LLM Research in the OHDSI Community (Session 1 of 3)

OHDSI Community Call
June 2, 2026 • 11 am ET

Upcoming Community Calls

Date	Topic
June 2	LLM Research Around The World, Session 1
June 9	LLM Research Around The World, Session 2
June 16	LLM Research Around The World, Session 3
June 23	CANCELLED: OHDSI Summer School at Columbia University
June 30	OMOP & OHDSI Research Spotlight

June 9: LLM Research in the OHDSI Community

Niko Moeller-Grell
King's College

FastOMOP - multi agent cohort creation

Ed Burn
University of Oxford

LLM integration in the PhenotypeR R package

Joel Swerdel
Johnson & Johnson

Phenelope – tool for developing concept sets using LLM

Jared Houghtaling
Johnson & Johnson

LLM-Based Phenotype Refinement via CAPR

Adam Johnson
Duke University

Using Synthetic Data and Claude Code to Develop Transportable Analytic Code

Join us for the third session in our LLM series during the June 16 community call.

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Matthew Spotnitz, Adam S Faye, John Giannini, Tamara R Litwin, Yechiam Ostchega, and Lew Berman** on the recent publication of **Assessing data quality of inflammatory bowel disease patients in the All of Us research program** in *JAMIA Open*.

JAMIA Open, 2026, 9(3), ooag084
https://doi.org/10.1093/jamiaopen/ooag084
Research and Applications

AMIA OXFORD

Assessing data quality of inflammatory bowel disease patients in the *All of Us* research program

Matthew Spotnitz, MD, MPH[1,*], Adam S. Faye, MD, MS[2], John Giannini, PhD[1], Tamara R. Litwin, PhD, MPH[1], Yechiam Ostchega, PhD, RN[1], Lew Berman, PhD, MS[1]

[1]All of United States Research Program, National Institutes of Health, Bethesda, MD, United States
[2]Division of Gastroenterology & Hepatology, NYU Langone Health, NYU School of Medicine, New York, NY, United States
*Corresponding author: Matthew Spotnitz, MD, MPH, *All of United States* Research Program, Office of the Director, National Institutes of Health, 6710B Rockledge Drive, Bethesda, MD, 20892, United States (matthew.spotnitz@nih.gov)

Abstract

Purpose: **Inflammatory bowel disease (IBD) consists of Crohn's disease (CD) and ulcerative colitis (UC) and is a spectrum autoimmune disease of the gastrointestinal tract. Large scale real-world evidence studies could provide valuable evidence about IBD for personalized healthcare recommendations. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) standardizes electronic health record (EHR) data, allowing for research that incorporates multiple data sources. We are interested in whether OMOP CDM data on IBD are fit-for-use.**

Methods: **We selected IBD diagnosis codes to define the phenotype. We used a data quality checklist to evaluate 5 domains: conformance, completeness, concordance, plausibility, and temporality. We also did sensitivity analyses for CD and UC that consisted of at least 2 diagnosis codes that were at least 30 days apart.**

Results: **All of the phenotype-defining ICD source codes mapped to SNOMED. Many concept prevalences were low. A total of 78 (30.1%) out of 253 concept correlations were above our strength threshold (□ > 0.5). The age distribution of concepts and relative frequency of IBD medications were plausible. The median time between diagnosis and biopsy for the cohort was 4.43 [-0.05, 104.29] weeks. For the subgroup of participants who had sufficient data for the timeline analysis, IBD diagnosis concepts tended to occur first. In our sensitivity analyses, the completeness percentages of many variables in the UC and CD subgroups were similar to IBD, except for disease specific workup and treatment concepts.**

Conclusion: **We have shown a novel implementation of our data quality framework on IBD cohorts.**

Key words: inflammatory bowel disease, Crohn's disease, ulcerative colitis, precision medicine, electronic health record, data quality

OHDSI Shoutouts!

Congratulations to the team of **Boris Delange, Mirna El Ghosh, Celia Alvarez-Romero, Maxim Moinat, Paul Hilders, Patrick Rockenschaub, Jan van den Brand, Michel E. van Genderen, Christian Jung, Denis Delamarre, Sylvain Robert, Christel Daniel, Marc Cuggia, and Carlos Luis Parra-Calderón** on the recent publication of **Standardizing ICU Data Across Europe: Development of the INDICATE Minimal Data Dictionary** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care 1207
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260390

Standardizing ICU Data Across Europe: Development of the INDICATE Minimal Data Dictionary

Boris DELANGE[a,1], Mirna EL GHOSH[b], Celia ALVAREZ-ROMERO[c], Maxim MOINAT[d], Paul HILDERS[e], Patrick ROCKENSCHAUB[f], Jan VAN DEN BRAND[g], Michel E. VAN GENDEREN[g], Christian JUNG[h], Denis DELAMARRE[a], Sylvain ROBERT[a], Christel DANIEL[b,i], Marc CUGGIA[a] and Carlos Luis PARRA-CALDERÓN[c]

[a]*Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, Rennes, France*
[b]*Sorbonne Université, Inserm, Université Sorbonne Paris Nord, LIMICS, Paris, France*
[c]*Computational Health Informatics Group. Institute of Biomedicine of Seville, IBIS/ Virgen del Rocío University Hospital/CSIC/University of Seville, Spain*
[d]*Erasmus MC University Medical Center, Department of Medical Informatics, Rotterdam, The Netherlands*
[e]*Amsterdam UMC, Department of Intensive Care Medicine, Amsterdam, The Netherlands*
[f]*Institute of Clinical Epidemiology, Public Health, Health Economics, Medical Statistics, and Informatics, Medical University of Innsbruck, Innsbruck, Austria*
[g]*Erasmus MC University Medical Center, Department of Adult Intensive Care, Rotterdam, The Netherlands*
[h]*Medical Faculty, Department of Cardiology, Pulmonology and Vascular Medicine, Heinrich-Heine-University Dusseldorf, Dusseldorf, German and CARID (Cardiovascular Research Institute Düsseldorf), Dusseldorf, Germany.*
[i]*Medical Information Department, Henri Mondor Teaching Hospital, Greater Paris Teaching Hospital (Assistance Publique – Hôpitaux de Paris), Creteil, France*

OHDSI Shoutouts!

Congratulations to the team of **Boris Delange, Mathilde Bories, Sylvain Robert, Arthur Simon, Claire Charamel, Catherine Duclos, and Marc Cuggia** on the recent publication of **A Hybrid Pipeline for Mapping French UCD Drug Codes to RxNorm with Dosage Preservation** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260385

A Hybrid Pipeline for Mapping French UCD Drug Codes to RxNorm with Dosage Preservation

Boris DELANGE[a,1], Mathilde BORIES [a], Sylvain ROBERT[a], Arthur SIMON[a], Claire CHARAMEL[a], Catherine DUCLOS[b] and Marc CUGGIA[a]
[a]CHU Rennes, INSERM, LTSI-UMR 1099, Univ Rennes, 35000 Rennes, France
[b]Université Sorbonne Paris Nord, APHP, Avicenne, Santé Publique, INSERM, Sorbonne Université, Laboratoire d'Informatique Médicale et d'Ingénierie des connaissances en e-Santé, LIMICS, 93017, Bobigny, France

ORCiD ID: Boris DELANGE https://orcid.org/0009-0002-6055-6935, Mathilde BORIES https://orcid.org/0000-0001-9092-4298, Sylvain ROBERT https://orcid.org/0009-0000-7591-4331, Claire CHARAMEL https://orcid.org/0009-0003-2670-9819, Catherine DUCLOS https://orcid.org/0000-0001-8745-378X, Marc CUGGIA https://orcid.org/0000-0001-6943-3937

Abstract. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) requires drug data to be encoded in RxNorm for interoperable and dose-aware analyses. In France, medications are identified using UCD and CIP codes, but no direct mapping exists between UCD and RxNorm. We developed an automated three-stage pipeline to map French UCD codes to RxNorm Clinical Drugs while preserving dosage information. The process combined reference-based mapping using the French Drug Nomenclature (RUIM) and OHDSI Standardized Vocabularies, rule-based similarity matching, and agentic AI reasoning, with expert validation at each stage. Applied to the 38,345 UCD codes from the French national nomenclature, the pipeline achieved a 97.4% mapping rate with 86.0% precision when evaluated on the 500 most frequently used codes. This approach provides a reproducible and scalable framework for drug terminology harmonization in OMOP ETL processes.

Keywords. OMOP CDM; RxNorm; UCD; Drug mapping; Semantic interoperability; LLM.

OHDSI Shoutouts!

