Vocabulary Refresh and Phenotype Phebruary Review

OHDSI Community Call
March 4, 2025 • 11 am ET

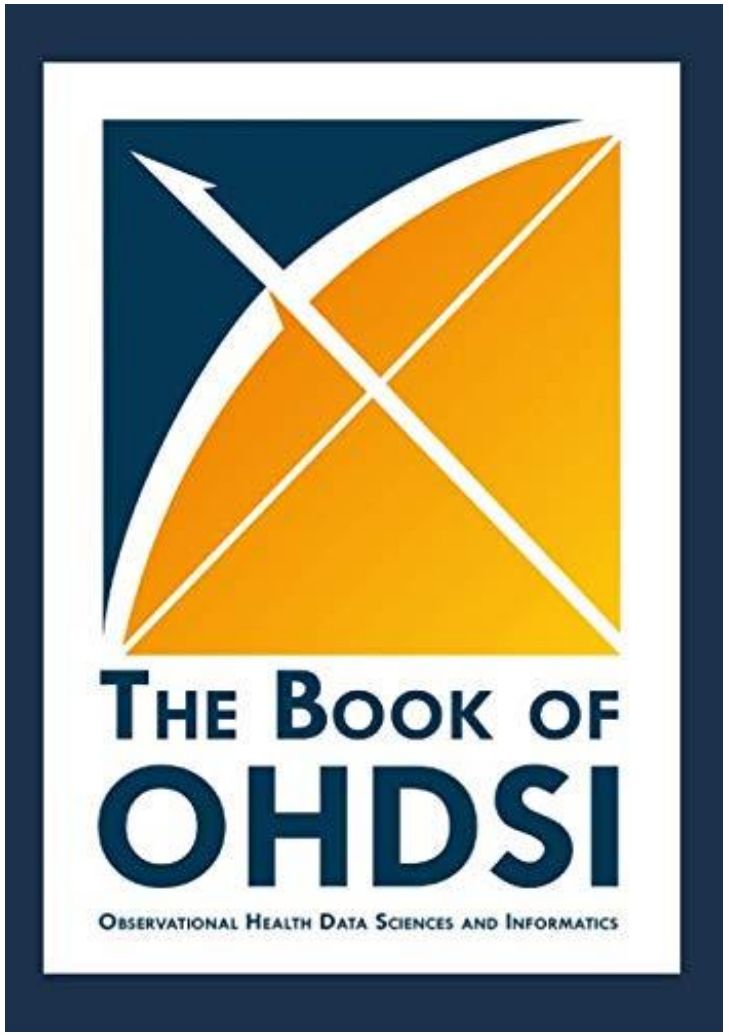
Upcoming Community Calls

Date	Topic
Mar. 4	Vocabulary Release Update, Winter 2025
Mar. 11	Book of OHDSI 2.0 Brainstorm and Planning Session
Mar. 18	OHDSI Evidence Network and Data Diagnostics Design
Mar. 25	Methods for Evaluating Data Fitness for Use
Apr. 1	Recent OHDSI Publications

March 11 Community Call

Christian Reich and Sarah Seager are leading the community efforts to publish a **2nd edition of the Book of OHDSI**.

Please join our March 11 call and take part inn a community brainstorm about what should be included in this edition, and how to get it done.



Three Stages of The Journey

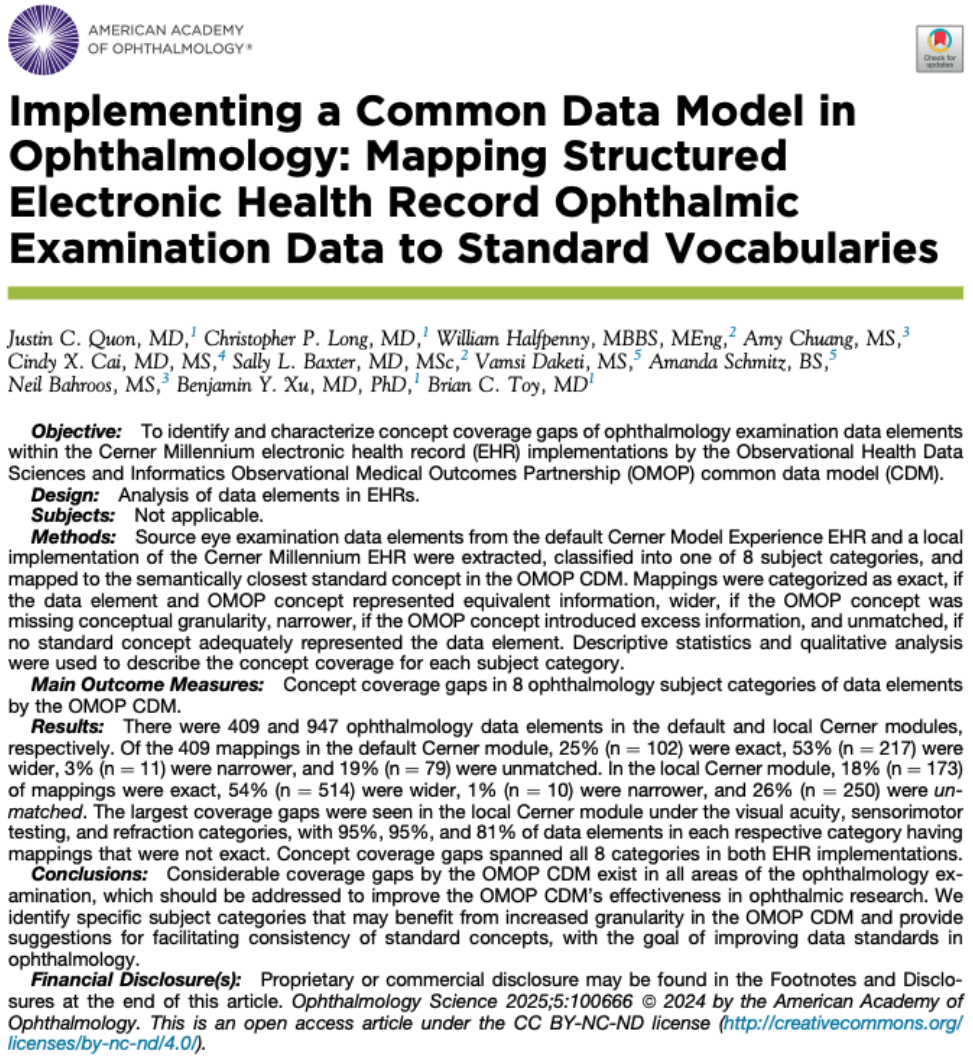
Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Justin Quon, Christopher Long, William Halfpenny, Amy Chuang, Cindy Cai, Sally Baxter, Vamsi Daketi, Amanda Schmitz, Neil Bahroos, Benjamin Xu, and Brian Toy** on the publication of **Implementing a Common Data Model in Ophthalmology: Mapping Structured Electronic Health Record Ophthalmic Examination Data to Standard Vocabularies** in *Ophthalmology Science*.



AMERICAN ACADEMY OF OPHTHALMOLOGY®

Implementing a Common Data Model in Ophthalmology: Mapping Structured Electronic Health Record Ophthalmic Examination Data to Standard Vocabularies

Justin C. Quon, MD,[1] Christopher P. Long, MD,[1] William Halfpenny, MBBS, MEng,[2] Amy Chuang, MS,[3] Cindy X. Cai, MD, MS,[4] Sally L. Baxter, MD, MSc,[2] Vamsi Daketi, MS,[5] Amanda Schmitz, BS,[5] Neil Bahroos, MS,[3] Benjamin Y. Xu, MD, PhD,[1] Brian C. Toy, MD[1]

Objective: To identify and characterize concept coverage gaps of ophthalmology examination data elements within the Cerner Millennium electronic health record (EHR) implementations by the Observational Health Data Sciences and Informatics Observational Medical Outcomes Partnership (OMOP) common data model (CDM).

Design: Analysis of data elements in EHRs.

Subjects: Not applicable.

Methods: Source eye examination data elements from the default Cerner Model Experience EHR and a local implementation of the Cerner Millennium EHR were extracted, classified into one of 8 subject categories, and mapped to the semantically closest standard concept in the OMOP CDM. Mappings were categorized as exact, if the data element and OMOP concept represented equivalent information, wider, if the OMOP concept was missing conceptual granularity, narrower, if the OMOP concept introduced excess information, and unmatched, if no standard concept adequately represented the data element. Descriptive statistics and qualitative analysis were used to describe the concept coverage for each subject category.

Main Outcome Measures: Concept coverage gaps in 8 ophthalmology subject categories of data elements by the OMOP CDM.

Results: There were 409 and 947 ophthalmology data elements in the default and local Cerner modules, respectively. Of the 409 mappings in the default Cerner module, 25% (n = 102) were exact, 53% (n = 217) were wider, 3% (n = 11) were narrower, and 19% (n = 79) were unmatched. In the local Cerner module, 18% (n = 173) of mappings were exact, 54% (n = 514) were wider, 1% (n = 10) were narrower, and 26% (n = 250) were *unmatched*. The largest coverage gaps were seen in the local Cerner module under the visual acuity, sensorimotor testing, and refraction categories, with 95%, 95%, and 81% of data elements in each respective category having mappings that were not exact. Concept coverage gaps spanned all 8 categories in both EHR implementations.

Conclusions: Considerable coverage gaps by the OMOP CDM exist in all areas of the ophthalmology examination, which should be addressed to improve the OMOP CDM's effectiveness in ophthalmic research. We identify specific subject categories that may benefit from increased granularity in the OMOP CDM and provide suggestions for facilitating consistency of standard concepts, with the goal of improving data standards in ophthalmology.

