theoretical predictions and observations in these datasets suggests that the proposed method is suitable for evaluating the performance of PU learning algorithms on empirical data, where the status of unlabeled instances cannot be directly assessed, as in cases of depression, bipolar disorder, or PTSD.

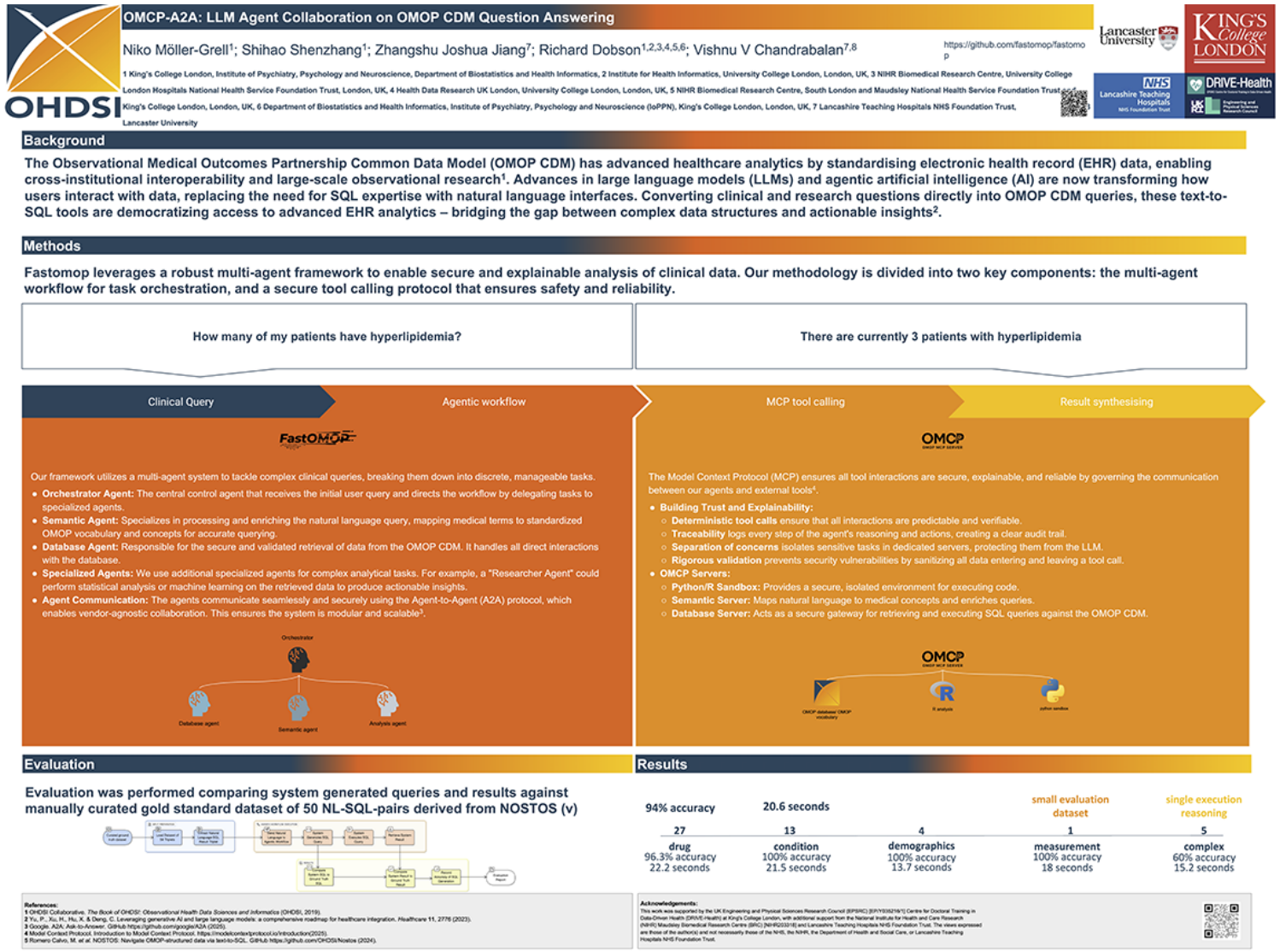
Acknowledgment: This research was supported by funding from the US National Institutes of Health, specifically, the National Institute of Mental Health grant #01MH129764, the National Library of Medicine grant #00LM013367, and infrastructure support from the National Center for Advancing Translational Sciences grant UL1TR001449.