Congratulations to the team of **Aly Khalifa, Alexander Berler, and Rada Hussein** on the recent publication of **EHDS Data Continuum: A Proposed IHE Integration Profile for Bridging Primary and Secondary Health Data Use** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260383

EHDS Data Continuum: A Proposed IHE Integration Profile for Bridging Primary and Secondary Health Data Use

Aly KHALIFA[a], Alexander BERLER[b], and Rada HUSSEIN[c,1]
[a] *Department of Artificial Intelligence and Informatics, Mayo Clinic, 200 1St Street SW, Rochester, Minnesota, 55905, USA*
[b] *IHE Catalyst AISBL BluePoint, Boulevard Reyers 80, 1030 Brussels, Belgium*
[c] *Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*
ORCiD ID: Aly Khalifa https://orcid.org/0000-0002-7084-1345, Rada Hussein https://orcid.org/0000-0003-1257-4848

Abstract. The European Health Data Space (EHDS) aims to enable both the primary use of health data for direct patient care (EHDS1) and the secondary use of data for research, innovation, and public health (EHDS2). While each pillar is supported by distinct infrastructures—MyHealth@EU for clinical continuity and HealthData@EU for trusted reuse—the lack of seamless integration between them poses significant technical, governance, and legal challenges. Existing standards such as HL7 CDA, HL7 FHIR, OHDSI OMOP, CDISC, EEHRxF, and HealthDCAT-AP, together with IHE integration profiles including XDS, XCA, XCPD, ATNA, and BPPC, provide a robust foundation but remain fragmented across domains. This paper introduces the EHDS Data Continuum, a proposed IHE integration profile designed to bridge EHDS1 and EHDS2 by harmonizing interoperability, data quality, privacy, and governance requirements across the full data lifecycle. By aligning primary and secondary use within a coherent technical and policy framework, the EHDS Data Continuum supports the realization of a trustworthy, interoperable, and citizen-centered European Health Data Space.

Keywords. European Health Data Space, healthcare standards, integration profiles

OHDSI Shoutouts!

Congratulations to the team of **Matisse Decilap, Anya Okhmatovskaia, Jean-Paul R. Soucy, Dave van Steirteghem, Santiago Marquez, Aman Verma, John D. Fletcher, and David L. Buckeridge** on the recent publication of **Frequency-Based Prioritization of ICD-10-CA/CCI to OMOP Mapping in a Canadian Hospital Data Warehouse: Coverage and Usagi Performance** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260381

Frequency-Based Prioritization of ICD-10-CA/CCI to OMOP Mapping in a Canadian Hospital Data Warehouse: Coverage and Usagi Performance

Matisse DECILAP [a,b,1], Anya OKHMATOVSKAIA[a], Jean-Paul R. SOUCY [c], Dave VAN STEIRTEGHEM [c], Santiago MARQUEZ [a,c,d], Aman VERMA[a], John D. FLETCHER [d], David L. BUCKERIDGE[a,c,d]
[a] *McGill Clinical & Health Informatics, Department of Epidemiology and Biostatistics, McGill University*
[b] *University of Bordeaux*
[c] *Research Institute of the McGill University Health Centre*
[d] *McGill University Health Centre*
ORCiD ID: Matisse DECILAP https://orcid.org/0009-0003-9376-3858

Abstract. The adoption of the OMOP Common Data Model (CDM) for the secondary use of health data in Canada requires rigorous mapping of local terminologies such as ICD-10-CA and CCI to standard concepts. This study evaluates the coverage of these codes within the McGill University Health Centre (MUHC) in 2024 and analyzes the performance of Usagi in facilitating the mapping process. Our results show that mapping only 481 CCI codes (2.8%) and 1,015 specific ICD-10-CA codes (6.5%) achieves 80% coverage of hospital procedures and 99.1% of diagnoses, respectively. Usagi scores, significantly higher with long labels ($p < 0.001$), remain relatively low for automatic adoption, highlighting the need for expert validation. A prioritized and collaborative approach is essential to optimize resources and ensure the interoperability of Canadian data within OMOP.

Keywords. OMOP CDM, ICD-10-CA/CCI, Terminology mapping, Data coverage, Health data interoperability.

OHDSI Shoutouts!

Congratulations to the team of **Joaquim Vertentes Rosa, Raquel Paradinha, João Rafael Almeida, and José Luís Oliveira** on the recent publication of **Simplifying Cohort Definition with a Conversational Query Builder** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260362

Simplifying Cohort Definition with a Conversational Query Builder

Joaquim Vertentes ROSA[a], Raquel PARADINHA[a], João Rafael ALMEIDA[a] and José Luís OLIVEIRA[a]
[a]*IEETA / DETI, LASI, University of Aveiro, Portugal*
ORCiD ID: JVR 0009-0000-6248-1469 ; RP 0009-0006-3983-7926; JRA 0000-0003-0729-2264; JLO 0000-0002-6672-6176

Abstract. The secondary use of clinical data has been widely studied, addressing many challenges in conducting observational studies. However, the complexity of dataset structures and detailed data requirements has led researchers to develop user-friendly query builders, enabling medical researchers to define cohorts more efficiently across the datasets. Although these tools simplify the workflow, their learning curve can sometimes be steep. Motivated by improving the usability to conducted observational studies, we propose a modular conversational assistant framework that would address these limitations. It can be integrated in any web application as an javascript component, improving the system usability. Additionally, the proposed framework would employ deterministic algorithms to reduce computational overhead. The system would enable integration into existing medical information systems through configuration files rather than code modifications. Validation within the OHDSI ecosystem would demonstrate practical applicability for real-world observational research scenarios.

Keywords. cohorts, observational, chatbot, conversational assistant, ATLAS

OHDSI Shoutouts!

Congratulations to the team of **Loreen Ruhm, Laura Purfürst, Michael Ahmadi, Jacques Ehret, Maria Rönnefarth, Falk Meyer-Eschenbach, Katharina Schönrath, Stefanie Rudolph, Joachim E. Weber, Christof von Kalle, Johanna Nothacker, and the Belove Study Group** on the recent publication of **OMOP Extraction of Medical Text Using LLMs: Preliminary Results** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260354

OMOP Extraction of Medical Text Using LLMs: Preliminary Results

Loreen RUHM[a, 1], Laura PURFÜRST[a], Michael AHMADI[a], Jacques EHRET[a], Maria RÖNNEFARTH[a], Falk MEYER-ESCHENBACH[a,b], Katharina SCHÖNRATH[a], Stefanie RUDOLPH[a], Joachim E. WEBER[a], Christof VON KALLE[a,c], and Johanna NOTHACKER[a], on behalf of the BeLOVE Study Group
[a] *Berlin Institute of Health (BIH), Charité Universitätsmedizin, Berlin, Germany*
[b] *Institute for Medical Informatics, Charité Universitätsmedizin, Berlin, Germany*
[c] *Luxembourg Institute of Health (LIH), Luxembourg*
ORCiD ID: Loreen Ruhm https://orcid.org/0000-0001-7217-2993

Abstract. We used medium-sized LLMs to extract medication data from German discharge letters into the OMOP CDM, achieving >85% accuracy and >75% F1-score, demonstrating the feasibility of OMOP extraction with medium-sized models.

Keywords. OMOP CDM, Large Language Models, GenAI, German Clinical Text

1. Introduction

LLMs enable the transformation of unstructured clinical text into structured data for secondary use [1][2], including mapping to Common Data Models (CDMs) [3][4]. We convert discharge letters from the longitudinal BeLOVE cohort [5] into an OMOP CDM instance, focusing on the medication domain (drug_exposure), using state-of-the-art open-weight LLMs. The OMOP CDM provides a widely adopted, interoperable standard for research. Contributions: (1) privacy-preserving with medium-sized open-weight LLMs; (2) extraction of real-world data for OMOP integration.

OHDSI Shoutouts!

Congratulations to the team of **Félix Berthou, Ghilsain Vaillant, Bastien Rance, and Adrien Coulet** on the recent publication of **Build and Query Indexes of Clinical Documents with Easy-to-Reuse Pipelines** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260325

Build and Query Indexes of Clinical Documents with Easy-to-Reuse Pipelines

Félix BERTHOU[a], Ghilsain VAILLANT[a], Bastien RANCE[a,b] and Adrien COULET[a,*]
[a] *Inria, Inserm, Université Paris Cité, HeKA U1346;* [b]*Centre de Recherche des Cordeliers, Inserm, Université Paris Cité, Sorbonne Université, U1338;* [c]*Assistance Publique - Hôpitaux de Paris, HEGP, Paris, France*
ORCiD ID: Ghilsain VAILLANT https://orcid.org/0000-0003-0267-3033, Bastien RANCE https://orcid.org/0000-0003-4417-1197, Adrien COULET https://orcid.org/0000-0002-1466-062X

Abstract. Electronic Health Records are a central source of healthcare data, containing structured data alongside unstructured clinical texts. The latter capture detailed reasoning, observations, treatment plans and clinical evolutions, which are crucial for phenotyping, and real-world evidence generation. Natural language processing enables the extraction, thus the subsequent use, of these crucial elements; however, these extractions remain one-off, study-specific efforts. This is detrimental as the extracted elements could be valuable for future research. We present `medkit Seshat`, an open-source Python pipeline that: (1) ingests free text, (2) recognizes relevant entities, (3) normalizes them with OMOP vocabularies, (4) builds an index that can either be searched by concept or by document. In addition, we share a flexible web UI to illustrate the interest of built indexes in terms of search, text analysis and export. `Seshat` aims at facilitating the reuse and adaptation of this prototypical pipeline to various purposes, with the main objective of enabling the secondary use of results of phenotyping campaigns.