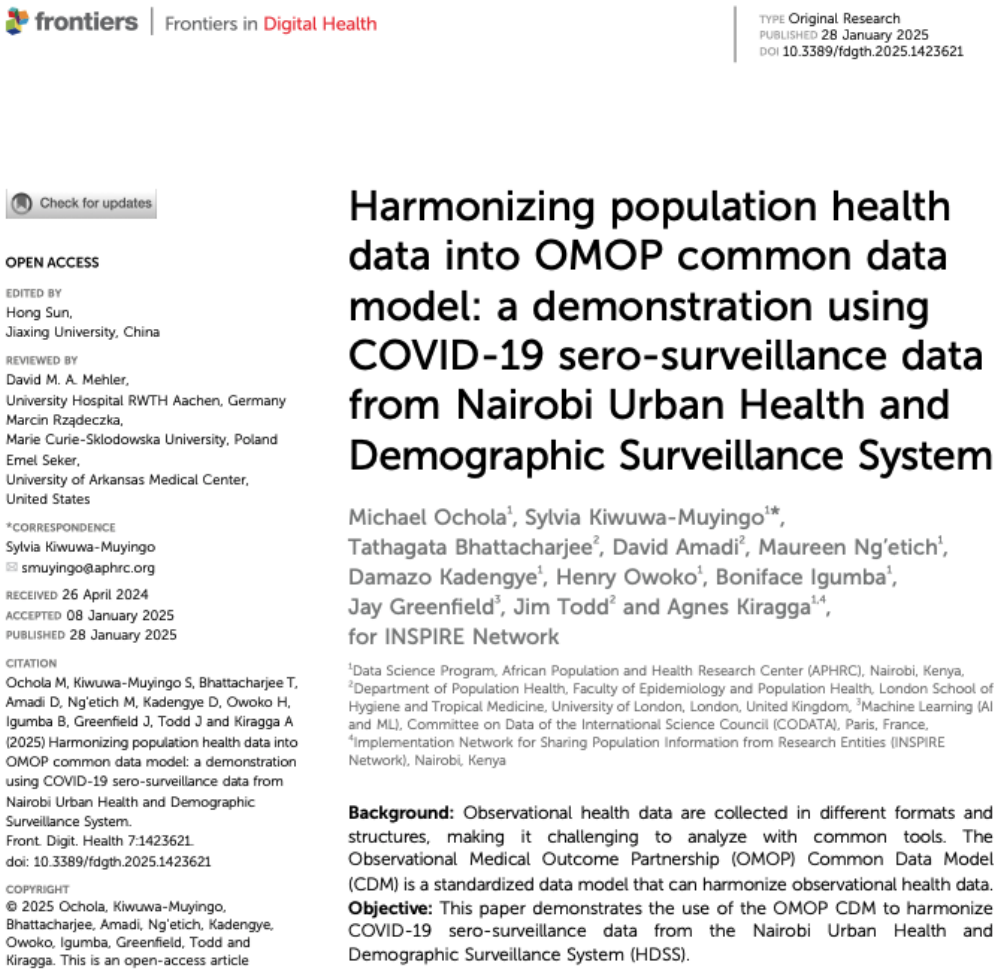
Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article.

OHDSI Shoutouts!

Congratulations to the team of **Aurora Quaye, John DiPalazzo, Kristin Kostka, Janelle Richard, Blaire Beers-Mulroy, Meredith Peck, Robert Krulee, and Yi Zhang** on the publication of **Identifying factors associated with persistent opioid use after total joint arthroplasty: a retrospective review** in *Pain Medicine.*

OHDSI Shoutouts!

Congratulations to the team of **Michael Ochola, Sylvia Kiwuwa-Muyingo, Tathagata Bhattacharjee, David Amadi, Maureen Ng'etich, Damazo Kadengye, Henry Owoko, Boniface Igumba, Jay Greenfield, Jim Todd, and Agnes Kiragga** on the publication of **Harmonizing population health data into OMOP common data model: a demonstration using COVID-19 sero-surveillance data from Nairobi Urban Health and Demographic Surveillance System** in *Frontiers in Digital Health.*



frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 28 January 2025
DOI 10.3389/fdgth.2025.1423621

Check for updates

OPEN ACCESS

EDITED BY
Hong Sun,
Jiaxing University, China

REVIEWED BY
David M. A. Mehler,
University Hospital RWTH Aachen, Germany
Marcin Rządeczka,
Marie Curie-Sklodowska University, Poland
Emel Seker,
University of Arkansas Medical Center, United States

*CORRESPONDENCE
Sylvia Kiwuwa-Muyingo
smuyingo@aphrc.org

RECEIVED 26 April 2024
ACCEPTED 08 January 2025
PUBLISHED 28 January 2025

CITATION
Ochola M, Kiwuwa-Muyingo S, Bhattacharjee T, Amadi D, Ng'etich M, Kadengye D, Owoko H, Igumba B, Greenfield J, Todd J and Kiragga A (2025) Harmonizing population health data into OMOP common data model: a demonstration using COVID-19 sero-surveillance data from Nairobi Urban Health and Demographic Surveillance System.
Front. Digit. Health 7:1423621.
doi: 10.3389/fdgth.2025.1423621

COPYRIGHT
© 2025 Ochola, Kiwuwa-Muyingo, Bhattacharjee, Amadi, Ng'etich, Kadengye, Owoko, Igumba, Greenfield, Todd and Kiragga. This is an open-access article

Harmonizing population health data into OMOP common data model: a demonstration using COVID-19 sero-surveillance data from Nairobi Urban Health and Demographic Surveillance System

Michael Ochola[1], Sylvia Kiwuwa-Muyingo[1]*, Tathagata Bhattacharjee[2], David Amadi[2], Maureen Ng'etich[1], Damazo Kadengye[1], Henry Owoko[1], Boniface Igumba[1], Jay Greenfield[3], Jim Todd[2] and Agnes Kiragga[1,4], for INSPIRE Network

[1]Data Science Program, African Population and Health Research Center (APHRC), Nairobi, Kenya, [2]Department of Population Health, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, University of London, London, United Kingdom, [3]Machine Learning (AI and ML), Committee on Data of the International Science Council (CODATA), Paris, France, [4]Implementation Network for Sharing Population Information from Research Entities (INSPIRE Network), Nairobi, Kenya

Background: Observational health data are collected in different formats and structures, making it challenging to analyze with common tools. The Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM) is a standardized data model that can harmonize observational health data.
Objective: This paper demonstrates the use of the OMOP CDM to harmonize COVID-19 sero-surveillance data from the Nairobi Urban Health and Demographic Surveillance System (HDSS).

OHDSI Shoutouts!

Congratulations to the team of **Kevin Ouazzani, Xavier Ansolabehere, Florence Journeau, Alexandre Vidal, Nicolas Jaubourg, Maxime Doublet, Raphael Thollot, Arnaud Fabre, and Nicolas Glatt** on the publication of **Project Victoria: A pragmatic data model to automate RWE generation from the national French claims database** in the *Health Informatics Journal.*

Health Informatics Journal
Volume 31, Issue 1, January 2025

https://doi.org/10.1177/14604582251318250

Sage Journals

Research Article

Project Victoria: A pragmatic data model to automate RWE generation from the national French claims database

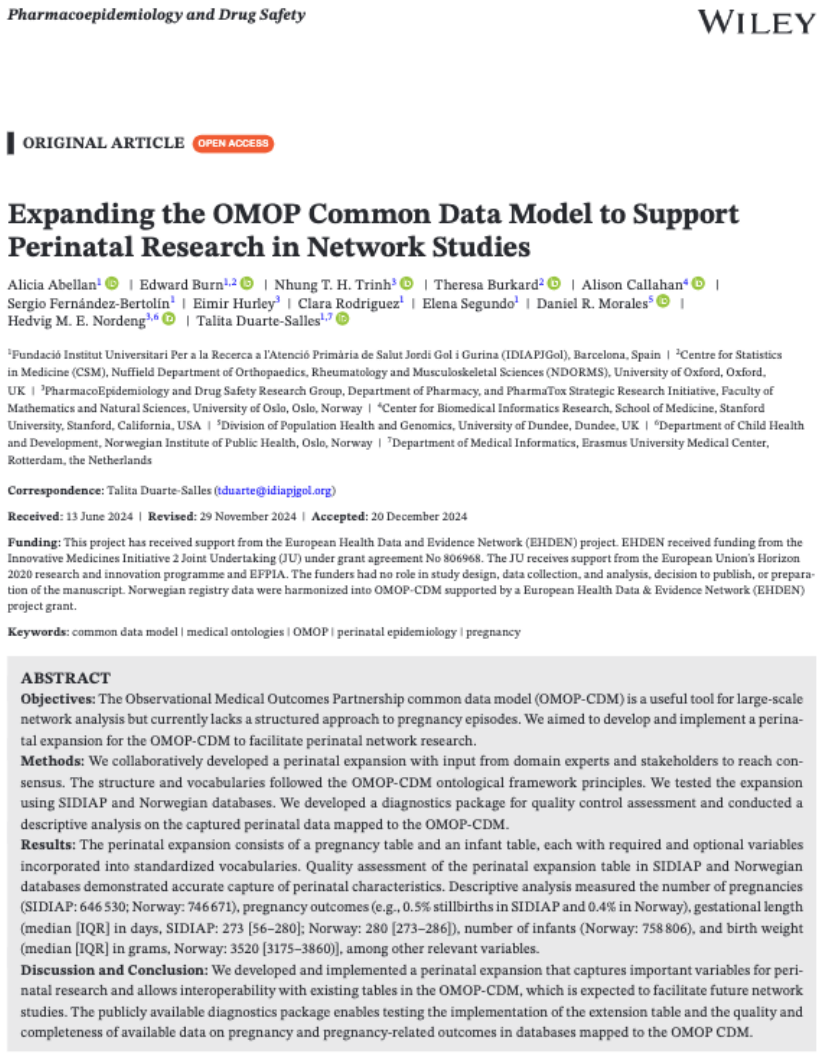
Kevin Ouazzani, Xavier Ansolabehere, Florence Journeau, Alexandre Vidal, Nicolas Jaubourg, Maxime Doublet, Raphael Thollot, Arnaud Fabre, and Nicolas Glatt

Abstract

Objective: This paper describes Victoria, an empirically built data pipeline for SNDS to: - Build an automated, scalable pipeline supporting changes to the data model inherent to the use of large databases, - Deliver a documented pipeline with clear processes, enabling scientific, epidemiological researches, - Ease access to SNDS data in compliance with regulatory requirements. **Methods:** This paper describes the 2-steps process of the Victoria pipeline and its final output. The initial cleaning step consists in formatting, deleting empty, error or duplicate records and renaming variables without changing their values, accordingly with the official SNDS documentation. The second step consists in creating 2 linearised data models: every line of each table is an event, and each table is indexed with a unique patient identifier, without the need for a central patient or identifier table. These 2 models are: - the epidemiological model, used for answering most of the research questions requiring population phenotyping (demography, diagnosis, procedures characteristics). - the medico-economic model is used for costs and healthcare consumption analyses. It contains more complex information about reimbursements rates and the data quality assessment is focused on costs rather than medico-administrative information. **Results:** The pipeline was executed on 2 different datasets representing ~85 000 and ~870 000 beneficiaries with the following configuration: one master with 4 cores and 16Go of RAM and respectively 4 and 6 workers. The total execution time for the smaller dataset was 25 h and 96 h for the larger one. The longest part of those times is represented by the format conversion to parquet. The cleaning step took only 4 h in both cases. The epidemiological model took 344 min for the smaller dataset and 1934 min for the larger one. The medico-economic model took the longest time with 704 min and 2145 min, respectively. **Conclusion:** Victoria pipeline is a successfully implemented SNDS pipeline. Compared to previous pipelines, reviewability is part of its design as unit tests and quality assessments can natively be developed to ensure data and analysis quality. The pipeline has been used for 2 published studies. The recent work toward OMOP conversion will be integrated in upcoming versions and, as Victoria is set to run on a CD platform, the potential evolution if SNDS format can be considered.

