2026 Best Community Contribution Honorees

OHDSI Community Call
Nov. 10, 2026 • 11 am ET

Upcoming Community Calls

Date	Topic
Nov. 10	Collaborator Showcase Honoree Presentations
Nov. 17	Early-Stage Researcher Presentations
Nov. 24	Software Demos
Dec. 1	OHDSI/OMOP Research Spotlight
Dec. 8	How Did OHDSI Do This Year?
Dec. 15	Holiday-Themed Fairwell to 2026
Dec. 22	NO CALL – Holiday Break

Early-Stage Researcher Contribution

ESR Rising Star Award

Celebrating three early-stage researchers and their award-winning work

Hsin Yi “Cindy” Chen

Aniek Markus

Nicholas Solages



Three Stages of The Journey

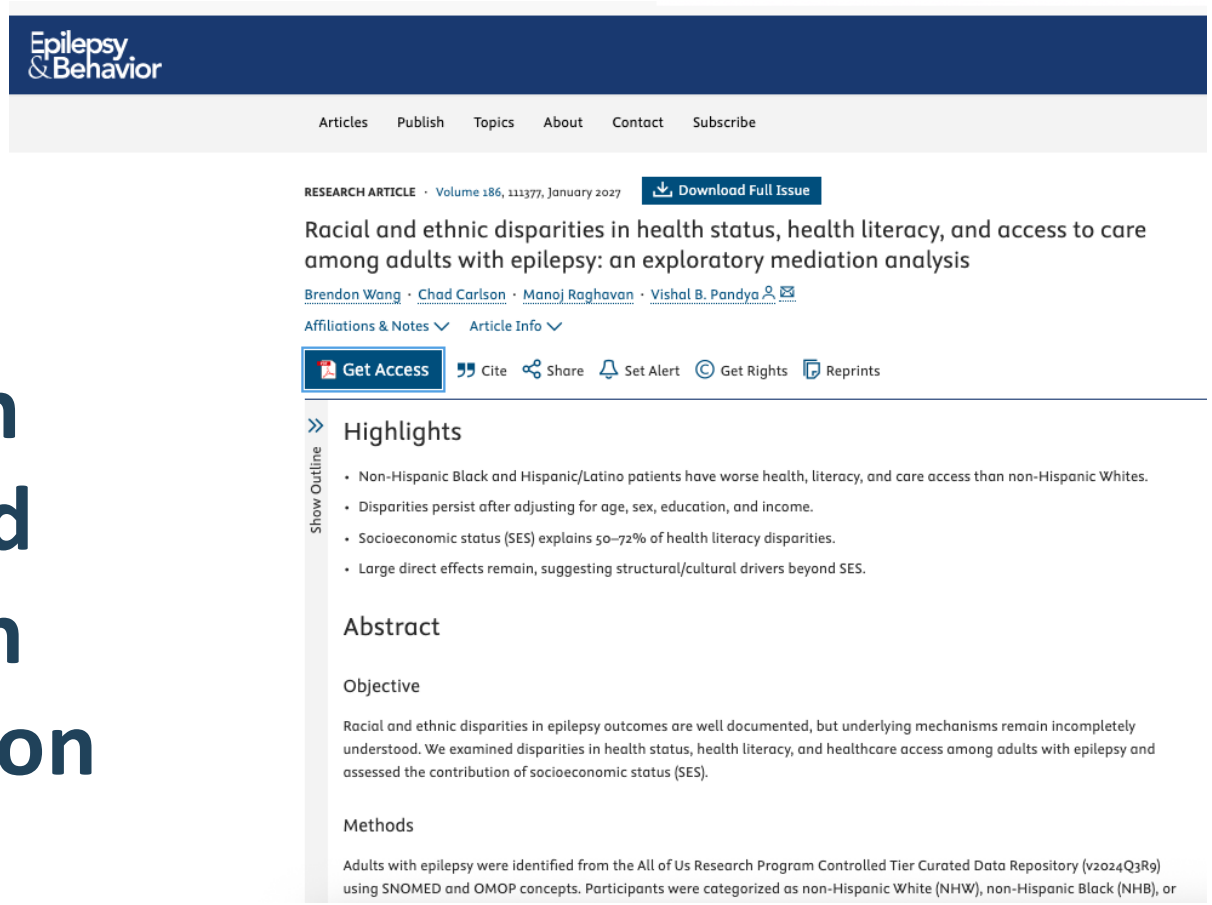
Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Brendon Wang, Chad Carlson, Manoj Raghavan, and Vishal B Pandya** on the recent publication of **Racial and ethnic disparities in health status, health literacy, and access to care among adults with epilepsy: an exploratory mediation analysis** in *Epilepsy & Behavior*.



Epilepsy & Behavior

Articles Publish Topics About Contact Subscribe

RESEARCH ARTICLE · Volume 186, 111377, January 2027 Download Full Issue

Racial and ethnic disparities in health status, health literacy, and access to care among adults with epilepsy: an exploratory mediation analysis

Brendon Wang · Chad Carlson · Manoj Raghavan · Vishal B. Pandya

Affiliations & Notes Article Info

Get Access Cite Share Set Alert Get Rights Reprints

Show Outline

Highlights

- Non-Hispanic Black and Hispanic/Latino patients have worse health, literacy, and care access than non-Hispanic Whites.
- Disparities persist after adjusting for age, sex, education, and income.
- Socioeconomic status (SES) explains 50–72% of health literacy disparities.
- Large direct effects remain, suggesting structural/cultural drivers beyond SES.

Abstract

Objective

Racial and ethnic disparities in epilepsy outcomes are well documented, but underlying mechanisms remain incompletely understood. We examined disparities in health status, health literacy, and healthcare access among adults with epilepsy and assessed the contribution of socioeconomic status (SES).

Methods

Adults with epilepsy were identified from the All of Us Research Program Controlled Tier Curated Data Repository (v2024Q3R9) using SNOMED and OMOP concepts. Participants were categorized as non-Hispanic White (NHW), non-Hispanic Black (NHB), or

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	Generative AI and Analytics
Wednesday	9 am	Patient-Level Prediction
Wednesday	2 pm	Natural Language Processing
Wednesday	7 pm	Medical Imaging
Wednesday	7 pm	Eyecare and Vision Research
Thursday	7 am	Europe Community Call
Thursday	10 am	GIS – Geographic Information System
Thursday	10 am	Rare Diseases
Thursday	7 pm	Dentistry
Friday	10 am	Transplant
Friday	11 am	Clinical Trials
Friday	11:30 am	Steering
Friday	11 pm	China Chapter
Tuesday	9 am	Data2Evidence

APAC Symposium: Nov. 13-15

The 2026 APAC Symposium will be held Nov. 13-15 at Yonsei University in Seoul, South Korea.