Keywords. clinical NLP, Phenotyping, Normalization, Indexing, Open Science

OHDSI Shoutouts!

Congratulations to the team of **Thomas Ruprecht, Eveline Prochaska, and Elisa Henke** on the recent publication of **Crique: A System for Automatic Extraction and Formalization of Eligibility Criteria for Clinical Trials** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260290

Crique: A System for Automatic Extraction and Formalization of Eligibility Criteria for Clinical Trials

Thomas RUPRECHT [a,1], Eveline PROCHASKA [a] and Elisa HENKE [a]
[a] *Institute for Medical Informatics and Biometry, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany*

Abstract. Defining formal and structured eligibility criteria for digital recruitment assistant systems can be a challenging interdisciplinary task requiring collaboration between technical and clinical experts. We introduce *crique,* a free and open source software capable of extracting formal eligibility criteria definitions from study protocols. The extracted eligibility criteria can be exported to OHDSI Atlas and used in conjunction with the standardized OMOP common data model. In this paper, we present our modular process for the extraction of eligibility criteria, describe our implementation in detail, and show examples of extracted eligibility criteria. We conclude with challenges and propose a strategy to evaluate the accuracy of extracted criteria using our software.

Keywords. OHDSI, OMOP, eligibility criteria, clinical trials, recruitment assistance

OHDSI Shoutouts!

Congratulations to the team of **Claire Charamel, Arthur Le Gall, Marc Cuggia, and Boris Delange** on the recent publication of **OPTIMA-DAW: Improving Cerebral Vasospasm Detection After Aneurysmal Subarachnoid Haemorrhage Using Machine Learning** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

doi:10.3233/SHTI260218

OPTIMA-DAW: Improving Cerebral Vasospasm Detection After Aneurysmal Subarachnoid Haemorrhage Using Machine Learning

Claire CHARAMEL[a,b] , Arthur LE GALL[b], Marc CUGGIA[a] and Boris DELANGE[a,1]
[a] *Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, Rennes, France*
[b] *Service d'Anesthésie - Réanimation et USC Chirurgicale, Trauma Center. 2 rue Henri Le Guilloux 35033 Rennes, France*
Claire CHARAMEL https://orcid.org/0009-0003-2670-9819 ,
Arthur LE GALL https://orcid.org/0000-0002-6742-7477 ,
Marc CUGGIA https://orcid.org/0000-0001-6943-3937 ,
Boris DELANGE https://orcid.org/0009-0002-6055-6935

Abstract. Cerebral vasospasm is a serious complication after aneurysmal subarachnoid haemorrhage (aSAH). We trained machine learning models on 168 patients (225 CTA timepoints) using standardized clinical data (OMOP CDM). XGBoost achieved the best performance (AUROC 0.79; 95% CI 0.65-0.91).

Keywords. Machine Learning (ML), Clinical Decision Support System, Cerebral vasospasm (CV), OMOP Common Data Model (OMOP CDM)

OHDSI Shoutouts!

Congratulations to the team of **Francisco Lozano, Julia Sánchez Esquivel, Sergio Paraíso-Medina, Raúl Alonso-Calvo, Paloma Jimeno, Inmaculada Luengo, and Víctor Maojo** on the recent publication of **Federated Multi-Agent Architecture for Harmonizing Public Health Datasets into OMOP and FHIR Standards** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260208

Federated Multi-Agent Architecture for Harmonizing Public Health Datasets into OMOP and FHIR Standards

Francisco LOZANO[a,1], Julia SÁNCHEZ ESQUIVEL[a], Sergio PARAÍSO-MEDINA[a], Raúl ALONSO-CALVO[a], Paloma JIMENO[b], Inmaculada LUENGO[b], Víctor MAOJO[a]
[a] *Biomedical Informatics Group, DIA & DLSIIS, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo S/N, Madrid, Spain*
[b] *Hi IBERIA, Juan Hurtado de Mendoza 14 28036 Madrid*

Abstract. This study presents a federated system developed within the SHIELD project, part of Horizon Europe's effort to reduce non-communicable diseases. It integrates retrospective and prospective clinical data using LLM-based multi-agent systems for automated ETL and natural language querying. Harmonization of MIMIC-IV, ELSA, and synthetic data into OMOP CDM and FHIR enabled validation. Results show high mapping accuracy and demonstrate the feasibility of scalable and interoperable data integration supporting clinical decision-making.

Keywords. Non-communicable diseases, OMOP CDM, FHIR, Federated learning, Multi-agent systems, Large language models

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

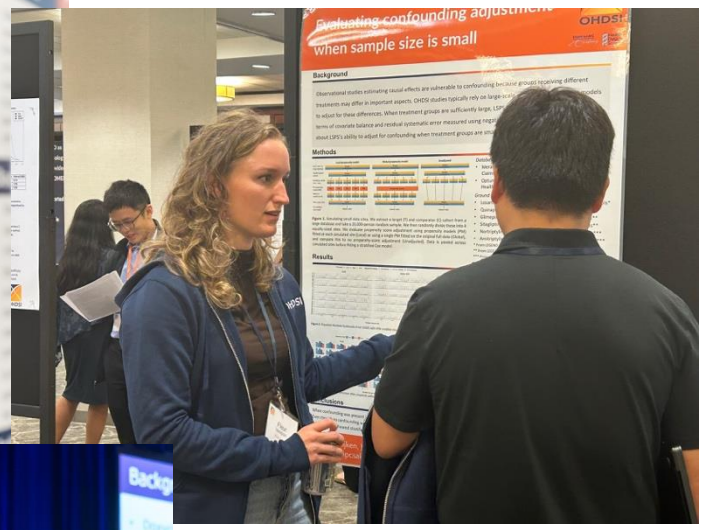
Upcoming Workgroup Calls

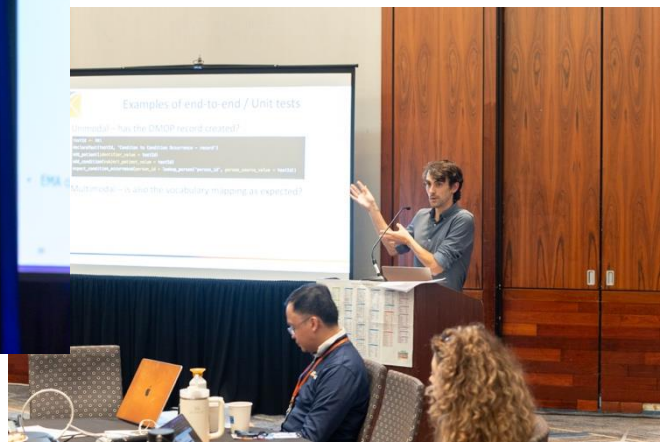
Date	Time (ET)	Meeting
Wednesday	9 am	Tidy R Programming with OMOP
Thursday	8 am	Medical Devices
Thursday	10 am	ATLAS/WebAPI
Thursday	10 am	Africa Chapter (ZOOM)
Thursday	10 am	GIS-Geographic Information System
Thursday	11 am	Industry
Thursday	11 am	Themis
Thursday	1 pm	Oncology Vocabulary/Development Subgroup
Thursday	2 pm	Early-Stage Researchers
Friday	11:30 am	Steering Group
Monday	9 am	Vaccine Vocabulary
Tuesday	9 am	Oncology Genomic Subgroup
Tuesday	10 am	CDM Survey Subgroup

Final Countdown: Showcase Deadline

The **call for participation** is open for the 2026 Global Symposium.

The submission deadline is June 5 at 8 pm ET.





3 Days Remaining!

ohdsi.org/OHDSI2026

2026 OHDSI Global Symposium

Registration is OPEN for the **2026 OHDSI Global Symposium**, which will be held Oct. 20-22 in New Brunswick, N.J., USA.