OHDSI Shoutouts!

Congratulations to the team of **Alicia Abellan, Edward Burn, Nhung T. H. Trinh, Theresa Burkard, Alison Callahan, Sergio Fernández-Bertolín, Eimir Hurley, Clara Rodriguez, Elena Segundo, Daniel R. Morales, Hedvig M. E. Nordeng, and Talita Duarte-Salles** on the publication of **Expanding the OMOP Common Data Model to Support Perinatal Research in Network Studies** in *Pharmacoepidemiology & Drug Safety.*



Pharmacoepidemiology and Drug Safety

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Expanding the OMOP Common Data Model to Support Perinatal Research in Network Studies

Alicia Abellan[1] | Edward Burn[1,2] | Nhung T. H. Trinh[3] | Theresa Burkard[2] | Alison Callahan[4] | Sergio Fernández-Bertolín[1] | Eimir Hurley[3] | Clara Rodriguez[1] | Elena Segundo[1] | Daniel R. Morales[5] | Hedvig M. E. Nordeng[3,6] | Talita Duarte-Salles[1,7]

[1]Fundació Institut Universitari Per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain | [2]Centre for Statistics in Medicine (CSM), Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), University of Oxford, Oxford, UK | [3]PharmacoEpidemiology and Drug Safety Research Group, Department of Pharmacy, and PharmaTox Strategic Research Initiative, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway | [4]Center for Biomedical Informatics Research, School of Medicine, Stanford University, Stanford, California, USA | [5]Division of Population Health and Genomics, University of Dundee, Dundee, UK | [6]Department of Child Health and Development, Norwegian Institute of Public Health, Oslo, Norway | [7]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands

Correspondence: Talita Duarte-Salles (tduarte@idiapjgol.org)

Received: 13 June 2024 | **Revised:** 29 November 2024 | **Accepted:** 20 December 2024

Funding: This project has received support from the European Health Data and Evidence Network (EHDEN) project. EHDEN received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript. Norwegian registry data were harmonized into OMOP-CDM supported by a European Health Data & Evidence Network (EHDEN) project grant.

Keywords: common data model | medical ontologies | OMOP | perinatal epidemiology | pregnancy

ABSTRACT

Objectives: The Observational Medical Outcomes Partnership common data model (OMOP-CDM) is a useful tool for large-scale network analysis but currently lacks a structured approach to pregnancy episodes. We aimed to develop and implement a perinatal expansion for the OMOP-CDM to facilitate perinatal network research.

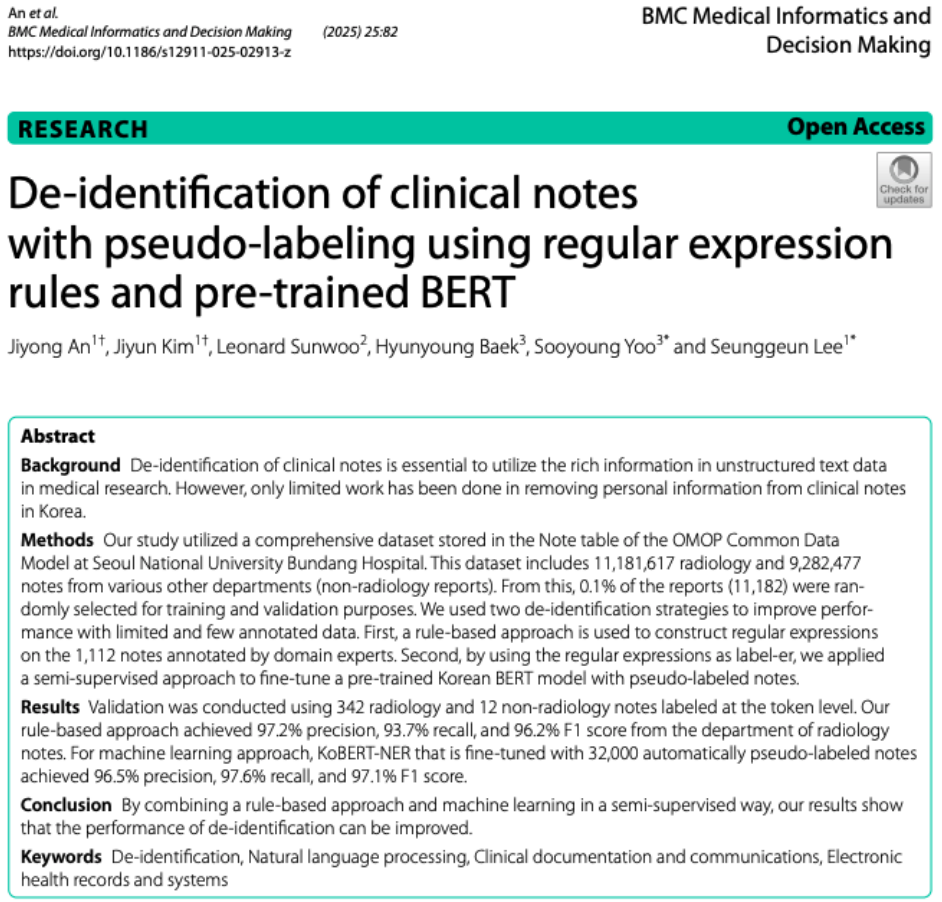
Methods: We collaboratively developed a perinatal expansion with input from domain experts and stakeholders to reach consensus. The structure and vocabularies followed the OMOP-CDM ontological framework principles. We tested the expansion using SIDIAP and Norwegian databases. We developed a diagnostics package for quality control assessment and conducted a descriptive analysis on the captured perinatal data mapped to the OMOP-CDM.

Results: The perinatal expansion consists of a pregnancy table and an infant table, each with required and optional variables incorporated into standardized vocabularies. Quality assessment of the perinatal expansion table in SIDIAP and Norwegian databases demonstrated accurate capture of perinatal characteristics. Descriptive analysis measured the number of pregnancies (SIDIAP: 646 530; Norway: 746 671), pregnancy outcomes (e.g., 0.5% stillbirths in SIDIAP and 0.4% in Norway), gestational length (median [IQR] in days, SIDIAP: 273 [56–280]; Norway: 280 [273–286]), number of infants (Norway: 758 806), and birth weight (median [IQR] in grams, Norway: 3520 [3175–3860]), among other relevant variables.

Discussion and Conclusion: We developed and implemented a perinatal expansion that captures important variables for perinatal research and allows interoperability with existing tables in the OMOP-CDM, which is expected to facilitate future network studies. The publicly available diagnostics package enables testing the implementation of the extension table and the quality and completeness of available data on pregnancy and pregnancy-related outcomes in databases mapped to the OMOP CDM.

OHDSI Shoutouts!

Congratulations to the team of **Jiyong An, Jiyun Kim, Leonard Sunwoo, Hyunyoung Baek, Sooyoung Yoo & Seunggeun Lee** on the publication of **De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT** in *BMC Medical Informatics and Decision Making.*



An *et al.*
BMC Medical Informatics and Decision Making (2025) 25:82
https://doi.org/10.1186/s12911-025-02913-z

BMC Medical Informatics and Decision Making

RESEARCH Open Access

De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT

Jiyong An[1†], Jiyun Kim[1†], Leonard Sunwoo[2], Hyunyoung Baek[3], Sooyoung Yoo[3*] and Seunggeun Lee[1*]

Abstract

Background De-identification of clinical notes is essential to utilize the rich information in unstructured text data in medical research. However, only limited work has been done in removing personal information from clinical notes in Korea.

Methods Our study utilized a comprehensive dataset stored in the Note table of the OMOP Common Data Model at Seoul National University Bundang Hospital. This dataset includes 11,181,617 radiology and 9,282,477 notes from various other departments (non-radiology reports). From this, 0.1% of the reports (11,182) were randomly selected for training and validation purposes. We used two de-identification strategies to improve performance with limited and few annotated data. First, a rule-based approach is used to construct regular expressions on the 1,112 notes annotated by domain experts. Second, by using the regular expressions as label-er, we applied a semi-supervised approach to fine-tune a pre-trained Korean BERT model with pseudo-labeled notes.