Nov. 13: Main Conference
Nov. 14: Tutorials, Datathon
Nov. 15: Datathon Team Activities

Registration is open! All details can be found on the event homepage.



ohdsi.org/APAC2026

Spotlight: Jiwon Um

Jiwon Um serves as the APAC Community Coordinator, leading the monthly APAC Community Calls. She also contributes to network studies and is part of the organizing team for the 2026 APAC Symposium.

Jiwon discusses how OHDSI has impacted her educational journey, child and adolescent development research, OHDSI and its relationship with medical AI, and more in the latest Collaborator Spotlight.



ohdsi.org/spotlight-jiwon-um

Canada Symposium: Nov. 17-18

Registration is now open for the 2026 OHDSI Canada Symposium, held Nov. 17-18 in Toronto, Ont.

Nov. 17: Workshops

Track A: From Source Data to OMOP

Track B: From OMOP to Evidence

Nov. 18: Symposium

The main symposium day will bring the Canadian OHDSI community together to share experiences, showcase emerging work, and explore opportunities for collaboration.



India Symposium: Dec. 11



OHDSI INDIA

WHY TO ATTEND?

- **UNDERSTAND INDIA'S HEALTHCARE DATA LANDSCAPE**
  Learn about the current state of healthcare data in India, emerging national priorities, and how global standards are shaping the future of healthcare and research.
- **EXPLORE OMOP AND HEALTHCARE DATA INITIATIVES**
  Discover the latest developments in OMOP, Real-World Evidence (RWE), FHIR, AI, and other data standardization initiatives being implemented across India.
- **LEARN HOW YOUR ORGANIZATION CAN PARTICIPATE**
  Understand how hospitals, medical colleges, and healthcare organizations can contribute to national data initiatives, multicenter research, training programs, and collaborative projects through OHDSI India.
- **LEARN FROM EXPERTS**
  Gain insights from national and international experts on healthcare data standards, OMOP, FHIR, Real-World Evidence, AI, and observational research.
- **NETWORK WITH THE HEALTHCARE COMMUNITY**
  Connect with clinicians, researchers, hospitals, academia, industry partners, and global OHDSI leaders to build meaningful collaborations
- **SHOWCASE YOURWORK**
  SHOWCASE YOURWORK Present your research through posters, lightning talks, or case studies and receive valuable feedback from the community
- **BUILD FUTURE COLLABORATIONS**
  Identify opportunities to participate in multicenter research, training programs, industry collaborations, and Real-World Evidence projects.

Upcoming Events Page

What We Do | Engage With OHDSI | OMOP CDM | Events | Community Forums

OHDSI Events

From events held in U.S., Canada, Latin America, Europe, Africa, Asia, or Australia, our annual symposia are among the most anticipated events of the year. Plenary talks, tutorials, workshops and the collaborator showcase create opportunities to share research, learn from each other, and collaborate on future opportunities.

These events also allow our global community an opportunity to connect face-to-face. Whether it is to share scientific breakthroughs, build massive Lego sets or dance on stage together, each symposium brings our community together and sets new paths for collaboration.

Upcoming Events

APAC Symposium

The 2026 APAC Symposium takes place Nov. 13-15 in Yonsei, Korea. Registration is open.

Event Homepage

Canada Symposium

The 2026 Canada Symposium takes place Nov. 17-18 in Toronto, Ont. Registration is open.

Event Homepage

India Symposium

The 2026 India Symposium takes place Dec. 11 in Chennai. Registration is open and limited to 60 seats.

Register Here

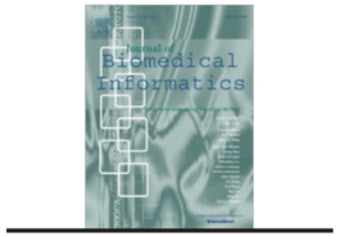
Europe Symposium

The 2027 Europe Symposium is scheduled for April 26 in Helsinki, Finland. More information to come.

OHDSI Europe

Call for Papers: Agentic AI for RWE Life Cycle Support

The Journal of Biomedical Informatics has opened a call for papers focused on Agentic AI For RWE Life-Cycle Support — scaling trustworthy evaluation for treatment safety and effectiveness.



Journal of Biomedical Informatics
Date: Available online 8 June 2026
Article: 105062
In Press, Journal Pre-proof

Published by
ELSEVIER

Editorial • Full text access Get rights and content ↗

Editorial

Call for papers: agentic AI for real-world evidence life cycle support – scaling trustworthy evaluation of treatment safety and effectiveness

Show more

View PDF Cite Add to Mendeley Share 10.1016/j.jbi.2026.105062 ↗

Article

Guest editors:

Richard Boyce (Contact) rdb20@pitt.edu; Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA.

Scott Alexander Malec, smalec@salud.unm.edu; Translational Informatics Division, School of Medicine, University of New Mexico, Albuquerque, NM, USA.

Chunhua Weng. cw2384@cumc.columbia.edu; Biomedical Informatics, Data Science Institute, Columbia University, New York, NY, USA.

Outline

Declaration of competing interest

References

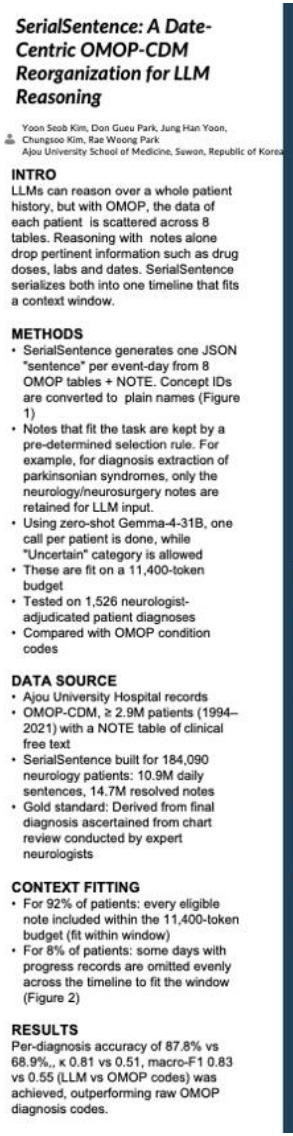
Submissions will be accepted from June 15 through **November 15, 2026,** with expected publication date of September 15, 2027.