Oct. 20: Tutorials
Oct. 21: Plenaries, Showcase
Oct. 22: Workgroup Activities

ohdsi.org/OHDSI2026

June Newsletter is Available

OHDSI

The Journey Newsletter (June 2026)

Welcome to the June edition of the OHDSI newsletter, where we highlight a trio of community calls dedicated to the rapidly evolving landscape of LLM research around the world. Time is running out to share your own work (LLM or otherwise), as **the deadline to submit to the 2026 Global Symposium Collaborator Showcase is less than a week away (June 5th, 8 pm ET)**! Inside this issue, you will also find spotlights on workgroup leads Anthony Sena and Benjamin Martin, key community updates, and a look at more than 25 recent OHDSI publications from May. **#JoinTheJourney**

Podcast: OMOP Efficiency, LLM Research

In the June 2026 On The Journey podcast, Patrick Ryan and Craig Sachson discuss milestones in the community's effort to build a global pharmacovigilance system, highlighting a week-long collaboration with the Uppsala Monitoring Center (UMC) that demonstrated the speed and efficiency of the OMOP CDMI and OHDSI tools. Patrick also provides a state-of-the-art overview of OHDSI's research into LLMs, exploring how generative AI can scale responsible evidence dissemination while mitigating the risks of unreliable medical data. *(If video does not appear, please click 'view this email in your browser.)*

Community Updates

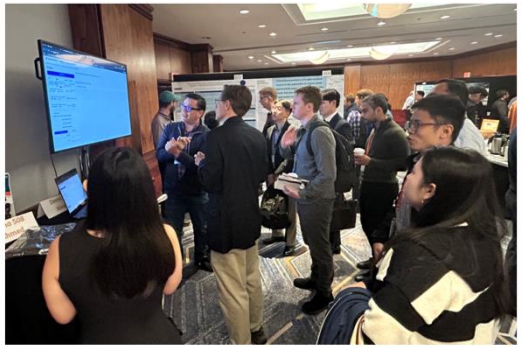
Where Have We Been?

- **Integrating Swedish data to OMOP:** In a recent week-long collaboration, OHDSI partnered with the Uppsala Monitoring Center (UMC)—the WHO Collaborating Centre for International Drug Monitoring—to help integrate Sweden's national registry data into the OMOP Common Data Model. Utilizing OHDSI tools and data quality packages, the team successfully designed and executed a full study using Strategus in just five days. This achievement highlights the incredible speed and efficiency of the OMOP CDM when onboarding new data partners to advance global pharmacovigilance. *Learn more in the podcast above.*
- **Workgroup Spotlights (Vocabulary, Evidence Network):** Leaders from both the Vocabulary and Evidence Network workgroups provided detailed presentations on recent work and upcoming initiatives during our community call series. These talks included a demo on the recent enhancements to the community contribution process for vocabulary updates, as well as a look at the current workstream model for the Evidence Network. *Check out the May presentations section below to learn more.*
- **OHDSI Europe Refresh:** During the 2026 OHDSI Europe Symposium, the new OHDSI Europe website was officially introduced. Congratulations to **Sicco den Otter, Aniek Markus** and **Ilse Vermeulen** on leading this initiative.

Where Are We Now?

- **Global Showcase Submission Deadline:** The clock is ticking! The 2026 Global Symposium Showcase submission deadline is this **Friday, June 5 at 8 pm ET**. We are accepting submissions for posters, software demos or lightning talks. Visit the showcase homepage to find all submission details and links to share your science, open-source tools, and real-world evidence with the global community.
- **Europe Symposium Research:** The #OHDSISocialShowcase will begin sharing research from the 2026 Europe Symposium via our social channels this month. Please follow us on LinkedIn, X/Twitter, Bluesky and Instagram, to learn more about the research happening in our community.
- **Columbia University Summer School:** The 2026 Summer School in Observational Health Data Science & Informatics, AI, and RWE takes place June 22-26, and we are nearing capacity for the event. If you want to take part in this intensive, in-person training led by **Patrick Ryan, George Hripcsak, Anna Ostropolets,** and **Karthik Natarajan**, please head to the event homepage and register soon.

Discover How Generative AI and LLMs Are Transforming Health Research in OHDSI



Generative AI and Large Language Models (LLMs) have the potential to massively accelerate our research mission, and we are seeing incredible breakthroughs across our global community. To highlight these exciting projects and spark new ideas, we are hosting a special, three-part community call series focused entirely on AI in OHDSI, kicking off on Tuesday, June 2, at 11:00 AM ET.

Over three straight weeks, you will hear from researchers who are breaking new ground and putting AI to work in real-world health data projects. Whether you are already building AI tools yourself or just curious about how this technology will change the future of medicine, these calls are the perfect place to learn what is happening right now. Check out our full three-week tentative agenda below, and be sure to save the dates for June 2, June 9, and June 16!

Tentative Schedule

June 2

- **Intro to OHDSI research in LLMs, GenAI WG updates** *(Martijn Schuemie, Johnson & Johnson / GenAI Workgroup Lead)*
- **Ariadne: Automated vocabulary mappings using AI** *(Anna Ostropolets, Johnson & Johnson / Columbia University; Martijn Schuemie, Johnson & Johnson)*
- **LLM–Based Classification of ICD-10-CM to SNOMED Mappings for Improved Semantic Fidelity in OHDSI** *(Dmytry Dymshyts, Johnson & Johnson)*

My Journey: Anthony Sena

In the latest installment of our "My Journey" series, Anthony Sena (Director of Observational Healthcare Data Analytics at Johnson & Johnson) shares how his background as a software engineer led him to the world of open-source healthcare analytics. Anthony discusses his pride in developing the OHDSI open-source tool stack (HADES and Atlas) and why a "federated data network" is essential to learning from the shared healthcare experiences of patients globally.

The Final Countdown is On! Submit Your #OHDSI2026 Showcase Report by Friday

ohdsi.org/spotlight-julio-oliveira

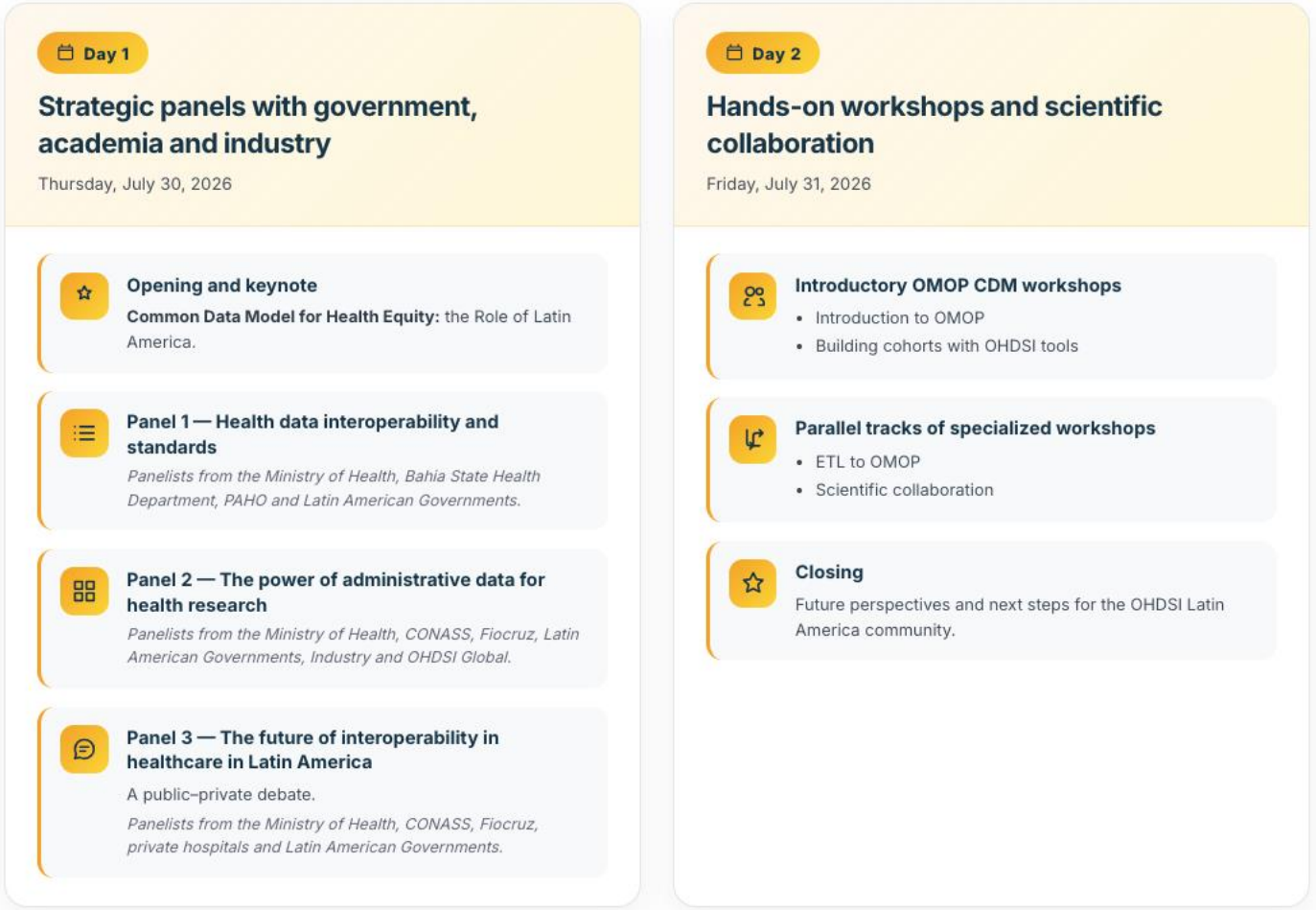
Spotlight: Benjamin Martin

In the latest edition of the Collaborator Spotlight, **Benjamin Martin** looks back at his journey to OHDSI, the OMOP impact on Johns Hopkins students, how the ESR workgroup tries to help junior researchers begin their OHDSI journey, and plenty more.

ohdsi.org/spotlight-benjamin-martin

First Latin America Symposium – July 30-31

Registration is open for the first OHDSI Latin America Symposium, taking place July 30-31 in Salvador, Brazil.



ohdsilatam.org

Columbia DBMI Summer School

The 2026 Summer School in Observational Health Data Science & Informatics, AI, and Real World Evidence

COLUMBIA | COLUMBIA UNIVERSITY DEPARTMENT OF BIOMEDICAL INFORMATICS

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

June 22–26, 2026, Columbia Biomedical Informatics

The Columbia OHDSI Summer School provides health professionals, researchers, and industry practitioners with an immersive, hands-on training to working with real-world health data and generating real-world evidence (RWE). Participants will explore the types of healthcare data captured during routine clinical care—such as electronic health records and administrative claims—and learn how to standardize these data using the OMOP Common Data Model to support collaborative, distributed research as part of a data network.