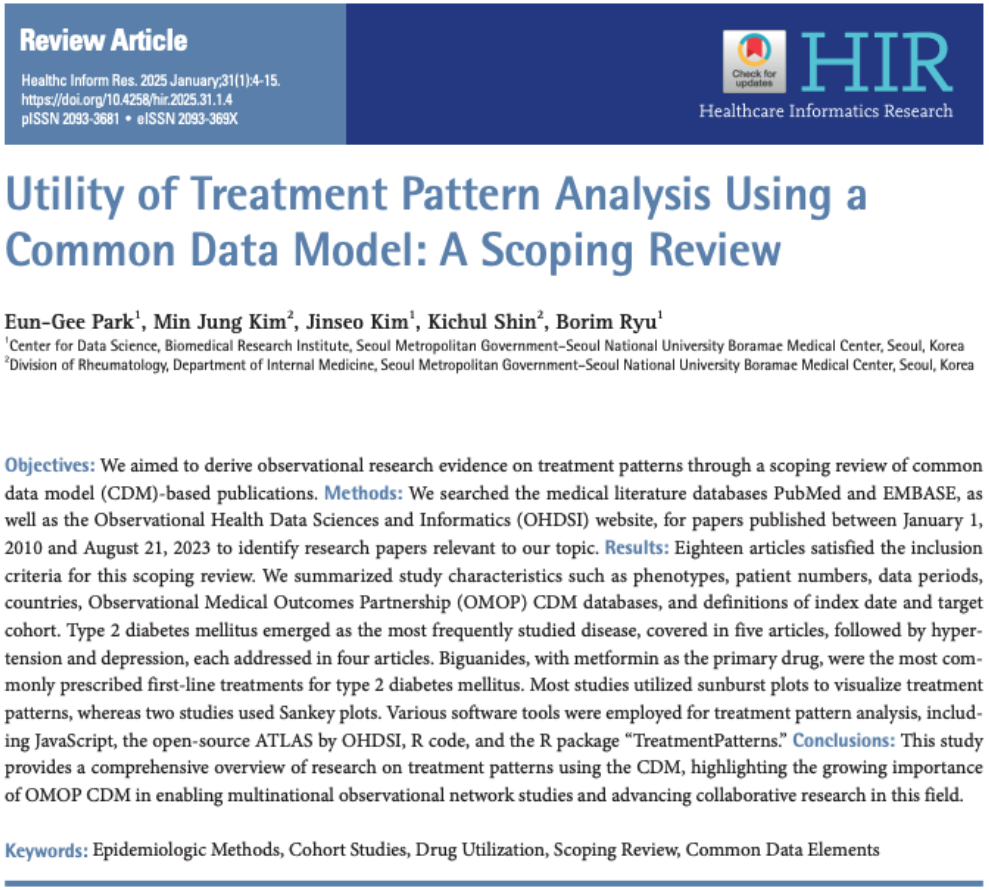
Results Validation was conducted using 342 radiology and 12 non-radiology notes labeled at the token level. Our rule-based approach achieved 97.2% precision, 93.7% recall, and 96.2% F1 score from the department of radiology notes. For machine learning approach, KoBERT-NER that is fine-tuned with 32,000 automatically pseudo-labeled notes achieved 96.5% precision, 97.6% recall, and 97.1% F1 score.

Conclusion By combining a rule-based approach and machine learning in a semi-supervised way, our results show that the performance of de-identification can be improved.

Keywords De-identification, Natural language processing, Clinical documentation and communications, Electronic health records and systems

OHDSI Shoutouts!

Congratulations to the team of **Eun-Gee Park, Min Jung Kim, Jinseo Kim, Kichul Shin, and Borim Ryu** on the publication of **Utility of Treatment Pattern Analysis Using a Common Data Model: A Scoping Review** in *Healthcare Informatics Research.*



Review Article
Healthc Inform Res. 2025 January;31(1):4-15.
https://doi.org/10.4258/hir.2025.31.1.4
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Utility of Treatment Pattern Analysis Using a Common Data Model: A Scoping Review

Eun-Gee Park[1], Min Jung Kim[2], Jinseo Kim[1], Kichul Shin[2], Borim Ryu[1]
[1]Center for Data Science, Biomedical Research Institute, Seoul Metropolitan Government–Seoul National University Boramae Medical Center, Seoul, Korea
[2]Division of Rheumatology, Department of Internal Medicine, Seoul Metropolitan Government–Seoul National University Boramae Medical Center, Seoul, Korea

Objectives: We aimed to derive observational research evidence on treatment patterns through a scoping review of common data model (CDM)-based publications. **Methods:** We searched the medical literature databases PubMed and EMBASE, as well as the Observational Health Data Sciences and Informatics (OHDSI) website, for papers published between January 1, 2010 and August 21, 2023 to identify research papers relevant to our topic. **Results:** Eighteen articles satisfied the inclusion criteria for this scoping review. We summarized study characteristics such as phenotypes, patient numbers, data periods, countries, Observational Medical Outcomes Partnership (OMOP) CDM databases, and definitions of index date and target cohort. Type 2 diabetes mellitus emerged as the most frequently studied disease, covered in five articles, followed by hypertension and depression, each addressed in four articles. Biguanides, with metformin as the primary drug, were the most commonly prescribed first-line treatments for type 2 diabetes mellitus. Most studies utilized sunburst plots to visualize treatment patterns, whereas two studies used Sankey plots. Various software tools were employed for treatment pattern analysis, including JavaScript, the open-source ATLAS by OHDSI, R code, and the R package "TreatmentPatterns." **Conclusions:** This study provides a comprehensive overview of research on treatment patterns using the CDM, highlighting the growing importance of OMOP CDM in enabling multinational observational network studies and advancing collaborative research in this field.

Keywords: Epidemiologic Methods, Cohort Studies, Drug Utilization, Scoping Review, Common Data Elements

OHDSI Shoutouts!

Congratulations to the team of **Chen Yanover, Ramit Magen-Rimon, Erica A. Voss, Joel Swerdel, Anna Sheahan, Nathan Hall, Jimyung Park, Rae Woong Park, Kwang Jae Lee, Sung Jae Shin, Seung In Seo, Kyung-Joo Lee, Thomas Falconer, Leonard Haas, Paul Nagy, Mary Grace Bowring, Michael Cook, Steven Miller, Tal El-Hay, Maytal Bivas-Benita, Pinchas Akiva, Yehuda Chowers & Roni Weisshof** on the publication of **Characteristics and Outcomes of Over a Million Patients with Inflammatory Bowel Disease in Seven Countries: Multinational Cohort Study and Open Data Resource** in *Digestive Diseases and Sciences.*

Digestive Diseases and Sciences (2025) 70:709–718
https://doi.org/10.1007/s10620-024-08787-x

ORIGINAL ARTICLE

Check for updates

Characteristics and Outcomes of Over a Million Patients with Inflammatory Bowel Disease in Seven Countries: Multinational Cohort Study and Open Data Resource

Chen Yanover[1,2] · Ramit Magen-Rimon[1,3] · Erica A. Voss[1,4] · Joel Swerdel[1,4] · Anna Sheahan[1,4] · Nathan Hall[1,4] · Jimyung Park[1,5,6] · Rae Woong Park[1,6] · Kwang Jae Lee[1,7] · Sung Jae Shin[1,7] · Seung In Seo[1,8] · Kyung-Joo Lee[1,9] · Thomas Falconer[1,5] · Leonard Haas[1,10] · Paul Nagy[1,10] · Mary Grace Bowring[1,10] · Michael Cook[1,10] · Steven Miller[1,10] · Tal El-Hay[1,2] · Maytal Bivas-Benita[1,2] · Pinchas Akiva[1,2] · Yehuda Chowers[1,11] · Roni Weisshof[1,11]

Received: 10 September 2024 / Accepted: 4 December 2024 / Published online: 26 December 2024

Abstract

Background and Aims Observational healthcare data are an important tool for delineating patients' inflammatory bowel disease (IBD) journey in real-world settings. However, studies that characterize IBD cohorts typically rely on a single resource, apply diverse eligibility criteria, and extract variable sets of attributes, making comparison between cohorts challenging. We aim to longitudinally describe and compare IBD patient cohorts across multiple geographic regions, employing unified data and analysis framework.

Methods We conducted a descriptive cohort study, using routinely collected healthcare data, from a federated network of data partners in sixteen databases from seven countries (USA, UK, France, Germany, Japan, Korea, and Australia); and computed the prevalence of thousands of attributes, across multiple baseline and follow-up time windows, for full disease cohorts and various strata.

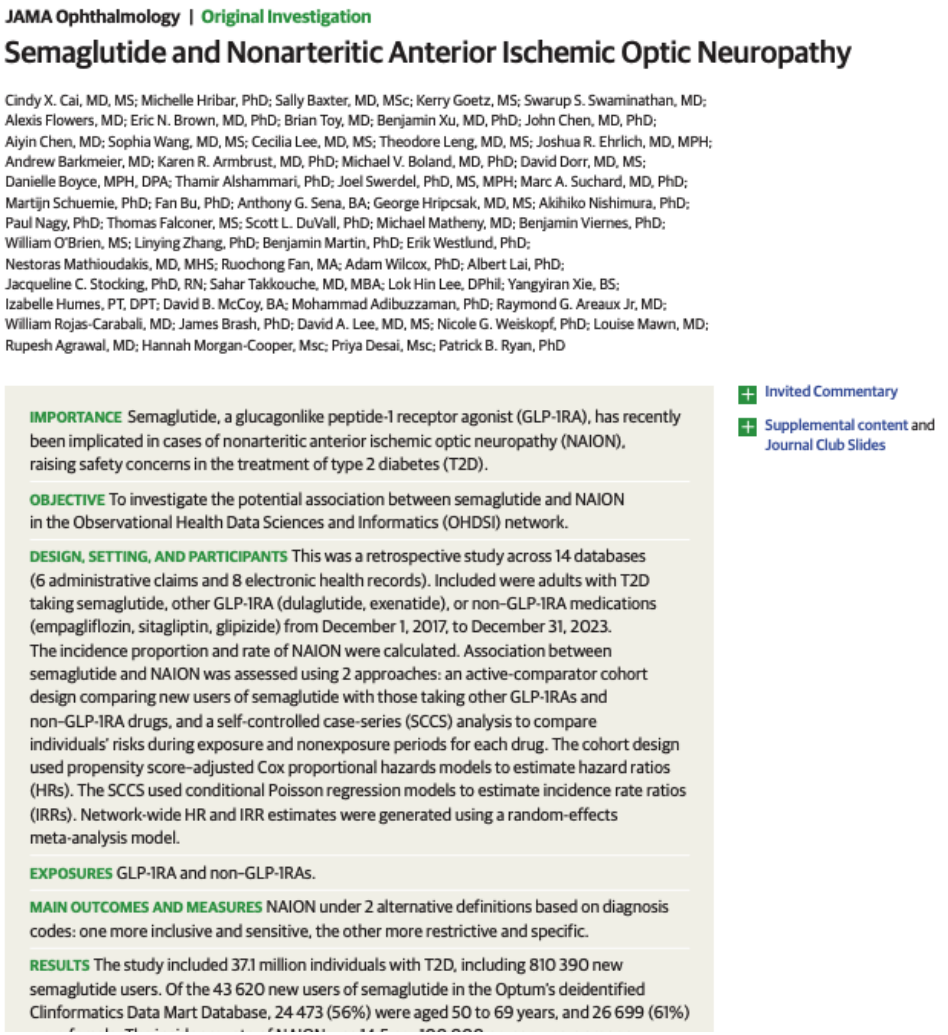
Results Characterizing the disease trajectory of 462,502 Crohn's disease (CD) and 589,118 ulcerative colitis (UC) subjects, we observed a decline over time in the average age at CD diagnosis in Europe and North America but less pronounced shifts in Japan and Korea; an uptick in the proportion of patients with anxiety diagnosis prior to CD diagnosis in European and US datasets; and stable rates of segmental colonic and small bowel resections within one and three years following UC and CD diagnosis, respectively, in most US databases.