#OHDSISocialShowcase This Week

Monday

SerialSentence: A Date-Centric OMOP-CDM Reorganization for LLM Reasoning

(Yoon Seob Kim, Don Gueu Park, Jung Han Yoon, Chungsoo Kim, Rae Woong Park)



SerialSentence: A Date-Centric OMOP-CDM Reorganization for LLM Reasoning

Yoon Seob Kim, Don Gueu Park, Jung Han Yoon, Chungsoo Kim, Rae Woong Park
Ajou University School of Medicine, Suwon, Republic of Korea

INTRO
LLMs can reason over a whole patient history, but with OMOP, the data of each patient is scattered across 8 tables. Reasoning with notes alone drop pertinent information such as drug doses, labs and dates. SerialSentence serializes both into one timeline that fits a context window.

METHODS
- SerialSentence generates one JSON "sentence" per event-day from 8 OMOP tables + NOTE. Concept IDs are converted to plain names (Figure 1)
- Notes that fit the task are kept by a pre-determined selection rule. For example, for diagnosis extraction of parkinsonian syndromes, only the neurology/neurosurgery notes are retained for LLM input.
- Using zero-shot Gemma-4-31B, one call per patient is done, while "Uncertain" category is allowed
- These are fit on a 11,400-token budget
- Tested on 1,526 neurologist-adjudicated patient diagnoses
- Compared with OMOP condition codes

DATA SOURCE
- Ajou University Hospital records
- OMOP-CDM, ≥ 2.9M patients (1994–2021) with a NOTE table of clinical free text
- SerialSentence built for 184,090 neurology patients: 10.9M daily sentences, 14.7M resolved notes
- Gold standard: Derived from final diagnosis ascertained from chart review conducted by expert neurologists

CONTEXT FITTING
- For 92% of patients: every eligible note included within the 11,400-token budget (fit within window)
- For 8% of patients: some days with progress records are omitted evenly across the timeline to fit the window (Figure 2)

RESULTS
Per-diagnosis accuracy of 87.8% vs 68.9%,, κ 0.81 vs 0.51, macro-F1 0.83 vs 0.55 (LLM vs OMOP codes) was achieved, outperforming raw OMOP diagnosis codes.

Serialize the whole OMOP record by date — and a zero-shot LLM recovers the diagnoses that condition codes miss.

87.8% vs 68.9%
primary-diagnosis accuracy · LLM vs OMOP codes

+18.9 pp
paired Δ [16.4–21.6] · McNemar p = 4.2×10⁻⁴⁷

κ 0.81 vs 0.51
agreement with neurologist adjudication



Figure 1: SerialSentence: eight OMOP tables converted to one JSON sentence per event-day



Figure 2: Note selection and cleanup, guaranteeing whole-document fit

Scan for the brief report (PDF)

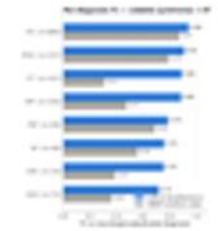


Figure 3: Per-class accuracy by parkinsonism related syndromes and essential tremor
PD: Parkinson's disease, MSA: multiple cyctem atophy, ET: essential tremor, DIP: drug induced parkinsonism, PSP: progressive supranuclear palsy, VP: vascular parkinsonism, CBS: corticobasal syndrome, DLB: demetia with Lewy bodies

LLM certainty	n (%)	Acc. / F1	FP	FN
All LLM	1,526	87.8% / 0.83	9.8%	12.2%
high	1,113 (73%)	96.2% / 0.92	3.8%	3.8%
medium	377 (25%)	71.4% / 0.72	28.6%	28.6%
low (abstained)	36 (2.4%)	—	0%	100%
OMOP codes	1,526	68.9% / 0.55	18.9%	31.1%

Table 1: Accuracy by LLM certainty (FP = incorrect diagnosis; FN = incorrect or abstained)

TIERED USE
High-certainty predictions (73%, 96.2% accurate) may be deployed as-is, while medium-or-lower may be deferred to human review as capacity allows. The model over-calls the diagnoses rather than abstains when uncertain (9.8% incorrect vs 2.4% abstained).

WHERE CODES FALL SHORT
OMOP codes are precise but under-recall (PSP P 0.93 / R 0.55; ET R 0.20) and leave 13% uncoded — the pipeline backfills such uncoded cases

CONCLUSION
SerialSentence bridges OMOP-CDM to whole-timeline LLM input by structuring notes and tabular data into a coherent input. With a zero-shot LLM it may recover diagnoses that ETL processes miss, and its self-rated certainty matches the measured accuracy by tier. This is done on single-site, single-specialty, diagnosis-extraction only so far, so extension as well as external replication is needed. All inference ran on a locally served open-weights model and therefore no PHI left the site.

PRESENTER:
Yoon Seob Kim

#OHDSISocialShowcase This Week

Tuesday

Analysis of Aggregate Incidence Rate Data in Real-World Evidence Studies

(Guy Cafri, Paul Coplan, Martijn Schuemie)

Summarizing Aggregate Incidence Rates Across Data Sources: A Simulation Study of Meta-Analysis and Poisson Regression Methods

Guy Cafri , Paul Coplan, Martijn Schuemie

Global Epidemiology Organization, Johnson and Johnson, Titusville, New Jersey

Background

- Incidence rates are widely used to characterize the risk of adverse events after drug exposure
- Comparative estimates are often preferred, but are not always feasible
- When several data sources are available, it is of interest to summarize incidence rates numerically
- Rare events make this difficult — individual data sources may have zero events
- Distinguish between two estimands
  - Marginal rate — the average incidence rate across data sources.
  - Conditional rate — the rate when the data-source random effect is zero; the median or typical study-specific rate.
- Use Monte Carlo simulation to determine how best to summarize incidence rates numerically across data sources under three challenges: few data sources, non-trivial between-source variability, and rare events.