Over the course of the week, participants will engage with three real-world analytic use cases:

- **Clinical characterization** – using descriptive epidemiology to study disease natural history and treatment patterns
- **Population-level estimation** – applying causal inference to assess drug safety and comparative effectiveness
- **Patient-level prediction** – leveraging machine learning for early disease detection and precision medicine

Participants will be guided through the full RWE study lifecycle: from designing observational studies tailored to each use case, to applying open-source tools form the OHDSI community, and executing analyses across real-world data sources.

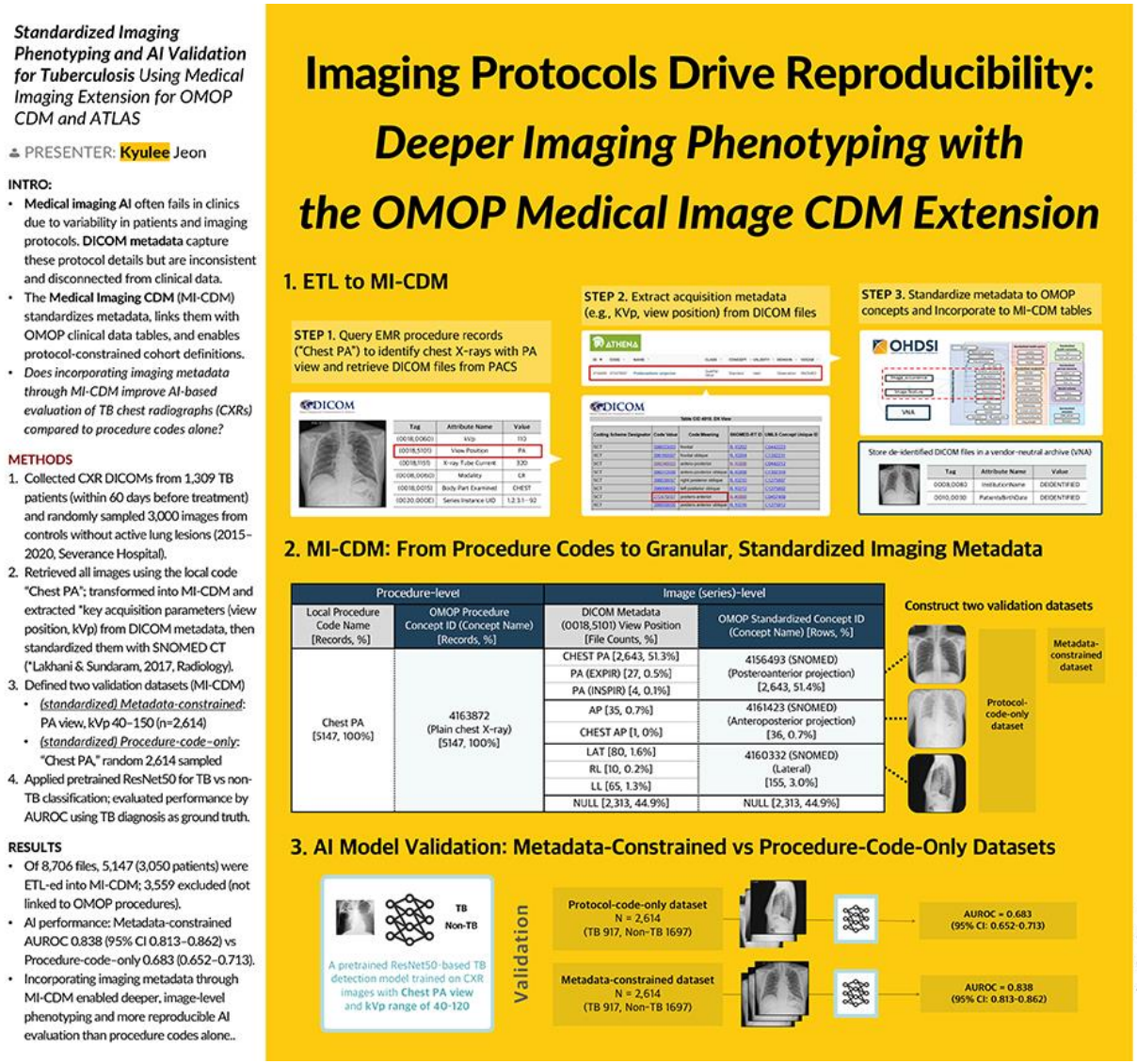
The curriculum combines foundational lectures on analytical methods with hands-on, interactive, faculty-led group exercises. In addition, participants will have dedicated time to develop and advance their own study concepts with personalized feedback and mentoring.

#OHDSISocialShowcase This Week

Monday

Standardized Imaging Phenotyping and AI Validation for Tuberculosis Using Medical Imaging Extension for OMOP CDM and ATLAS

(Kyulee Jeon, Minseong Kim, Soon Ho Yoon, Seng Chan You)



Standardized Imaging Phenotyping and AI Validation for Tuberculosis Using Medical Imaging Extension for OMOP CDM and ATLAS

PRESENTER: Kyulee Jeon

INTRO:

- **Medical imaging AI** often fails in clinics due to variability in patients and imaging protocols. **DICOM metadata** capture these protocol details but are inconsistent and disconnected from clinical data.
- The **Medical Imaging CDM** (MI-CDM) standardizes metadata, links them with OMOP clinical data tables, and enables protocol-constrained cohort definitions.
- *Does incorporating imaging metadata through MI-CDM improve AI-based evaluation of TB chest radiographs (CXRs) compared to procedure codes alone?*

METHODS

1. Collected CXR DICOMs from 1,309 TB patients (within 60 days before treatment) and randomly sampled 3,000 images from controls without active lung lesions (2015–2020, Severance Hospital).
2. Retrieved all images using the local code "Chest PA"; transformed into MI-CDM and extracted "key acquisition parameters (view position, kVp) from DICOM metadata, then standardized them with SNOMED CT ("Lakhani & Sundaram, 2017, Radiology).
3. Defined two validation datasets (MI-CDM)
   - *(standardized) Metadata-constrained*: PA view, kVp 40–150 (n=2,614)
   - *(standardized) Procedure-code–only*: "Chest PA," random 2,614 sampled
4. Applied pretrained ResNet50 for TB vs non-TB classification; evaluated performance by AUROC using TB diagnosis as ground truth.

RESULTS

- Of 8,706 files, 5,147 (3,050 patients) were ETL-ed into MI-CDM; 3,559 excluded (not linked to OMOP procedures).
- AI performance: Metadata-constrained AUROC 0.838 (95% CI 0.813–0.862) vs Procedure-code–only 0.683 (0.652–0.713).
- Incorporating imaging metadata through MI-CDM enabled deeper, image-level phenotyping and more reproducible AI evaluation than procedure codes alone..