Conclusions The study provides a comprehensive characterization of IBD patient cohorts from various countries including insights into disease trends, demographics, and pre-diagnosis symptoms. All characteristics and outcomes are publicly available, providing an unprecedented, comprehensive open resource for clinicians and researchers.

Keywords Crohn's disease · Ulcerative colitis · Routinely collected health data · Cohort study

OHDSI Shoutouts!

Congratulations to the team of Cindy Cai, Michelle Hribar, Sally Baxter, Kerry Goetz, Swarup S. Swaminathan, Alexis Flowers, Eric N. Brown, Brian Toy, Benjamin Xu, John Chen, Aiyin Chen, Sophia Wang, Cecilia Lee, Theodore Leng, Joshua R. Ehrlich, Andrew Barkmeier, Karen R. Armbrust, Michael V. Boland, David Dorr, Danielle Boyce, Thamir Alshammari, Joel Swerdel, Marc A. Suchard, Martijn Schuemie, Fan Bu, Anthony G. Sena, George Hripcsak, Akihiko Nishimura, Paul Nagy, Thomas Falconer, Scott L. DuVall, Michael Matheny, Benjamin Viernes, William O'Brien, Linying Zhang, Benjamin Martin, Erik Westlund, Nestoras Mathioudakis, Ruochong Fan, Adam Wilcox, Albert Lai, Jacqueline C. Stocking, Sahar Takkouche, Lok Hin Lee, Yangyiran Xie, Izabelle Humes, David B. McCoy, Mohammad Adibuzzaman, Raymond G. Areaux Jr, William Rojas-Carabali, James Brash, David A. Lee, Nicole G. Weiskopf, Louise Mawn, Rupesh Agrawal, Hannah Morgan-Cooper, Priya Desai, and Patrick Ryan on the publication of **Semaglutide and Nonarteritic Anterior Ischemic Optic Neuropathy** in *JAMA Ophthalmology.*



JAMA Ophthalmology | Original Investigation

Semaglutide and Nonarteritic Anterior Ischemic Optic Neuropathy

Cindy X. Cai, MD, MS; Michelle Hribar, PhD; Sally Baxter, MD, MSc; Kerry Goetz, MS; Swarup S. Swaminathan, MD; Alexis Flowers, MD; Eric N. Brown, MD, PhD; Brian Toy, MD; Benjamin Xu, MD, PhD; John Chen, MD, PhD; Aiyin Chen, MD; Sophia Wang, MD, MS; Cecilia Lee, MD, MS; Theodore Leng, MD, MS; Joshua R. Ehrlich, MD, MPH; Andrew Barkmeier, MD; Karen R. Armbrust, MD, PhD; Michael V. Boland, MD, PhD; David Dorr, MD, MS; Danielle Boyce, MPH, DPA; Thamir Alshammari, PhD; Joel Swerdel, PhD, MS, MPH; Marc A. Suchard, MD, PhD; Martijn Schuemie, PhD; Fan Bu, PhD; Anthony G. Sena, BA; George Hripcsak, MD, MS; Akihiko Nishimura, PhD; Paul Nagy, PhD; Thomas Falconer, MS; Scott L. DuVall, PhD; Michael Matheny, MD; Benjamin Viernes, PhD; William O'Brien, MS; Linying Zhang, PhD; Benjamin Martin, PhD; Erik Westlund, PhD; Nestoras Mathioudakis, MD, MHS; Ruochong Fan, MA; Adam Wilcox, PhD; Albert Lai, PhD; Jacqueline C. Stocking, PhD, RN; Sahar Takkouche, MD, MBA; Lok Hin Lee, DPhil; Yangyiran Xie, BS; Izabelle Humes, PT, DPT; David B. McCoy, BA; Mohammad Adibuzzaman, PhD; Raymond G. Areaux Jr, MD; William Rojas-Carabali, MD; James Brash, PhD; David A. Lee, MD, MS; Nicole G. Weiskopf, PhD; Louise Mawn, MD; Rupesh Agrawal, MD; Hannah Morgan-Cooper, Msc; Priya Desai, Msc; Patrick B. Ryan, PhD

IMPORTANCE Semaglutide, a glucagonlike peptide-1 receptor agonist (GLP-1RA), has recently been implicated in cases of nonarteritic anterior ischemic optic neuropathy (NAION), raising safety concerns in the treatment of type 2 diabetes (T2D).

OBJECTIVE To investigate the potential association between semaglutide and NAION in the Observational Health Data Sciences and Informatics (OHDSI) network.

DESIGN, SETTING, AND PARTICIPANTS This was a retrospective study across 14 databases (6 administrative claims and 8 electronic health records). Included were adults with T2D taking semaglutide, other GLP-1RA (dulaglutide, exenatide), or non-GLP-1RA medications (empagliflozin, sitagliptin, glipizide) from December 1, 2017, to December 31, 2023. The incidence proportion and rate of NAION were calculated. Association between semaglutide and NAION was assessed using 2 approaches: an active-comparator cohort design comparing new users of semaglutide with those taking other GLP-1RAs and non–GLP-1RA drugs, and a self-controlled case-series (SCCS) analysis to compare individuals' risks during exposure and nonexposure periods for each drug. The cohort design used propensity score–adjusted Cox proportional hazards models to estimate hazard ratios (HRs). The SCCS used conditional Poisson regression models to estimate incidence rate ratios (IRRs). Network-wide HR and IRR estimates were generated using a random-effects meta-analysis model.

EXPOSURES GLP-1RA and non–GLP-1RAs.

MAIN OUTCOMES AND MEASURES NAION under 2 alternative definitions based on diagnosis codes: one more inclusive and sensitive, the other more restrictive and specific.

RESULTS The study included 37.1 million individuals with T2D, including 810 390 new semaglutide users. Of the 43 620 new users of semaglutide in the Optum's deidentified Clinformatics Data Mart Database, 24 473 (56%) were aged 50 to 69 years, and 26 699 (61%)

Invited Commentary

Supplemental content and Journal Club Slides

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Tuesday	12 pm	Atlas
Wednesday	8 am	Psychiatry
Wednesday	7 pm	Medical Imaging
Thursday	8 am	Medical Devices
Thursday	10 am	Themis
Thursday	11 am	Industry
Thursday	12 pm	Methods Research
Thursday	1 pm	Oncology Vocabulary/Development
Thursday	2 pm	Early-Stage Researchers
Friday	10 am	Transplant
Friday	10 am	GIS - Geographic Information System
Friday	11:30 am	Steering
Monday	9 am	Vaccine Vocabulary
Monday	10 am	Healthcare Systems Interest Group

March Newsletter is Available

OHDSI

The Journey Newsletter (March 2025)

One community focus in 2025 is to generate reliable evidence that can fill the evidence gaps identified by clinical guidelines. That process continued through our fourth Phenotype Phebruary, when leads for our 14 network studies, our Phenotype workgroup, and other members of the community collaborated to develop and evaluate phenotypes for indications, exposures, and outcomes of interest. We also learned about the mission, achievements and 2025 goals for our 30+ workgroups, which provide a home for the talents and passions for each member of our global community. **#JoinTheJourney**

Podcast: Workgroups, Phenotypes, Next Steps

Workgroups and phenotypes were the community focus over the last month. In the latest On The Journey podcast, Patrick Ryan and Craig Sachson discuss the 2025 goals for workgroups and how they align with the overall community focus. Then they discuss the impact of the latest Phenotype Phebruary, the lessons the community can take from it, and how it fits in the guideline-driven evidence studies ongoing. *(if video does not appear, please click View this email in your browser)*

Community Updates

Where Have We Been?

- The Winter 2025 vocabulary refresh was released last week and includes several domain changes, newly added concepts, concept changes and more. Please join our March 4 community call for a full update on this recent refresh. Thank you to the vocabulary team for your hard work on this.
- Phenotype Phebruary provided an opportunity to develop phenotypes and make them analysis-ready for the guideline-driven network studies generated by our community. Check out the forum thread and our community calls page for updates and video demos, including one on concept set creation.
- Representatives from 30+ workgroups joined February community calls to share their respective objectives and key results for 2025. These presentations and the accompanying slides have been updated on our workgroups homepage.

Where Are We Now?

- March to Data Fitness is our theme this month as we work to build up the Evidence Network and determine which data partners are appropriate to generate evidence for our clinical guideline studies. Our March 18 community call will be focused on the Evidence Network and will include a mini tutorial on data diagnostics, while the March 25 community call will focus on methods for evaluating data fitness for use.
- Christian Reich and Sarah Seager are leading an effort to publish a second edition of the Book of OHDSI, which will include updates to previous text and new paragraphs/chapters. This work is taking place within the Education workgroup; if you would like to join this effort, please sign up here.
- The #OHDSISocialShowcase features posters, software demos and lightning talks from the 2024 Global Symposium. Please make sure you are following OHDSI's LinkedIn, Twitter/X and Instagram feeds to receive daily updates on the research presented by our community.