Methods

Monte Carlo Simulation Design

- Data generated from a Poisson model with data-source random effects; 1,000 simulated datasets per condition.
- Number of data sources: 5 or 10.
- Conditional incidence rate: 0.01, 0.001, or 0.0001 events per year.
- SD of the random effect: 0 or 1. With SD = 1 the marginal rates are 0.016490, 0.001650, and 0.000160.
- Exposure times fixed per data source: 500, 1,000, 2,000, 3,000 and 4,000 years (repeated for 10 sources).

Six Estimation Methods

- Pooled events divided by pooled person-time (Pooled model)
- Frequentist fixed-effect Poisson regression
- Frequentist random-effects Poisson regression
- Frequentist random-effects meta-analysis
- Bayesian random-effects meta-analysis
- Bayesian random-effects Poisson regression

Outcomes

- Percent relative bias of the incidence rate, absolute bias of the between-source SD, and coverage of 95% confidence or credible intervals.

Results

Marginal Incidence Rate

- Calculate using the intercept in a model with no random effects or in a random-effects model: $exp(\hat{a}+0.5*\hat{\tau}^2)$
- Least biased point estimates came from pooling, frequentist fixed-effect Poisson regression, and Bayesian random-effects Poisson regression using the posterior median (Figure 1).
- Other random-effects approaches showed large positive bias, driven by unstable, right-skewed estimates of the between-source SD combined with the convexity of the exponential map (Jensen's inequality).

Conditional Incidence Rate

- Frequentist and Bayesian random-effects Poisson regression had the least bias across all conditions (Figure 2)
- For both marginal and conditional incidence rate same pattern with 10 data sources (not displayed)

Results

Figure 1. Percent Relative Bias of Marginal Estimates for the Incidence Rate by Method and Standard Deviation of the Random Effects (5 Data Sources)

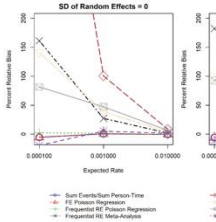
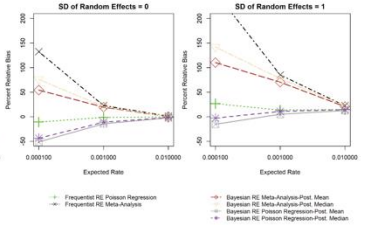


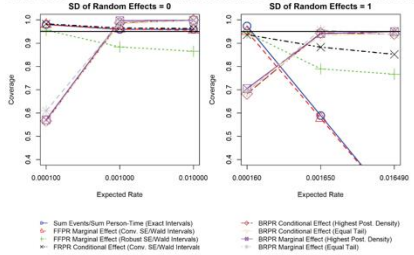
Figure 2. Percent Relative Bias of Conditional Estimates for the Incidence Rate by Method and Standard Deviation of the Random Effects (5 Data Sources)



Coverage

- Pooling, frequentist fixed-effect or random-effects Poisson regression interval estimates for the marginal or conditional incidence rates were below nominal coverage when $\tau = 1$ and the rate was at one of the two larger values, but close to nominal otherwise.
- Bayesian estimates of either the marginal or conditional rate from a Poisson regression only fell below nominal coverage at the rarest population rate simulated, and close to nominal otherwise.
- Bayesian and robust standard errors improved coverage with an increasing number of data sources (ie, 10-not displayed)

Figure 7. Coverage for Methods Demonstrating the Least Bias for Estimating Marginal and Conditional Incidence Rates (5 Data Sources)



Note: Solid black line represents nominal coverage level. FFPR=Frequentist Fixed-effect Poisson Regression. BRPR=Bayesian Random-effects Poisson Regression. FRPR=Frequentist Random-effects Poisson Regression.

Conclusions

Marginal Rate

- Point Estimation : pooling/frequentist fixed-effect Poisson regression or Bayesian random-effects Poisson regression models
- Interval Estimation: If rare, pooled data with exact confidence intervals or a fixed-effect Poisson regression with robust standard errors and Wald intervals. If not as rare, then a Bayesian random-effect Poisson regression with equal-tail or highest density posterior credibility intervals.

Conditional Rate

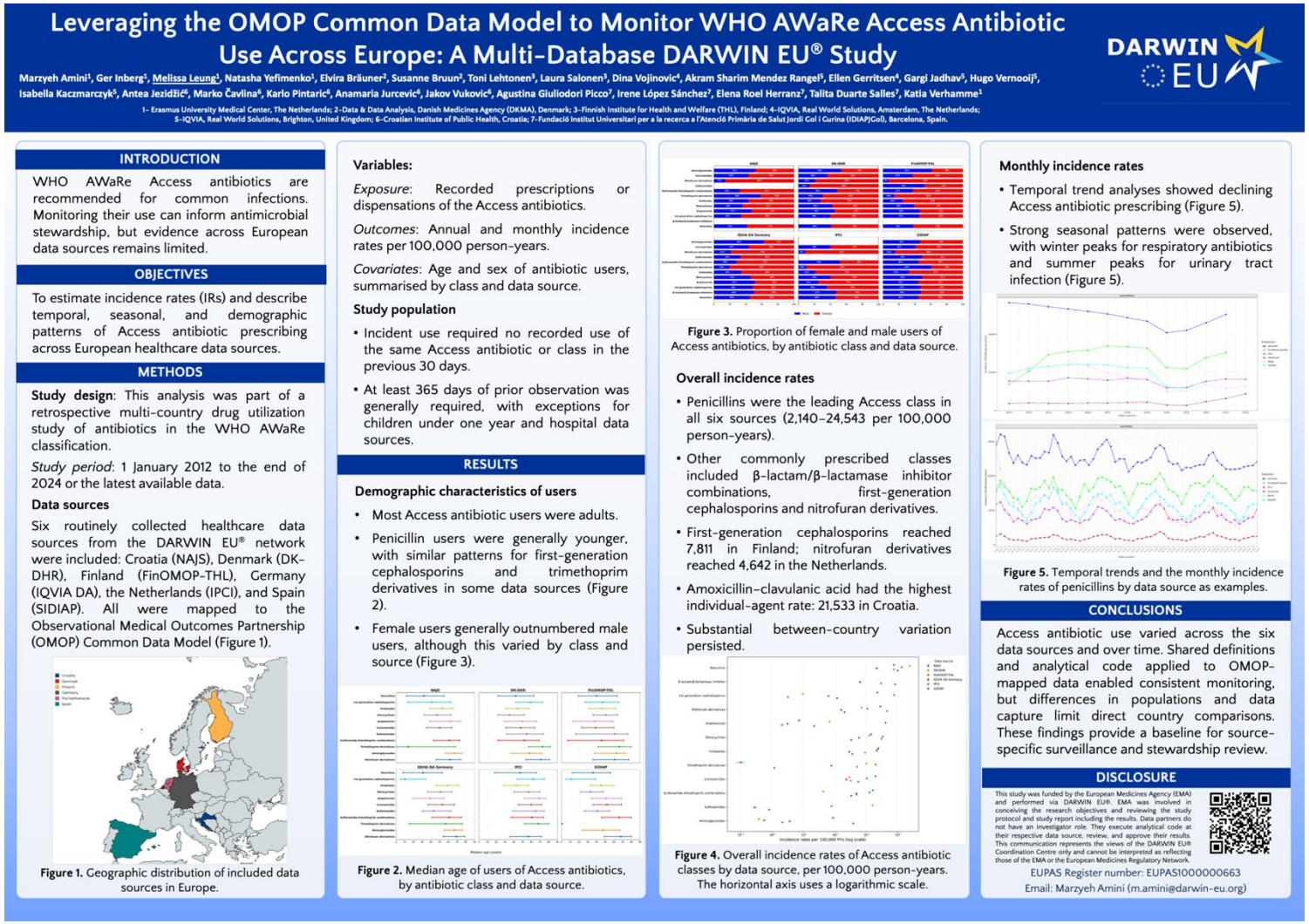
- Point Estimation : Random-effects Poisson regression (frequentist or Bayesian)
- Interval Estimation: If rare, frequentist random-effect Poisson regression with conventional standard errors and Wald intervals. If not as rare, then a Bayesian random-effect Poisson regression with equal-tail or highest density posterior credibility intervals.