Imaging Protocols Drive Reproducibility:
Deeper Imaging Phenotyping with the OMOP Medical Image CDM Extension

1. ETL to MI-CDM

STEP 1. Query EMR procedure records ("Chest PA") to identify chest X-rays with PA view and retrieve DICOM files from PACS

STEP 2. Extract acquisition metadata (e.g., KVp, view position) from DICOM files

STEP 3. Standardize metadata to OMOP concepts and Incorporate to MI-CDM tables

2. MI-CDM: From Procedure Codes to Granular, Standardized Imaging Metadata

Procedure-level		Image (series)-level	
Local Procedure Code Name [Records, %]	OMOP Procedure Concept ID (Concept Name) [Records, %]	DICOM Metadata (0018,5101) View Position [File Counts, %]	OMOP Standardized Concept ID (Concept Name) [Rows, %]
Chest PA [5147, 100%]	4163872 (Plain chest X-ray) [5147, 100%]	CHEST PA [2,643, 51.3%]	4156493 (SNOMED) (Posteroanterior projection) [2,643, 51.4%]
		PA (EXPIR) [27, 0.5%]	
		PA (INSPIR) [4, 0.1%]	
		AP [35, 0.7%]	4161423 (SNOMED) (Anteroposterior projection) [36, 0.7%]
		CHEST AP [1, 0%]	
		LAT [80, 1.6%]	4160332 (SNOMED) (Lateral) [155, 3.0%]
		RL [10, 0.2%]	
		LL [65, 1.3%]	
		NULL [2,313, 44.9%]	NULL [2,313, 44.9%]

Construct two validation datasets

Metadata-constrained dataset

Protocol-code-only dataset

3. AI Model Validation: Metadata-Constrained vs Procedure-Code-Only Datasets

TB / Non-TB

A pretrained ResNet50-based TB detection model trained on CXR images with **Chest PA view** and **kVp range of 40-120**

Validation

Protocol-code-only dataset
N = 2,614
(TB 917, Non-TB 1697)

AUROC = 0.683
(95% CI: 0.652-0.713)

Metadata-constrained dataset
N = 2,614
(TB 917, Non-TB 1697)

AUROC = 0.838
(95% CI: 0.813-0.862)

OMOP Medical Image CDM Extension:

Example of how a chest X-ray taken on 2017-02-18 is represented in OMOP CDM with the imaging extension:

- **Procedure_occurrence:** records the clinical procedure
- **Image_occurrence:** links the patient and procedure to the actual image file or pixel data (e.g., file path, series UID).
- **Measurement:** stores metadata values extracted from the DICOM header (e.g., KVP = 110, tube voltage).
- **Image_feature:** connects the image with the extracted metadata as standardized imaging features.
- In **ATLAS**, MI-CDM can be used to define protocol-constrained cohorts without burdensome preprocessing.

Kyulee Jeon[12*], Minseong Kim[3*], Soon Ho Yoon[4], Seng Chan You[12]

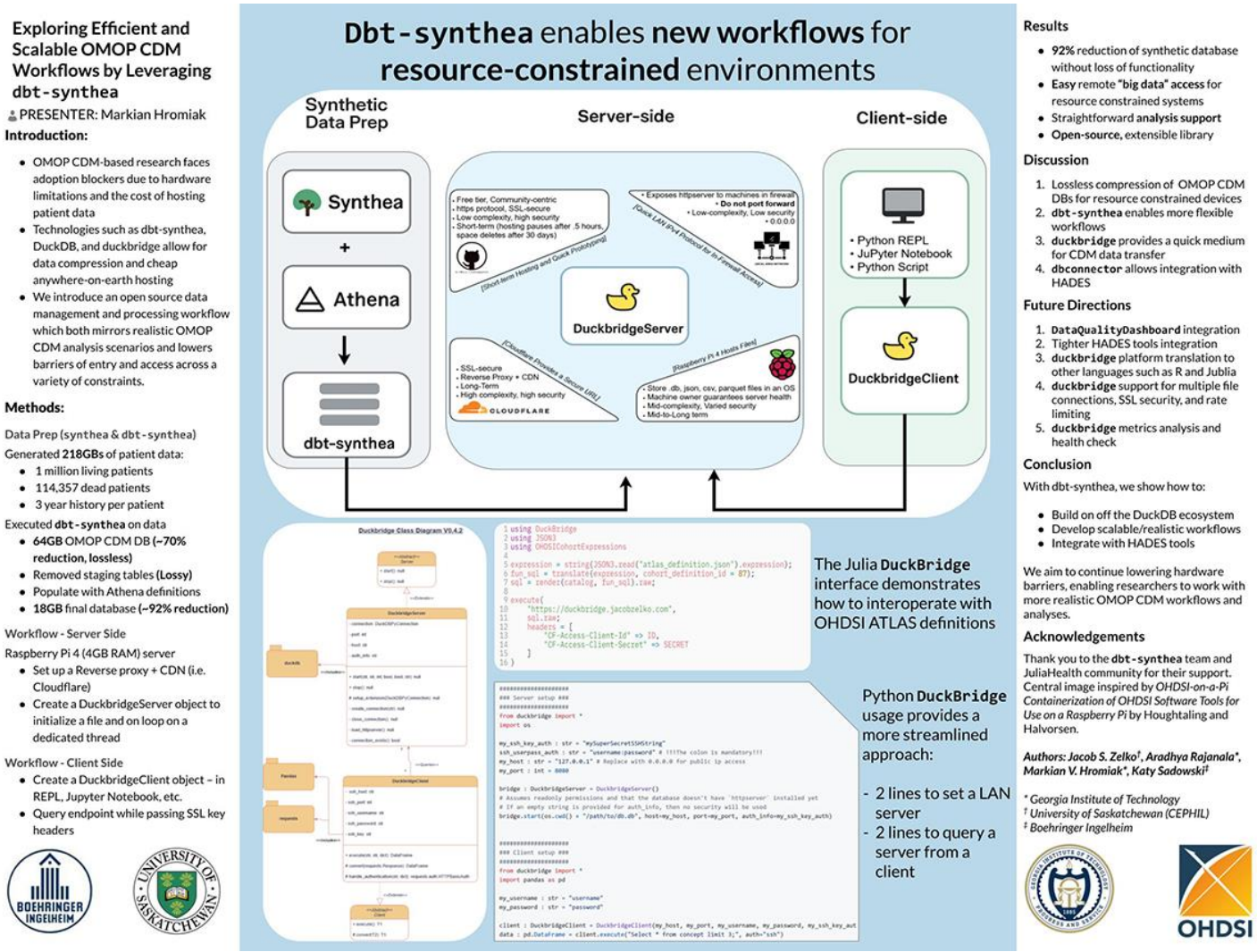
1. *Dept. of Biomedical Systems Informatics, Yonsei University College of Medicine*
2. *Yonsei Institute for Digital Health, Yonsei University*
3. *Dept. of Material Sciences and Engineering, Yonsei University College of Engineering*
4. *Dept. of Radiology, Seoul National University College of Medicine*

#OHDSISocialShowcase This Week

Tuesday

Exploring Efficient and Scalable OMOP CDM Workflows by Leveraging dbt-synthea

(Markian Hromiak, Aradhya Rajanala, Jacob S. Zelko, Katy Sadowski)



Exploring Efficient and Scalable OMOP CDM Workflows by Leveraging dbt-synthea

PRESENTER: Markian Hromiak

Introduction:

- OMOP CDM-based research faces adoption blockers due to hardware limitations and the cost of hosting patient data
- Technologies such as dbt-synthea, DuckDB, and duckbridge allow for data compression and cheap anywhere-on-earth hosting
- We introduce an open source data management and processing workflow which both mirrors realistic OMOP CDM analysis scenarios and lowers barriers of entry and access across a variety of constraints.

Methods:

Data Prep (synthea & dbt-synthea)

Generated **218GBs** of patient data:

- 1 million living patients
- 114,357 dead patients
- 3 year history per patient

Executed **dbt-synthea** on data

- **64GB** OMOP CDM DB **(~70% reduction, lossless)**
- Removed staging tables **(Lossy)**
- Populate with Athena definitions
- **18GB** final database **(~92% reduction)**

Workflow - Server Side

Raspberry Pi 4 (4GB RAM) server

- Set up a Reverse proxy + CDN (i.e. Cloudflare)
- Create a DuckbridgeServer object to initialize a file and on loop on a dedicated thread

Workflow - Client Side

- Create a DuckbridgeClient object – in REPL, Jupyter Notebook, etc.
- Query endpoint while passing SSL key headers

Dbt-synthea enables **new workflows** for **resource-constrained** environments

Synthetic Data Prep

Server-side

Client-side

Synthea + Athena → dbt-synthea

- Free tier, Community-centric
- https protocol, SSL-secure
- Low complexity, high security
- Short-term (hosting pauses after .5 hours, space deletes after 30 days)

[Short-term Hosting and Quick Prototyping]

- Exposes httpserver to machines in firewall
- Do not port forward
- Low-complexity, Low security
- 0.0.0.0

[Quick LAN IPv4 Protocol for In-Firewall Access]

DuckbridgeServer

[Cloudflare Provides a Secure URL]

- SSL-secure
- Reverse Proxy + CDN
- Long-Term
- High complexity, high security

CLOUDFLARE

[Raspberry Pi 4 Hosts Files]

- Store .db, .json, .csv, parquet files in an OS
- Machine owner guarantees server health
- Mid-complexity, Varied security
- Mid-to-Long term

- Python REPL
- JuPyter Notebook
- Python Script

DuckbridgeClient

Duckbridge Class Diagram V0.4.2

The Julia **DuckBridge** interface demonstrates how to interoperate with OHDSI ATLAS definitions

Python **DuckBridge** usage provides a more streamlined approach:

- 2 lines to set a LAN server
- 2 lines to query a server from a client

Results

- **92%** reduction of synthetic database without loss of functionality
- **Easy** remote **"big data" access** for resource constrained systems
- Straightforward **analysis support**
- **Open-source,** extensible library

Discussion

1. Lossless compression of OMOP CDM DBs for resource constrained devices
2. **dbt-synthea** enables more flexible workflows
3. **duckbridge** provides a quick medium for CDM data transfer
4. **dbconnector** allows integration with HADES

Future Directions

1. **DataQualityDashboard** integration
2. Tighter HADES tools integration
3. **duckbridge** platform translation to other languages such as R and Julia
4. **duckbridge** support for multiple file connections, SSL security, and rate limiting
5. **duckbridge** metrics analysis and health check

Conclusion

With dbt-synthea, we show how to:

- Build on off the DuckDB ecosystem
- Develop scalable/realistic workflows
- Integrate with HADES tools

We aim to continue lowering hardware barriers, enabling researchers to work with more realistic OMOP CDM workflows and analyses.