Where Are We Going?

- The OHDSI Global Symposium will be held Oct. 7-9 at the Hyatt Regency Hotel in New Brunswick, NJ, USA. Agenda and registration information will be shared when available.
- Registration is open for the Europe Symposium, which will be held July 5-7 in the "Old Prison" building of Hasselt University in Hasselt, Belgium. More information is available later in this newsletter.

Get To Know The OHDSI Workgroups

OHDSI has a central mission to improve health globally, but there are countless areas where our community can be of service. Work around data, methods, open-source tools, and clinical applications are all pieces of the puzzle. Within OHDSI, there are opportunities to work in—or learn from—any or many of these areas.

Leaders from over 30+ workgroups presented opportunities for all community members to find a home for their talents and passions. Newcomers and veterans can both make meaningful contributions to our community by collaborating in workgroups. Throughout February, workgroup representatives shared the mission, recent achievements and 2025 goals. You can find those presentations and see if there is a home for you on our workgroups homepage.

OHDSI Workgroups Homepage

Join A Workgroup

February Publications

Jones N, Shih MC, Healey E, Zhai CW, Advani S, Smith-McLallen A, Sontag D, Kanjilal S. Use of Machine Learning to Assess the Management of Uncomplicated Urinary Tract Infection. JAMA Netw Open. 2025 Jan 2;8(1):e2456950. doi: 10.1001/jamanetworkopen.2024.56950. PMID: 39888618; PMCID: PMC11786233.

Quon JC, Long CP, Halfpenny W, Chuang A, Cai CX, Baxter SL, Daketi V, Schmitz A, Bahroos N, Xu BY, Toy BC. Implementing a Common Data Model in Ophthalmology: Mapping Structured Electronic Health Record Ophthalmic Examination Data to Standard Vocabularies. Ophthalmol Sci. 2024 Nov 28;5(2):100666. doi: 10.1016/j.xops.2024.100666. PMID: 39896425; PMCID: PMC11783105.

Ouazzani K, Ansolabehere X, Journeau F, Vidal A, Jaubourg N, Doublet M, Thollot R, Fabre A, Glatt N. Project Victoria: A pragmatic data model to automate RWE generation from the national French claims database. Health Informatics J. 2025 Jan-Mar;31(1):14604582251318250. doi: 10.1177/14604582251318250. PMID: 39913942.

Popat A, Yadav S, Obholz J, Hwang EA, Rehman AU, Sharma P. The Efficacy of Artificial Intelligence in the Detection and Management of Atrial Fibrillation. Cureus. 2025 Jan 8;17(1):e77135. doi: 10.7759/cureus.77135. PMID: 39925585; PMCID: PMC11805596.

Ochola M, Kiwuwa-Muyingo S, Bhattacharjee T, Amadi D, Ng'etich M, Kadengye D, Owoko H, Igumba B, Greenfield J, Todd J, Kiragga A. Harmonizing population health data into OMOP common data model: a demonstration using COVID-19 sero-surveillance data from Nairobi Urban Health and Demographic Surveillance System. Front Digit Health. 2025 Jan 28;7:1423621. doi: 10.3389/fdgth.2025.1423621. PMID: 39949611; PMCID: PMC11822943.

Abellan A, Burn E, Trinh NTH, Burkard T, Callahan A, Fernández-Bertolín S, Hurley E, Rodriguez C, Segundo E, Morales DR, M E Nordeng H, Duarte-Salles T. Expanding the OMOP Common Data Model to Support Perinatal Research in Network Studies. Pharmacoepidemiol Drug Saf. 2025 Feb;34(2):e70106. doi: 10.1002/pds.70106. PMID: 39950235; PMCID: PMC11826376.

OHDSI Europe Symposium - Save-the-date!

OHDSI BELGIUM

EUROPE

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Save-the-date

5-7 July 2025

Location

Old Prison - Hasselt University
Martelarenlaan
Hasselt - BELGIUM

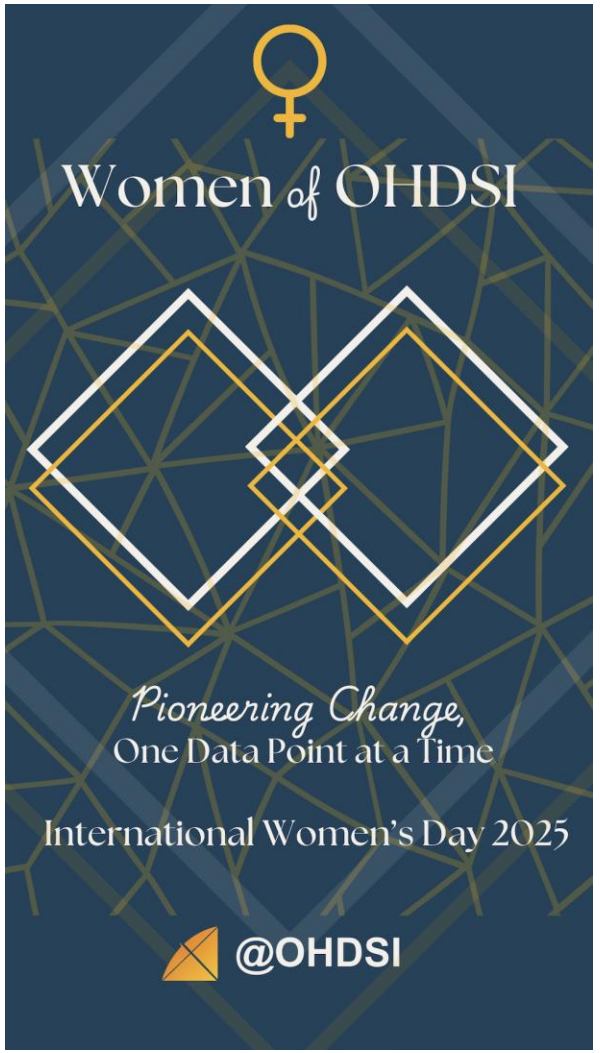
Global Symposium: Oct. 7-9

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

Today is the deadline to join the Scientific Review Committee!

International Women's Day

This Saturday is International Women's Day.



The Women of OHDSI are building an Instagram template for collaborators to share their photos in. Be on the lookout for this and help showcase the amazing work of women in our community throughout the world.

iCAN mNSCLC Studyathon 2025

📍 **March 25-28, Helsinki, Finland**

Exploring the Real-World Treatment Landscape of mNSCLC

In this studyathon, we will **characterize real-world treatment patterns of metastatic NSCLC**, with a focus on **the adoption and impact of immune checkpoint inhibitors (ICIs) across different regions.**

🔗 **Study GitHub Repository:** https://github.com/ohdsi-studies/MNSCLCStudyathon

📌 **If you're interested in contributing, please reach out:**

- **Asieh Golozar** – golozar@ohdsi.org
- **Kimmo Porkka** – kimmo.porkka@helsinki.fi
- **Eric Fey** – eric.fey@hus.fi

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

Hongfang Liu, PhD

D. Bradley McWilliams Chair Professor of Biomedical Informatics, Vice President of Learning Health System, University of Texas Health Science Center at Houston

'A Translational Science Framework in Advancing Healthcare AI'

March 27, 2025, 11am-12pm EST
Virtually via Zoom

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

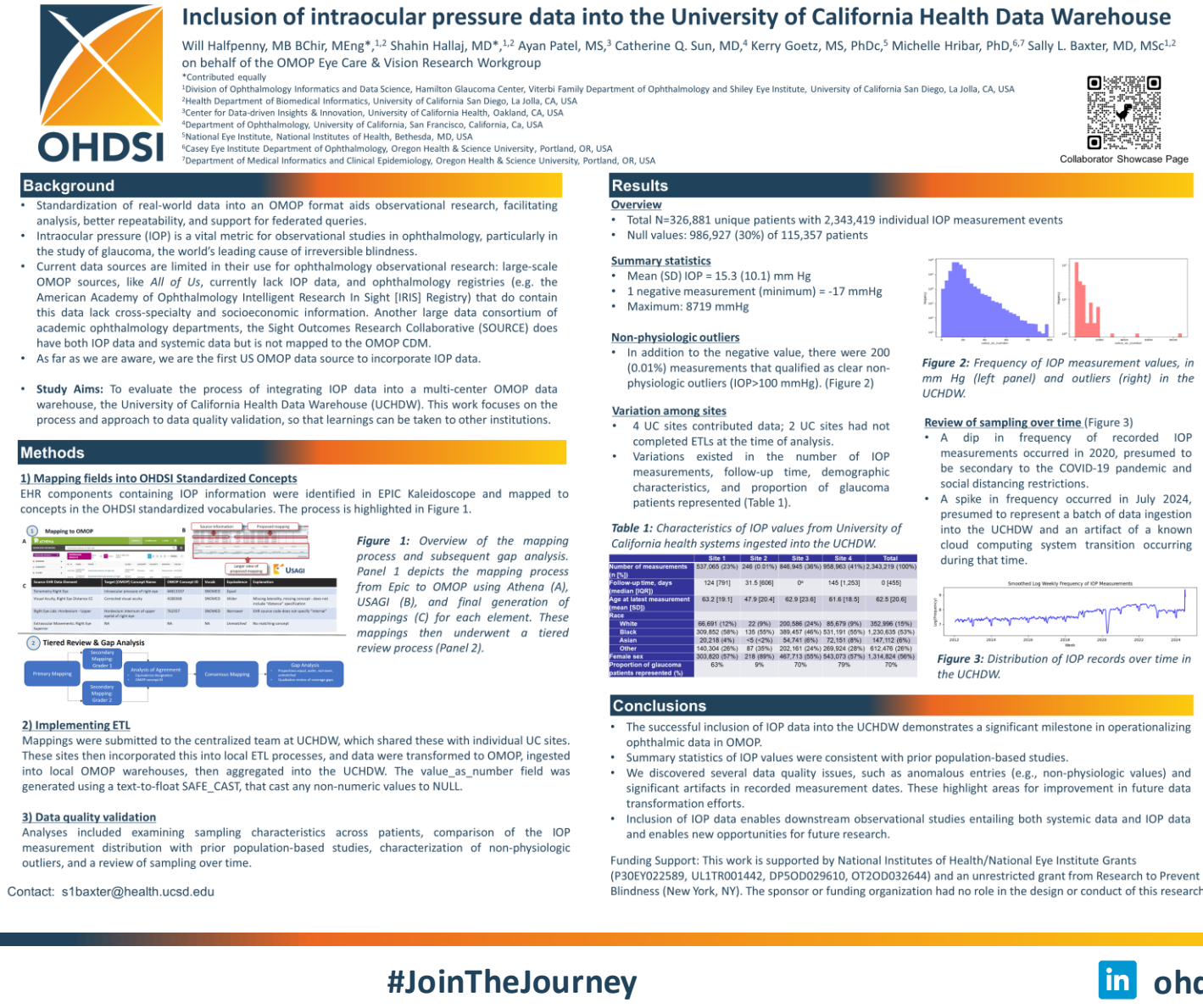
TuftsMedicine
Tufts Medical Center

#OHDSISocialShowcase This Week

Monday

Inclusion of intraocular pressure data into the University of California Health Data Warehouse