#OHDSISocialShowcase This Week

Wednesday

Leveraging the OMOP Common Data Model to Monitor WHO AWaRe Access Antibiotic Use Across Europe: A Multi-Database DARWIN EU® Study

(Marzyeh Amini, Ger Inberg, Melissa Leung, Elvira Bräuner, Susanne Bruun, Toni Lehtonen, Laura Salonen, Dina Vojinovic, Akram Sharim Mendez Rangel, Ellen Gerritsen, Gargi Jadhav, Hugo Vernooij, Isabella Kaczmarczyk, Antea Jezidžić, Marko Čavlina, Karlo Pintaric, Anamaria Jurcevic, Jakov Vukovic, Agustina Giuliodori Picco, Irene López Sánchez, Elena Roel Herranz, Talita Duarte Salles, Katia Verhamme)



Leveraging the OMOP Common Data Model to Monitor WHO AWaRe Access Antibiotic Use Across Europe: A Multi-Database DARWIN EU® Study

Marzyeh Amini[1], Ger Inberg[1], Melissa Leung[1], Natasha Yefimenko[1], Elvira Bräuner[2], Susanne Bruun[2], Toni Lehtonen[3], Laura Salonen[3], Dina Vojinovic[4], Akram Sharim Mendez Rangel[5], Ellen Gerritsen[4], Gargi Jadhav[5], Hugo Vernooij[5], Isabella Kaczmarczyk[5], Antea Jezidžić[6], Marko Čavlina[6], Karlo Pintaric[6], Anamaria Jurcevic[6], Jakov Vukovic[6], Agustina Giuliodori Picco[7], Irene López Sánchez[7], Elena Roel Herranz[7], Talita Duarte Salles[7], Katia Verhamme[1]

1- Erasmus University Medical Center, The Netherlands; 2-Data & Data Analysis, Danish Medicines Agency (DKMA), Denmark; 3-Finnish Institute for Health and Welfare (THL), Finland; 4-IQVIA, Real World Solutions, Amsterdam, The Netherlands; 5-IQVIA, Real World Solutions, Brighton, United Kingdom; 6-Croatian Institute of Public Health, Croatia; 7-Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain.

DARWIN EU

INTRODUCTION

WHO AWaRe Access antibiotics are recommended for common infections. Monitoring their use can inform antimicrobial stewardship, but evidence across European data sources remains limited.

OBJECTIVES

To estimate incidence rates (IRs) and describe temporal, seasonal, and demographic patterns of Access antibiotic prescribing across European healthcare data sources.

METHODS

Study design: This analysis was part of a retrospective multi-country drug utilization study of antibiotics in the WHO AWaRe classification.

Study period: 1 January 2012 to the end of 2024 or the latest available data.

Data sources

Six routinely collected healthcare data sources from the DARWIN EU® network were included: Croatia (NAJS), Denmark (DK-DHR), Finland (FinOMOP-THL), Germany (IQVIA DA), the Netherlands (IPCI), and Spain (SIDIAP). All were mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (Figure 1).

Figure 1. Geographic distribution of included data sources in Europe.

Variables:

Exposure: Recorded prescriptions or dispensations of the Access antibiotics.

Outcomes: Annual and monthly incidence rates per 100,000 person-years.

Covariates: Age and sex of antibiotic users, summarised by class and data source.

Study population

- Incident use required no recorded use of the same Access antibiotic or class in the previous 30 days.
- At least 365 days of prior observation was generally required, with exceptions for children under one year and hospital data sources.

RESULTS

Demographic characteristics of users

- Most Access antibiotic users were adults.
- Penicillin users were generally younger, with similar patterns for first-generation cephalosporins and trimethoprim derivatives in some data sources (Figure 2).
- Female users generally outnumbered male users, although this varied by class and source (Figure 3).

Figure 2. Median age of users of Access antibiotics, by antibiotic class and data source.

Figure 3. Proportion of female and male users of Access antibiotics, by antibiotic class and data source.

Overall incidence rates

- Penicillins were the leading Access class in all six sources (2,140–24,543 per 100,000 person-years).
- Other commonly prescribed classes included β-lactam/β-lactamase inhibitor combinations, first-generation cephalosporins and nitrofuran derivatives.
- First-generation cephalosporins reached 7,811 in Finland; nitrofuran derivatives reached 4,642 in the Netherlands.
- Amoxicillin-clavulanic acid had the highest individual-agent rate: 21,533 in Croatia.
- Substantial between-country variation persisted.

Figure 4. Overall incidence rates of Access antibiotic classes by data source, per 100,000 person-years. The horizontal axis uses a logarithmic scale.