Acknowledgements

Thank you to the **dbt-synthea** team and JuliaHealth community for their support. Central image inspired by *OHDSI-on-a-Pi Containerization of OHDSI Software Tools for Use on a Raspberry Pi* by Houghtaling and Halvorsen.

Authors: Jacob S. Zelko†, Aradhya Rajanala*, Markian V. Hromiak*, Katy Sadowski‡

** Georgia Institute of Technology*
† University of Saskatchewan (CEPHIL)
‡ Boehringer Ingelheim

#OHDSISocialShowcase This Week

Wednesday

Quantifying Condition Completeness using Medications in the All of Us Research Program

(**Lina Sulieman**, Xinzhuo Jiang, Joshua Smith, Karthik Natarajan, Paul Harris)

Quantifying Condition Completeness using Medications in the All of Us Research Program

PRESENTER: **Lina Sulieman**

INTRODUCTION:

- Existing EHR completeness metrics: assessing the presence of expected clinical events
  - General metrics
  - Lack specificity
- Medications: prescribed to treat specific conditions
- **Objectives:** Using Medication-Indications to assess condition completeness

METHODS

- Creating Medication-Indication :
  - OMOP concept relationship: National Drug File-Reference Terminology (NDF-RT)
  - Drug Evidence Base (DEB): MEDLINE, MedlinePLUS, NDF-RT, SIDER, DrugBank
  - Mapping NDF-RT in OMOP and CUI in DEB to SNOMED codes
- Dataset: All of Us curated dataset version-8 (CDR-8)
- Assessing condition documentation completeness:
  - At least two medication entries
  - At least one condition
  - e.g. Two Albuterol for any asthma or bronchospasms entries
- Condition Completeness: participants with medication-indications/participants with the medication
- Sensitivity Analysis: Evidence strength in relation to Condition Completeness

RESULTS

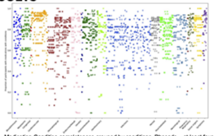


Figure 1. Medication-Condition completeness grouped by conditions Phecode – at least four resources

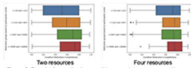


Figure 2. Box plot for medication-condition completeness grouped by number of participants

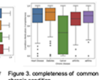


Figure 3. completeness of common chronic condition

Using **medication-indication evidence** reported in the OMOP and Drug Evidence Base datasets can **quantify the completeness of conditions documentation at scale.**

High Completion: Chronic diseases

High Missingness: Genetics and pregnancy

VANDERBILT UNIVERSITY MEDICAL CENTER

Department of Biomedical Informatics

AMMO BAR

- Reasons for missing conditions treated by medication:
  - Mapping errors
  - EHR Fragmentation
  - Missing non-billable codes
- Keywords and definition:
  - Medication-Indication: The medication that is used to treat a condition
  - Evidence strength: The number of resources that mentioned a drug to treat the studied condition
- Dataset CDR-8:
  - DEB2: 6,431 medication-conditions relationships, 1130 conditions
  - 362,134 participants
  - 326,073 participants (90.04%) had both drugs and conditions
  - 8,089 participants: medications without any condition

Medication	Conditions treated by medication	Participant num	Condition completeness
Acetaminophen	Fever;Headache;Toothache;Degenerative polyarthritis; Myalgia ;Hyperthemia; Migraine Disorders; Common Cold; Back Pain; Pharyngitis; Arthritis; Dysmenorrhea; Influenza;Pain;	165794	0.729
lidocaine	Poisoning; Pain finding; Unspecified disorder of esophagus; Coronary arteriosclerosis in patient with history of previous myocardial infarction; Ventricular tachycardia, unspecified;Other type of myocardial infarction;	124291	0.504
Sodium chloride	Adiposity; Dry eye; Other obesity not elsewhere classified; Tear film insufficiency, unspecified;Overweight and obesity; Tear film insufficiency;Sodium deficiency;Dehydration;Genetic susceptibility to obesity	123176	0.536
Ondansetron	Alcoholic Intoxication, Chronic;Nausea;Vomiting;Postoperative Nausea	112985	0.363
fentanyl	Pain, Postoperative;Pain, not elsewhere classified;Observation of pattern of pain;Discomforting present pain;Chronic pain;Pain, unspecified;Acute onset pain;Pain finding;Pain	96776	0.569
oxycodone	Pain, Postoperative;Chronic pain;Severe pain;Pain;Postoperative pain	89375	0.594
ibuprofen	Rheumatoid bursitis, unspecified hip;Ankylosing spondylitis of cervicothoracic region;Fever;Chronic gout, unspecified ;Rheumatoid Arthritis;Ankylosing spondylitis;Common Cold; Migraine Disorders;Hyperthemia;Dysmenorrhea (finding);Pain, Postoperative;	84651	0.815
midazolam	Agitated ;Absence epileptic syndrome, intractable, with status epilepticus;Unspecified dementia;Dravet syndrome, intractable, with status epilepticus;Anxiety disorder; Lennox-Gastaut syndrome, intractable, with status epilepticus; Psychomotor agitation;	78495	0.373
albuterol	Asthma;Bronchial Spasm;Bronchospasm; Chronic Obstructive Airway Disease;Pneumopathy;Lung diseases;	71425	0.823
propofol	Dementia;Epileptic;Agitation;Delirium caused by substance or medication;Lennox-Gastaut syndrome, with status epilepticus; Pain;Organic brain syndrome;Lafora progressive myoclonus epilepsy, not intractable, with status epilepticus;Psychomotor agitation;	69251	0.232
aspirin	Rheumatoid bursitis;Fever ;Rheumatoid arthritis; Cerebrovascular accident ; Inflammation; Pharyngitis; Rheumatoid arthritis; Degenerative polyarthritis; Streptococcal Infections; Observation of pattern of pain; Myocardial Infarction; Angina, Unstable; Ankylosing spondylitis; Pre-Eclampsia; Coughing; Gout; Headache; Transient Ischemic Attack ;Mucocutaneous Lymph Node Syndrome; Heart Diseases;	65326	0.898
docusate	Constipation	64307	0.366

Lina Sulieman, PhD[1], Xinzhuo Jiang, MS[2], Joshua Smith, PhD[1], Karthik Natarajan, PhD[2], Paul Harris, PhD[1]
1Vanderbilt University Medical Center, Nashville, TN;
2Columbia University Irving Medical Center, New York, NY

All of Us DATA & RESEARCH CENTER

OHDSI

#OHDSISocialShowcase This Week

Thursday

NLP based Extraction and OMOP Standardization of Breast Cancer Clinical Data from Indian Discharge Summaries

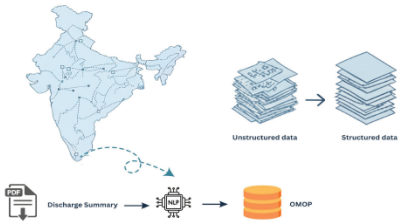
(Swetha Kiranmayi Jakkuva, Khansa Fathima, M R Sai Dileep, Shreema S Rao, Sanjay R, Sai Pattabhiram L)

NLP-based Extraction and OMOP Standardization of Breast Cancer Clinical Data from Indian Discharge Summaries

Dr. Swetha Kiranmayi Jakkuva[1], Khansa Fathima[2], M. R. Sai Dileep[1], Shreema S. Rao[1], Sanjay R[1], Sai Pattabhiram L[1]GVW Technologies | [2]JSS AHER

Background

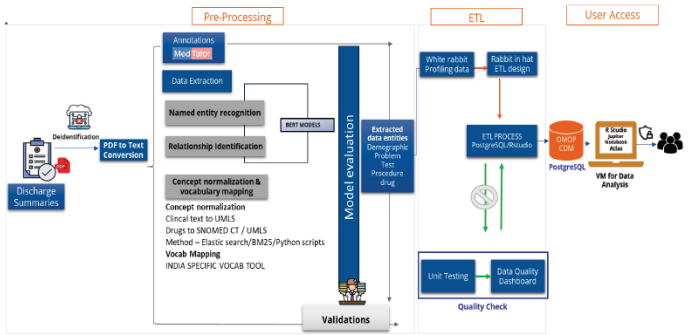
India's healthcare system is digitizing, but much of the data remains unstructured, fragmented, and not research-ready, limiting real-world evidence generation. To address this, We initiated the Breast Cancer Study to demonstrate how NLP can extract unstructured clinical data (e.g., discharge summaries) and standardize it into OMOP CDM for scalable research.