(William Halfpenny, Shahin Hallaj, Ayan Patel, Catherine Q. Sun, Kerry Goetz, Michelle Hribar, Sally L. Baxter, on behalf of the OMOP Eye Care & Vision Research Workgroup)



Inclusion of intraocular pressure data into the University of California Health Data Warehouse

Will Halfpenny, MB BChir, MEng*,[1,2] Shahin Hallaj, MD*,[1,2] Ayan Patel, MS,[3] Catherine Q. Sun, MD,[4] Kerry Goetz, MS, PhDc,[5] Michelle Hribar, PhD,[6,7] Sally L. Baxter, MD, MSc[1,2]
on behalf of the OMOP Eye Care & Vision Research Workgroup
*Contributed equally
[1]Division of Ophthalmology Informatics and Data Science, Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California San Diego, La Jolla, CA, USA
[2]Health Department of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA
[3]Center for Data-driven Insights & Innovation, University of California Health, Oakland, CA, USA
[4]Department of Ophthalmology, University of California, San Francisco, California, Ca, USA
[5]National Eye Institute, National Institutes of Health, Bethesda, MD, USA
[6]Casey Eye Institute Department of Ophthalmology, Oregon Health & Science University, Portland, OR, USA
[7]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Collaborator Showcase Page

Background

- Standardization of real-world data into an OMOP format aids observational research, facilitating analysis, better repeatability, and support for federated queries.
- Intraocular pressure (IOP) is a vital metric for observational studies in ophthalmology, particularly in the study of glaucoma, the world's leading cause of irreversible blindness.
- Current data sources are limited in their use for ophthalmology observational research: large-scale OMOP sources, like *All of Us*, currently lack IOP data, and ophthalmology registries (e.g. the American Academy of Ophthalmology Intelligent Research In Sight [IRIS] Registry) that do contain this data lack cross-specialty and socioeconomic information. Another large data consortium of academic ophthalmology departments, the Sight Outcomes Research Collaborative (SOURCE) does have both IOP data and systemic data but is not mapped to the OMOP CDM.
- As far as we are aware, we are the first US OMOP data source to incorporate IOP data.

- **Study Aims:** To evaluate the process of integrating IOP data into a multi-center OMOP data warehouse, the University of California Health Data Warehouse (UCHDW). This work focuses on the process and approach to data quality validation, so that learnings can be taken to other institutions.

Methods

1) Mapping fields into OHDSI Standardized Concepts

EHR components containing IOP information were identified in EPIC Kaleidoscope and mapped to concepts in the OHDSI standardized vocabularies. The process is highlighted in Figure 1.

Figure 1: *Overview of the mapping process and subsequent gap analysis. Panel 1 depicts the mapping process from Epic to OMOP using Athena (A), USAGI (B), and final generation of mappings (C) for each element. These mappings then underwent a tiered review process (Panel 2).*

2) Implementing ETL

Mappings were submitted to the centralized team at UCHDW, which shared these with individual UC sites. These sites then incorporated this into local ETL processes, and data were transformed to OMOP, ingested into local OMOP warehouses, then aggregated into the UCHDW. The value_as_number field was generated using a text-to-float SAFE_CAST, that cast any non-numeric values to NULL.

3) Data quality validation

Analyses included examining sampling characteristics across patients, comparison of the IOP measurement distribution with prior population-based studies, characterization of non-physiologic outliers, and a review of sampling over time.

Contact: s1baxter@health.ucsd.edu

Results

Overview

- Total N=326,881 unique patients with 2,343,419 individual IOP measurement events
- Null values: 986,927 (30%) of 115,357 patients

Summary statistics

- Mean (SD) IOP = 15.3 (10.1) mm Hg
- 1 negative measurement (minimum) = -17 mmHg
- Maximum: 8719 mmHg

Non-physiologic outliers

- In addition to the negative value, there were 200 (0.01%) measurements that qualified as clear non-physiologic outliers (IOP>100 mmHg). (Figure 2)

Variation among sites

- 4 UC sites contributed data; 2 UC sites had not completed ETLs at the time of analysis.
- Variations existed in the number of IOP measurements, follow-up time, demographic characteristics, and proportion of glaucoma patients represented (Table 1).

Table 1: *Characteristics of IOP values from University of California health systems ingested into the UCHDW.*

	Site 1	Site 2	Site 3	Site 4	Total
Number of measurements (n [%])	537,065 (23%)	246 (0.01%)	846,945 (36%)	958,963 (41%)	2,343,219 (100%)
Follow-up time, days (median [IQR])	124 [791]	31.5 [606]	0*	145 [1,253]	0 [455]
Age at latest measurement (mean [SD])	63.2 [19.1]	47.9 [20.4]	62.9 [23.6]	61.6 [18.5]	62.5 [20.6]
Race					
White	66,691 (12%)	22 (9%)	200,586 (24%)	85,679 (9%)	352,996 (15%)
Black	309,852 (58%)	135 (55%)	389,457 (46%)	531,191 (55%)	1,230,635 (53%)
Asian	20,218 (4%)	<5 (<2%)	54,741 (6%)	72,151 (8%)	147,112 (6%)
Other	140,304 (26%)	87 (35%)	202,161 (24%)	269,924 (28%)	612,476 (26%)
Female sex	303,820 (57%)	218 (89%)	467,713 (55%)	543,073 (57%)	1,314,824 (56%)
Proportion of glaucoma patients represented (%)	63%	9%	70%	79%	70%

Figure 2: *Frequency of IOP measurement values, in mm Hg (left panel) and outliers (right) in the UCHDW.*

Review of sampling over time (Figure 3)

- A dip in frequency of recorded IOP measurements occurred in 2020, presumed to be secondary to the COVID-19 pandemic and social distancing restrictions.
- A spike in frequency occurred in July 2024, presumed to represent a batch of data ingestion into the UCHDW and an artifact of a known cloud computing system transition occurring during that time.

Figure 3: *Distribution of IOP records over time in the UCHDW.*

Conclusions

- The successful inclusion of IOP data into the UCHDW demonstrates a significant milestone in operationalizing ophthalmic data in OMOP.
- Summary statistics of IOP values were consistent with prior population-based studies.
- We discovered several data quality issues, such as anomalous entries (e.g., non-physiologic values) and significant artifacts in recorded measurement dates. These highlight areas for improvement in future data transformation efforts.
- Inclusion of IOP data enables downstream observational studies entailing both systemic data and IOP data and enables new opportunities for future research.

Funding Support: This work is supported by National Institutes of Health/National Eye Institute Grants (P30EY022589, UL1TR001442, DP5OD029610, OT2OD032644) and an unrestricted grant from Research to Prevent Blindness (New York, NY). The sponsor or funding organization had no role in the design or conduct of this research.

#OHDSISocialShowcase This Week

Tuesday

Hierarchical Algorithms for Querying Physiologically Distinct Groups in Adult Congenital Heart Disease Using OMOP CDM

(Seohu Lee, Jong Ko, Haeun Lee, Ari Cedars)

Hierarchical Algorithms for Querying Physiologically Distinct Groups in Adult Congenital Heart Disease Using OMOP CDM

PRESENTER: Seohu Lee

INTRO:

- Adult Congenital Heart Disease (ACHD) is a rare, heterogeneous condition requiring large, multicenter datasets for effective study.
- Understanding ACHD through multicenter research can lead to better clinical insights, improved patient management, and the identification of rare subtypes.
- However, using the OMOP CDM for large-scale multicenter observational ACHD studies requires preliminary groundwork.

METHODS

- We developed hierarchical algorithms by mapping ICD-10-CM codes to SNOMED CT codes for different ACHD physiological groups. These mappings were validated with I-MAGIC and SNOMED International to ensure accuracy.
- We gathered relevant ICD-10-CM and SNOMED CT codes from the Johns Hopkins OMOP CDM dataset, covering approximately 2.1 million patients.
- The algorithms were tested by calculating patient counts for each physiological group in the JHM ATLAS. We also assessed the consistency of the ICD-10-CM to SNOMED mappings across I-MAGIC and SNOMED International.