Monthly incidence rates

- Temporal trend analyses showed declining Access antibiotic prescribing (Figure 5).
- Strong seasonal patterns were observed, with winter peaks for respiratory antibiotics and summer peaks for urinary tract infection (Figure 5).

Figure 5. Temporal trends and the monthly incidence rates of penicillins by data source as examples.

CONCLUSIONS

Access antibiotic use varied across the six data sources and over time. Shared definitions and analytical code applied to OMOP-mapped data enabled consistent monitoring, but differences in populations and data capture limit direct country comparisons. These findings provide a baseline for source-specific surveillance and stewardship review.

DISCLOSURE

This study was funded by the European Medicines Agency (EMA) and performed via DARWIN EU®. EMA was involved in conceiving the research objectives and reviewing the study protocol and study report including the results. Data partners do not have an investigator role. They execute analytical code at their respective data source, review, and approve their results. This communication represents the views of the DARWIN EU® Coordination Centre only and cannot be interpreted as reflecting those of the EMA or the European Medicines Regulatory Network.

EUPAS Register number: EUPAS1000000663

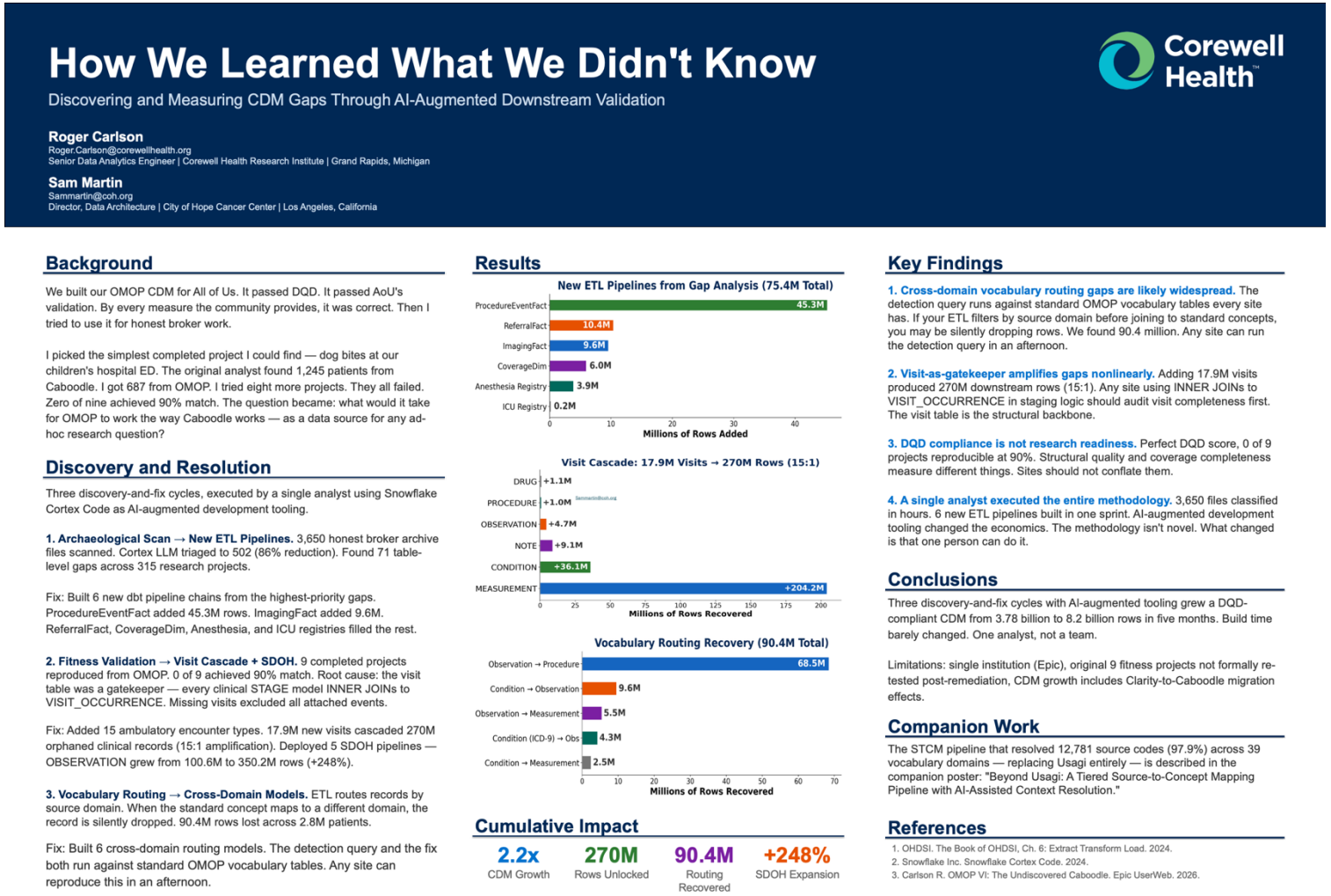
Email: Marzyeh Amini (m.amini@darwin-eu.org)

#OHDSISocialShowcase This Week

Thursday

How AI Helped Us Learn What We Didn't Know: Discovering CDM Gaps Through AI-Augmented Downstream Validation

(Roger Carlson, Sam Martin)



How We Learned What We Didn't Know

Discovering and Measuring CDM Gaps Through AI-Augmented Downstream Validation

Roger Carlson
Roger.Carlson@corewellhealth.org
Senior Data Analytics Engineer | Corewell Health Research Institute | Grand Rapids, Michigan

Sam Martin
Sammartin@coh.org
Director, Data Architecture | City of Hope Cancer Center | Los Angeles, California

Corewell Health

Background

We built our OMOP CDM for All of Us. It passed DQD. It passed AoU's validation. By every measure the community provides, it was correct. Then I tried to use it for honest broker work.

I picked the simplest completed project I could find — dog bites at our children's hospital ED. The original analyst found 1,245 patients from Caboodle. I got 687 from OMOP. I tried eight more projects. They all failed. Zero of nine achieved 90% match. The question became: what would it take for OMOP to work the way Caboodle works — as a data source for any ad-hoc research question?

Discovery and Resolution

Three discovery-and-fix cycles, executed by a single analyst using Snowflake Cortex Code as AI-augmented development tooling.