#OHDSISocialShowcase This Week

Friday

Agentic conversation on OMOP CDM: the OMCP-A2A foundation library

(Niko Möller-Grell, Shihao Shenzhang, Zhangshu Joshua Jiang, Vishnu V Chandrabalan, Richard Dobson)



Where Are We Going?

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

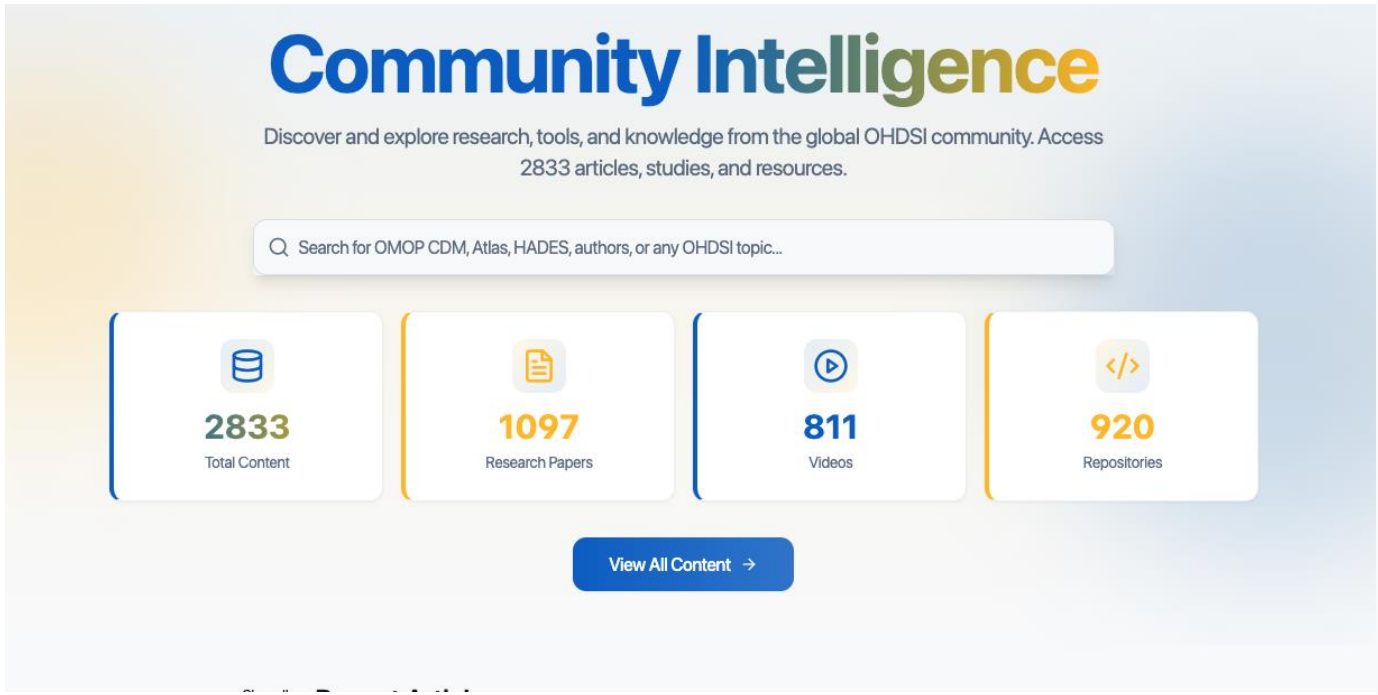
Three Stages of The Journey

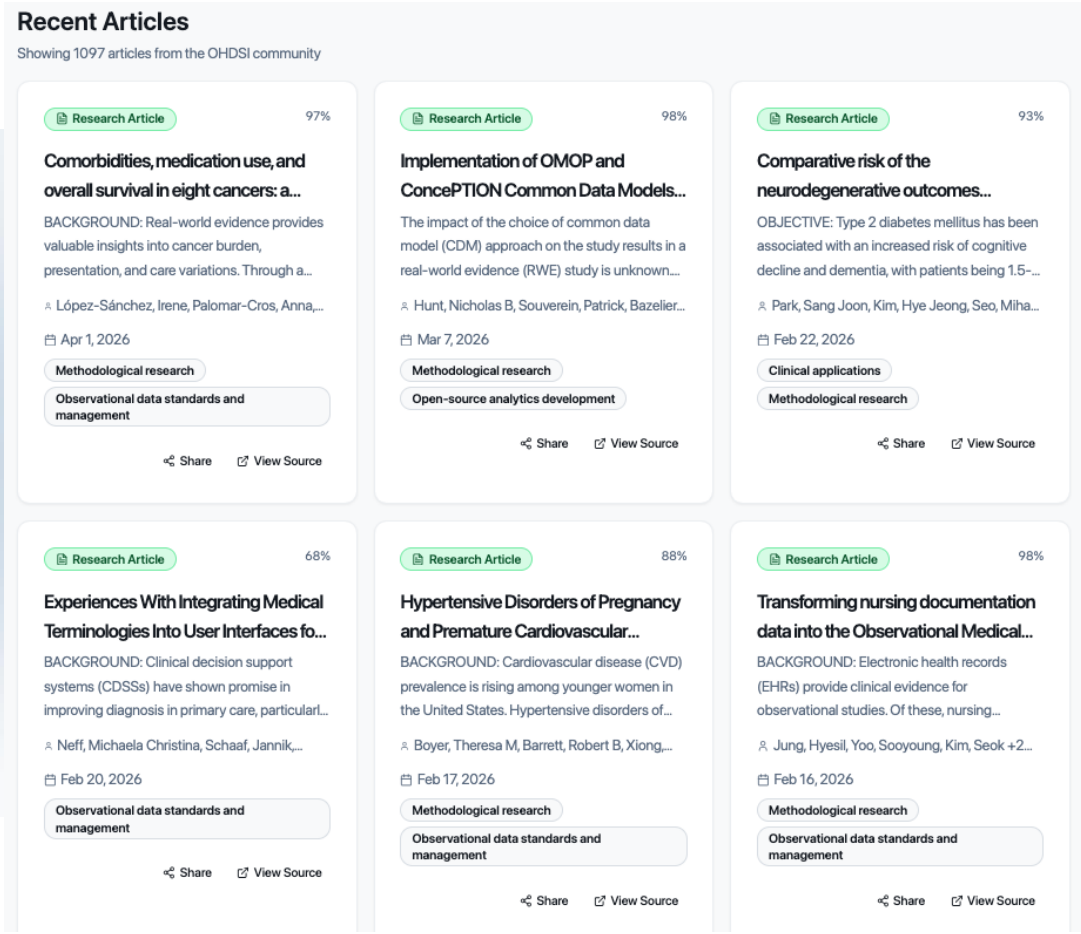
Where Have We Been?

Where Are We Now?

Where Are We Going?

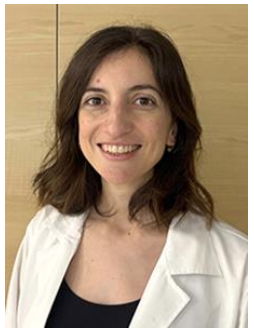
Community Dashboard





dash.ohdsi.org

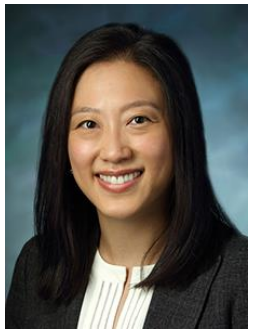
March 24: OHDSI/OMOP Research Spotlight



Berta Cuyàs

Hospital de la Santa Creu i Sant Pau, Autonomous University of Barcelona, CIBERehd

Trends in incidence, prevalence, and survival of primary liver cancer in the United Kingdom (2000–2021)
• *European Journal of Public Health*



Cindy Cai

Johns Hopkins University

Semaglutide and diabetic retinopathy: an OHDSI network study • *BMJ Open Diabetes Research & Care*

Matthew Spotnitz

Department of Health and Human Services, National Institutes of Health

Trends in incidence, prevalence, and survival of primary liver cancer in the United Kingdom (2000–2021)
• *European Journal of Public Health*

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-2026

Find your workgroup.

Fuel our mission.

ohdsi.org/workgroups