Challenges:

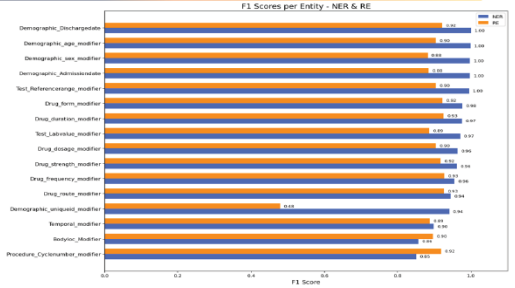
- ❑ Privacy laws
- ❑ Infra needs
- ❑ Lean team
- ❑ Manual annotation
- ❑ Drug mapping

Methods



Results

OMOP CDM tables extracted
Person
observation
observation_period
visit_occurrence
condition_occurrence
drug_exposure
procedure_occurrence
measurement
death
Care_site
note



	NER			RE		
	Precision	Recall	F1	Precision	Recall	F1
MACRO-AVERAGE	0.848	0.857	0.849	0.846	0.865	0.852
MICRO-AVERAGE	0.94	0.94	0.94	0.932	0.900	0.916

Conclusions

This initiative underscores the need for regional customization and cross-standard collaboration to enable inclusive, scalable, and reproducible research in India.
By aligning OHDSI and FHIR, we lay the foundation for robust evidence generation, capacity building, and global-standard health informatics in resource-limited settings.

Contact: contact@ohdsi.org

#OHDSISocialShowcase This Week

Friday

Comparing Timeline and Challenges of OMOP CDM Implementation in Brazil

(Juliana Araújo Prata de Faria, Danilo Luis Cerqueira Dias, Valentina Martufi, Julio Barbour Oliveira, Ricardo Felix Monteiro Neto, Karine Brito Beck da Silva Magalhães, Roberto Perez Carreiro, Maurício L. Barreto, Elzo Pereira Pinto Junior, Pablo Ivan Pereira Ramos)

Comparing Timeline and Challenges of OMOP CDM Implementation in Brazil

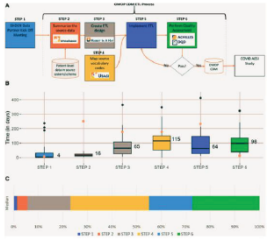
PRESENTER: **Valentina Martufi**

INTRO

The Data Interoperability and Federated Analyses (IDAF) group at the Centre for the Integration of Data and Knowledge for Health (CIDACS/Fiocruz-Bahia) applied the workflow proposed by the EHDEN to support the standardization into the OMOP CDM of Brazilian administrative data related to gestational syphilis. Comparative analysis of the duration (days) of the stages of the ETL process between the Brazilian and European experiences, highlighting the challenges faced, the lessons learned, and the solutions adopted in the local context.

METHODS

The workflow followed to implement ETL in the Brazilian context was adapted from the EHDEN and OHDSI Community methodological guide, including the following steps:



Source: adapted from Voss, E. A., Blackster, C., van Sandijk, S., Moinat, M., Kalliefz, M., van Speybrouck, M., Prieto-Alhambra, D., Schuemie, M. J., & Rijnbeek, P. R. (2024). European Health Data & Evidence Network--learnings from building out a standardized international health data network. Journal of the American Medical Informatics Association, 31(1), 209–219.

*Orange dots mark CIDACS' experience

RESULTS

The ETL process lasted 443 days, more than the EHDEN network average of 358 days but less than the maximum time observed of 622 days.

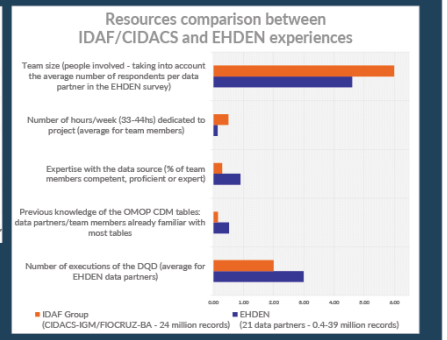
DQD: 1,363 standard, automated validation tests, only 37 (2.71%) resulted in failures [see poster 153 for more tech details!].

The comparison of execution times between Brazil and the EHDEN institutions shows not only the **contextual and operational disparities**, but also the **ability** of the Brazilian team to **adapt** and **learn** when applying **good international practices**.

		ACTIVITY DESCRIPTION	EHDEN (21 Data Partners) median length in days for each step across all data partners	BRAZIL (group IDAF CIDACS/Fiocruz-BA) length in days for each step
OMOP CDM ETL development process	STEP 1	Project Kick-off Meeting and team composition	4	6
	STEP 2	Summarize the source data (WhiteRabbit)	16	251
	STEP 3 AND STEP 4	Create ETL design (Rabbit in a Hat) and Map source vocabulary codes	180	179
	STEP 5	Implement ETL	64	223
	STEP 6	Perform Quality Assessment (DQD)	98	13

Figure 1: The figure shows a comparison between the European network and the IDAF group in terms of the duration, in days, of the stages in the ETL development process.

Compares the duration, in days, of ETL process stages between the EHDEN and the IDAF group, highlighting differences in OMOP CDM expertise, knowledge of source data, data volume and size, project dedication, and number of DQD iterations.



Compares corresponding resources available

Although the **Brazilian process was challenging** and took **longer** in some stages, it is **technically feasible** and can generate products **compatible with international standards**. The adaptations made and the **lessons learned** reinforce the **applicability** of the **OMOP CDM** in the **Global South***, even within a context of **restricted resources**.

The Global South is vast and highly heterogeneous! Brazil has similar resource constraints to other countries, but a very peculiar and unique data ecosystem, with a differential breadth and depth.

Take a picture to download our ETL documentation

Acknowledgements:
This work was funded by the Gates Foundation

Contact email: valentina.martufi@fiocruz.br

cidacs Centro de Integração de Dados e Conhecimentos para Saúde

FIOCRUZ | Bahia

AMMO BAR

The project was approved by the Fiocruz Research Ethics Committee - approval No. 7.045.566/2024 (CAAE 82529724.6.0000.0040), following national research ethics guidelines. Furthermore, the institution's data governance frameworks were respected, with the use of a secure environment and anonymized data to protect individual privacy.

CHALLENGES

Challenges faced: mappeability of vocabularies from the Brazilian context, the need for technical training, predominantly clinical nature of the model, which required adaptations to accommodate CIDACS' predominantly registry data. Additionally, our rich socioeconomic variables all ended up in the Observations table, undermining the efficiency of the ETL for this table.

Another inherent challenge is the sustainability of OMOPping interventions, strongly dependent on the availability of dedicated project funding, within the context of restricted resources of the Global South.

On the other hand, the adoption of the EHDEN workflow, the use of open-source OHDSI tools, the application of good knowledge management from the start of the project by recording decisions, changes, experience reports, documentation and sharing stand out as facilitators.

The findings contribute to strengthening the reference model of the EHDEN network and provide relevant input for other data standardization initiatives in countries of the Global South.

Juliana Araújo Prata de Faria, Danilo Luis Cerqueira Dias, Valentina Martufi, Julio Barbour Oliveira, Ricardo Felix Monteiro Neto, Karine Brito Beck da Silva Magalhães, Roberto Perez Carreiro, Maurício L. Barreto, Elzo Pereira Pinto Junior, Pablo Ivan Pereira Ramos

UNIVERSITY OF NORTH STATE

OHDSI

Where Are We Going?

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

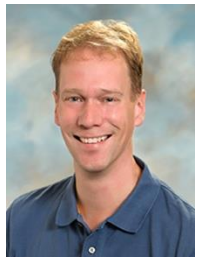
Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

June 2: LLM Research in the OHDSI Community



Martijn Schuemie

Johnson & Johnson / GenAI Workgroup Lead

Intro to OHDSI research in LLMs, GenAI WG updates

Anna Ostropolets

Johnson & Johnson / Columbia University

Ariadne: Automated vocabulary mappings using AI *(with Martijn Schuemie)*

Dmytry Dymshyts

Johnson & Johnson

LLM–Based Classification of ICD-10-CM to SNOMED Mappings for Improved Semantic Fidelity in OHDSI

Roger Carlson

Corewell Health

ACQIRE: Extending the OMOP Lifecycle

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