1. Archaeological Scan → New ETL Pipelines. 3,650 honest broker archive files scanned. Cortex LLM triaged to 502 (86% reduction). Found 71 table-level gaps across 315 research projects.

Fix: Built 6 new dbt pipeline chains from the highest-priority gaps. ProcedureEventFact added 45.3M rows. ImagingFact added 9.6M. ReferralFact, CoverageDim, Anesthesia, and ICU registries filled the rest.

2. Fitness Validation → Visit Cascade + SDOH. 9 completed projects reproduced from OMOP. 0 of 9 achieved 90% match. Root cause: the visit table was a gatekeeper — every clinical STAGE model INNER JOINs to VISIT_OCCURRENCE. Missing visits excluded all attached events.

Fix: Added 15 ambulatory encounter types. 17.9M new visits cascaded 270M orphaned clinical records (15:1 amplification). Deployed 5 SDOH pipelines — OBSERVATION grew from 100.6M to 350.2M rows (+248%).

3. Vocabulary Routing → Cross-Domain Models. ETL routes records by source domain. When the standard concept maps to a different domain, the record is silently dropped. 90.4M rows lost across 2.8M patients.

Fix: Built 6 cross-domain routing models. The detection query and the fix both run against standard OMOP vocabulary tables. Any site can reproduce this in an afternoon.

Results

Cumulative Impact

2.2x	270M	90.4M	+248%
CDM Growth	Rows Unlocked	Routing Recovered	SDOH Expansion

Key Findings

1. Cross-domain vocabulary routing gaps are likely widespread. The detection query runs against standard OMOP vocabulary tables every site has. If your ETL filters by source domain before joining to standard concepts, you may be silently dropping rows. We found 90.4 million. Any site can run the detection query in an afternoon.

2. Visit-as-gatekeeper amplifies gaps nonlinearly. Adding 17.9M visits produced 270M downstream rows (15:1). Any site using INNER JOINs to VISIT_OCCURRENCE in staging logic should audit visit completeness first. The visit table is the structural backbone.

3. DQD compliance is not research readiness. Perfect DQD score, 0 of 9 projects reproducible at 90%. Structural quality and coverage completeness measure different things. Sites should not conflate them.

4. A single analyst executed the entire methodology. 3,650 files classified in hours. 6 new ETL pipelines built in one sprint. AI-augmented development tooling changed the economics. The methodology isn't novel. What changed is that one person can do it.

Conclusions

Three discovery-and-fix cycles with AI-augmented tooling grew a DQD-compliant CDM from 3.78 billion to 8.2 billion rows in five months. Build time barely changed. One analyst, not a team.

Limitations: single institution (Epic), original 9 fitness projects not formally re-tested post-remediation, CDM growth includes Clarity-to-Caboodle migration effects.

Companion Work

The STCM pipeline that resolved 12,781 source codes (97.9%) across 39 vocabulary domains — replacing Usagi entirely — is described in the companion poster: "Beyond Usagi: A Tiered Source-to-Concept Mapping Pipeline with AI-Assisted Context Resolution."

References

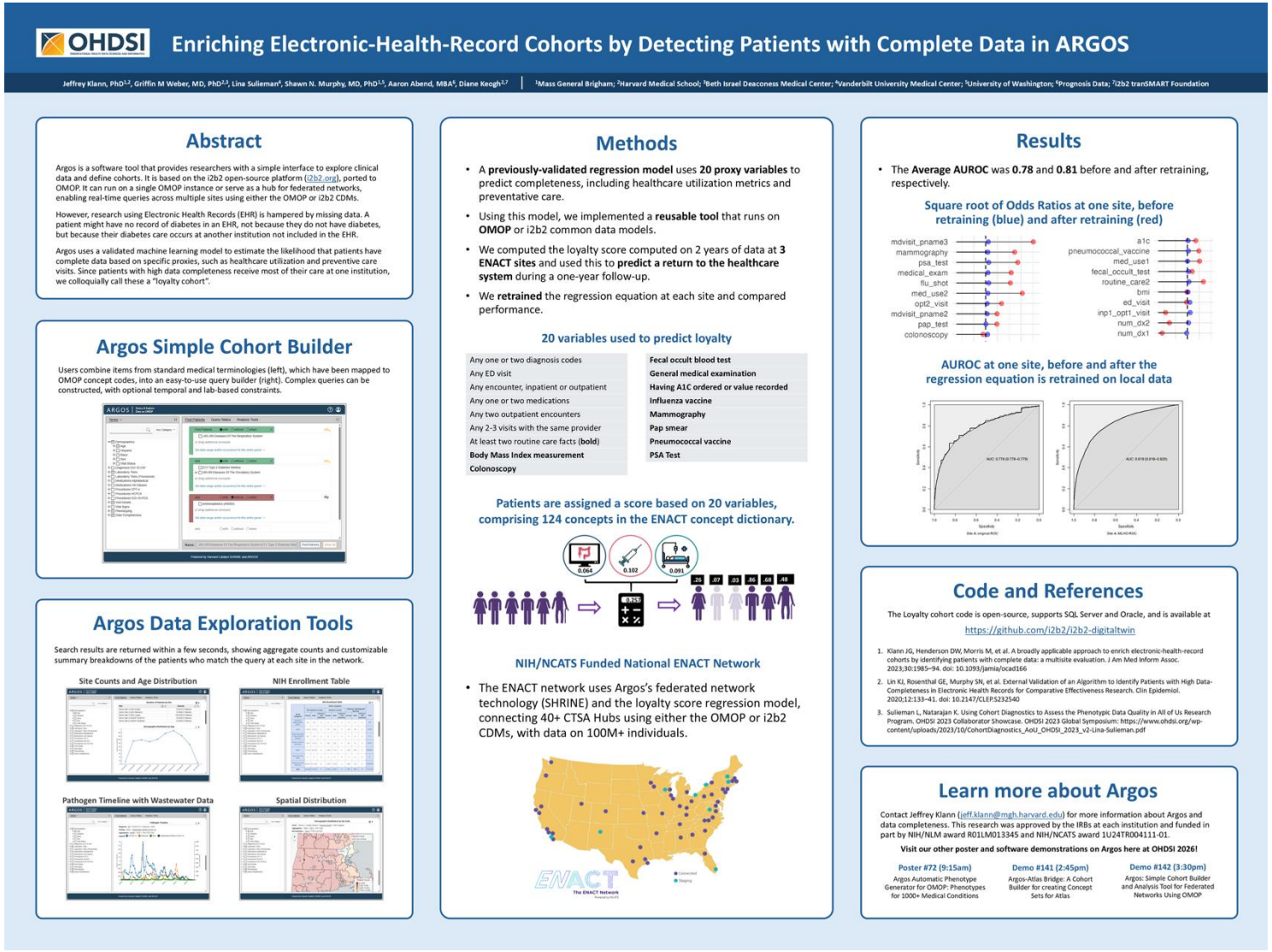
1. OHDSI. The Book of OHDSI, Ch. 6: Extract Transform Load. 2024.
2. Snowflake Inc. Snowflake Cortex Code. 2024.
3. Carlson R. OMOP VI: The Undiscovered Caboodle. Epic UserWeb. 2026.

#OHDSISocialShowcase This Week

Friday

Enriching Electronic-Health-Record Cohorts by Detecting Patients with Complete Data in Argos

(Jeffrey Klann, Griffin M Weber, Lina Sulieman, Shawn N. Murphy, Aaron Abend, Diane Keogh)



OHDSI

Enriching Electronic-Health-Record Cohorts by Detecting Patients with Complete Data in ARGOS

Jeffrey Klann, PhD[1,2], Griffin M Weber, MD, PhD[2,3], Lina Sulieman[4], Shawn N. Murphy, MD, PhD[1,5], Aaron Abend, MBA[6], Diane Keogh[2,7] | [1]Mass General Brigham; [2]Harvard Medical School; [3]Beth Israel Deaconess Medical Center; [4]Vanderbilt University Medical Center; [5]University of Washington; [6]Prognosis Data; [7]i2b2 tranSMART Foundation

Abstract

Argos is a software tool that provides researchers with a simple interface to explore clinical data and define cohorts. It is based on the i2b2 open-source platform (i2b2.org), ported to OMOP. It can run on a single OMOP instance or serve as a hub for federated networks, enabling real-time queries across multiple sites using either the OMOP or i2b2 CDMs.

However, research using Electronic Health Records (EHR) is hampered by missing data. A patient might have no record of diabetes in an EHR, not because they do not have diabetes, but because their diabetes care occurs at another institution not included in the EHR.

Argos uses a validated machine learning model to estimate the likelihood that patients have complete data based on specific proxies, such as healthcare utilization and preventive care visits. Since patients with high data completeness receive most of their care at one institution, we colloquially call these a "loyalty cohort".

Argos Simple Cohort Builder

Users combine items from standard medical terminologies (left), which have been mapped to OMOP concept codes, into an easy-to-use query builder (right). Complex queries can be constructed, with optional temporal and lab-based constraints.

Argos Data Exploration Tools

Search results are returned within a few seconds, showing aggregate counts and customizable summary breakdowns of the patients who match the query at each site in the network.

Site Counts and Age Distribution

NIH Enrollment Table

Pathogen Timeline with Wastewater Data

Spatial Distribution

Methods

- A **previously-validated regression model** uses **20 proxy variables** to predict completeness, including healthcare utilization metrics and preventative care.
- Using this model, we implemented a **reusable tool** that runs on **OMOP** or i2b2 common data models.
- We computed the loyalty score computed on 2 years of data at **3 ENACT sites** and used this to **predict a return to the healthcare system** during a one-year follow-up.
- We **retrained** the regression equation at each site and compared performance.

20 variables used to predict loyalty

Any one or two diagnosis codes	**Fecal occult blood test**
Any ED visit	**General medical examination**
Any encounter, inpatient or outpatient	**Having A1C ordered or value recorded**
Any one or two medications	**Influenza vaccine**
Any two outpatient encounters	**Mammography**
Any 2-3 visits with the same provider	**Pap smear**
At least two routine care facts (**bold**)	**Pneumococcal vaccine**
Body Mass Index measurement	**PSA Test**
Colonoscopy	

Patients are assigned a score based on 20 variables, comprising 124 concepts in the ENACT concept dictionary.

NIH/NCATS Funded National ENACT Network

- The ENACT network uses Argos's federated network technology (SHRINE) and the loyalty score regression model, connecting 40+ CTSA Hubs using either the OMOP or i2b2 CDMs, with data on 100M+ individuals.

Results

- The **Average AUROC** was **0.78** and **0.81** before and after retraining, respectively.

Square root of Odds Ratios at one site, before retraining (blue) and after retraining (red)

AUROC at one site, before and after the regression equation is retrained on local data

Code and References

The Loyalty cohort code is open-source, supports SQL Server and Oracle, and is available at https://github.com/i2b2/i2b2-digitaltwin

1. Klann JG, Henderson DW, Morris M, et al. A broadly applicable approach to enrich electronic-health-record cohorts by identifying patients with complete data: a multisite evaluation. J Am Med Inform Assoc. 2023;30:1985–94. doi: 10.1093/jamia/ocad166
2. Lin KJ, Rosenthal GE, Murphy SN, et al. External Validation of an Algorithm to Identify Patients with High Data-Completeness in Electronic Health Records for Comparative Effectiveness Research. Clin Epidemiol. 2020;12:133–41. doi: 10.2147/CLEP.S232540
3. Sulieman L, Natarajan K. Using Cohort Diagnostics to Assess the Phenotypic Data Quality in All of Us Research Program. OHDSI 2023 Collaborator Showcase. OHDSI 2023 Global Symposium: https://www.ohdsi.org/wp-content/uploads/2023/10/CohortDiagnostics_AoU_OHDSI_2023_v2-Lina-Sulieman.pdf

Learn more about Argos

Contact Jeffrey Klann (jeff.klann@mgh.harvard.edu) for more information about Argos and data completeness. This research was approved by the IRBs at each institution and funded in part by NIH/NLM award R01LM013345 and NIH/NCATS award 1U24TR004111-01.

Visit our other poster and software demonstrations on Argos here at OHDSI 2026!

Poster #72 (9:15am)
Argos Automatic Phenotype Generator for OMOP: Phenotypes for 1000+ Medical Conditions

Demo #141 (2:45pm)
Argos-Atlas Bridge: A Cohort Builder for creating Concept Sets for Atlas

Demo #142 (3:30pm)
Argos: Simple Cohort Builder and Analysis Tool for Federated Networks Using OMOP

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-2026



Find your workgroup.

Fuel our mission.

ohdsi.org/workgroups