Groundwork for ACHD OHDSI Network Study:

10 Hierarchical Algorithms for ACHD Physiological Groups

Identification of ACHD Codes → Conversion to SNOMED CT → Validation → Development of Hierarchical Algorithms → Application for Patient Counts → Results & Refinement

Discrepancy Table of Original and Reconverted Codes of ICD-10-CM

No.	Original Codes	Reconverted Codes	
	ICD-10-CM	I-MAGIC	SNOMED International
1	Q20 Congenital malformations of cardiac chambers and connections	Q24.9 Congenital malformation of heart, unspecified	Q24.9 Congenital malformation of heart, unspecified
9	Q20.8 Other congenital malformations of cardiac chambers and connections	Q24.9 Congenital malformation of heart, unspecified	Q24.9 Congenital malformation of heart, unspecified
10	Q20.9 Congenital malformation of cardiac chambers and connections, unspecified	Q24.9 Congenital malformation of heart, unspecified	Q24.9 Congenital malformation of heart, unspecified
11	Q21 Congenital malformations of cardiac septa	Q21.9 Congenital malformation of cardiac septum, unspecified	Q21.9 Congenital malformation of cardiac septum, unspecified
18	Q21.14 Superior sinus venosus atrial septal defect	Q21.16 Sinus venosus atrial septal defect, unspecified	Q21.1 Atrial septal defect
19	Q21.15 Inferior sinus venosus atrial septal defect	Q21.16 Sinus venosus atrial septal defect, unspecified	Q21.1 Atrial septal defect
21	Q21.19 Other specified atrial septal defect	Q21.10 Atrial septal defect, unspecified	Q21.1 Atrial septal defect
23	Q21.20 Atrioventricular septal defect, unspecified as to partial or complete	Q21.23 Complete atrioventricular septal defect	Q21.2 Atrioventricular septal defect
24	Q21.21 Partial atrioventricular septal defect	Q21.23 Complete atrioventricular septal defect	Q21.2 Atrioventricular septal defect
25	Q21.22 Transitional atrioventricular septal defect	Q21.23 Complete atrioventricular septal defect	Q21.2 Atrioventricular septal defect
...	...	...	...
83	Q25.8 Other congenital malformations of other great arteries	Q27.9 Congenital malformation of peripheral vascular system, unspecified	Q27.9 Congenital malformation of peripheral vascular system, unspecified
84	Q25.9 Congenital malformation of great arteries, unspecified	Q27.9 Congenital malformation of peripheral vascular system, unspecified	Q27.9 Congenital malformation of peripheral vascular system, unspecified
85	Q26 Congenital malformations of great veins	Q26.9 Congenital malformation of great vein, unspecified	Q26.9 Congenital malformation of great vein, unspecified
90	Q26.4 Anomalous pulmonary venous connection, unspecified	Q26.9 Congenital malformation of great vein, unspecified	Q26.8 Other congenital malformations of great veins
91	Q26.8 Other congenital malformations of great veins	Q26.9 Congenital malformation of great vein, unspecified	Q26.9 Congenital malformation of great vein, unspecified

Hierarchical Algorithms for ACHD Physiological Groups and Corresponding Patient Count from JHM ATLAS

ACHD Physiological Groups	Hierarchical Algorithms with SNOMED CT Code	Patient Count (n)
1. Eisenmenger Syndrome/Shunt with pulmonary hypertension	[(434462 and/or 4099995 and/or 4100152 and/or 4289309 and/or 315922 and/or 4061819) and (4322024 or 4339214)] or 40493243	785
2. Fontan/Glenn/Single Ventricle	4339962 and/or 4208834 and/or 4050559 and/or 2107269 and/or 4051948 and/or 40491942	2
3-1. D-Transposition of the great arteries with atrial switch	(432431 and/or 40456182 and/or 313867) and (4221982 and/or 4075541 and/or 2107361)	0
3-2. D-Transposition of the great arteries with arterial switch	(432431 and/or 40456182 and/or 313867) and (4019932 and/or 4286184 and/or 4077745 and/or 4122006)	0
4. L-Transposition of the great arteries	(432431 and/or 40456182 and/or 313867) and (4100733 and/or 4101005)	59
5. Tetralogy of Fallot/DORV TOF type	313867 and/or 4101618 and/or 320835 and/or 4109337 and/or 4101619	771
6. Truncus arteriosus	441950 and/or 45766266	46
7. AV Canal defects	4100152 and/or 4235784 and/or 435912 and/or 37164933	311
8. Ebstein's anomaly	4069182	100
9. Shone Complex	(313006 and/or 441108 and/or 40404007 and/or 4100869) and (4324704 and/or 4062247 and/or 321119 and/or 4253808 and/or 314457 and/or 259123 and/or 4147787)	18
10. Sinus Venosus	4316879	45

Conclusions

Key Takeaways

- The hierarchical algorithms demonstrate the potential of using OMOP CDM to categorize ACHD physiological groups effectively. However, low patient counts in certain groups (e.g., Fontan/Glenn/Single Ventricle and D-Transposition) indicate a need for further refinement and validation.

Future Directions

- Future studies should apply these algorithms to real patient data from multiple institutions to confirm their accuracy and broader applicability. This groundwork supports larger multicenter studies, enhancing our understanding of ACHD through standardized data analysis.

	I-MAGIC	SNOMED International
Number of Agreement	55	47
Percentage (%)	59.78	51.09

Take a picture to **download** the **full paper**

Seohu Lee, Jong Ko, Haeun Lee, Ari Cedars

slee619@jhu.edu

JOHNS HOPKINS MEDICINE SCHOOL OF MEDICINE

OHDSI

#OHDSISocialShowcase This Week

Wednesday

Vasculitis without phlebitis phenotype development using real-world data: development and evaluation study

(Jill Hardin, Amir Sarayani, Dina Gifkins, Tara Beaulieu, James Gilbert, Joel Swerdel)

Vasculitis without phlebitis phenotype development using real-world data: development and evaluation study

PRESENTER: **Jill Hardin**

INTRODUCTION:

1. Phenotypes with high positive predictive value (PPV) and sensitivity are needed for safety studies using real world data (RWD).
2. Health authorities have requested safety studies to be conducted on the clinical outcome of vasculitis without phlebitis.
3. **This study aimed to develop a RWD phenotype for vasculitis without phlebitis.**

METHODS:

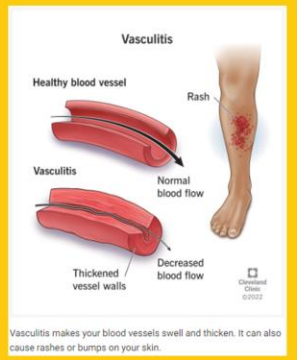
1. Comprehensive literature review to define the clinical concept and identify previously developed phenotypes.
2. OHDSI software tools ie, PHEOBE, ATLAS, Cohort Diagnostics, and Phevaluator facilitated this phenotype development project.
3. Data sources:

Database	Years	Country	Data Type	Visit Types	# of Persons (millions)	Mean Age at 1st Observation	Median Length of Follow-up (years)
Merative® MarketScan Commercial Claims and Encounters (CCAE)	2000-2024	US	Claims	IP/OP	172	31	2.84
Merative® MarketScan Medicare Supplemental (MDCR)	2000-2024	US	Claims	IP/OP	11	71	3.98
Optum's Clinformatics® Data Mart + Date of Death (Optum)	2000-2023	US	Claims	IP/OP	99	36	3.21
Optum® Electronic Health Record dataset (Optum EHR)	2007-2024	US	EHR	IP/OP	114	37	4.91
IQVIA® Pharmetrics	2015-2023	US	Claims	IP/OP	163	34	2.62

RESULTS:

1. Algorithm used an occurrence of a diagnosis code for vasculitis.
2. The concept set expression had 175 standard SNOMED concepts for vasculitis and 4 nonstandard concepts from the observation and condition domains.

A phenotype for vasculitis without phlebitis showed acceptable performance metrics.



Vasculitis makes your blood vessels swell and thicken. It can also cause rashes or bumps on your skin.

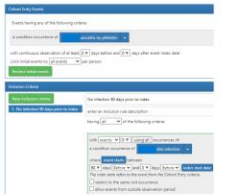
Database Name	PheValuator Sensitivity	PheValuator PPV	PheValuator specificity	PheValuator NPV
Merative CCAE	0.804 (0.791 - 0.817)	0.775 (0.761 - 0.788)	1.000 (1.000 - 1.000)	1.000 (1.000 - 1.000)
Merative MDCR	0.535 (0.528 - 0.542)	0.858 (0.851 - 0.864)	0.999 (0.999 - 0.999)	0.996 (0.995 - 0.996)
Optum EHR	0.839 (0.827 - 0.850)	0.704 (0.691 - 0.717)	0.999 (0.999 - 0.999)	1.000 (1.000 - 1.000)
Pharmetrics	0.898 (0.888 - 0.908)	0.741 (0.728 - 0.754)	0.999 (0.999 - 0.999)	1.000 (1.000 - 1.000)
Optum DOD	0.781 (0.772 - 0.790)	0.809 (0.800 - 0.818)	0.999 (0.999 - 0.999)	0.999 (0.999 - 0.999)

Summary of phenotype performance metrics estimated via the Phevaluator tool. The metrics with the best performance is highlighted in green, while the worst is highlighted in red.

Take a picture to **download** the **poster**

ADDITIONAL RESULTS:

3. Algorithm required no skin infections in the 90 days prior to and including the index date of vasculitis.
4. Skin infection was defined using the SNOMED 'infection of skin' concept id 4029043 from the condition domain and included 1289 descendant concepts.



5. Counts of persons identified with phenotype ranged from 101,912 in MDCR to 361,601 in Optum EHR.
6. Between 60% (Pharmetrics) to 65% (MDCR) persons were female.
7. Between 31% (Pharmetrics) and 53% (Optum EHR) had a drug dispensing for corticosteroids in 1 to 30 days after the index event for vasculitis.

CONCLUSIONS:

1. Our literature review found only one study[1] that provided performance metrics. This study included 446 persons with vasculitis and used Canadian administrative data across Nova Scotia, limiting the generalizability of the findings.
2. Our study included a larger number of individuals with vasculitis and used large administrative claims and EHR databases, resulting in greater generalizability.

REFERENCES:

1. Bernatsky S, Linehan T, Hanly JG. The accuracy of administrative data diagnoses of systemic autoimmune rheumatic diseases. J Rheumatol. 2011 Aug;38(8):1612-6. doi: 10.3899/jrheum.101149.

Jill Hardin[1,2], Eva-maria Didden[3], Amir Sarayani[4], Dina Gifkins[4], Tara Beaulieu[4], James Gilbert[1,2], Joel Swerdel[1,2]
[1]Observational Health Data Analytics, Global Epidemiology, Janssen Research and Development, Titusville, NJ, USA; [2]Observational Health Data Sciences and Informatics, New York, NY, USA; [3]Global Epidemiology, Janssen Research and Development, Basel, Switzerland; [4]Global Epidemiology, Janssen Research and Development, Raritan, NJ, USA

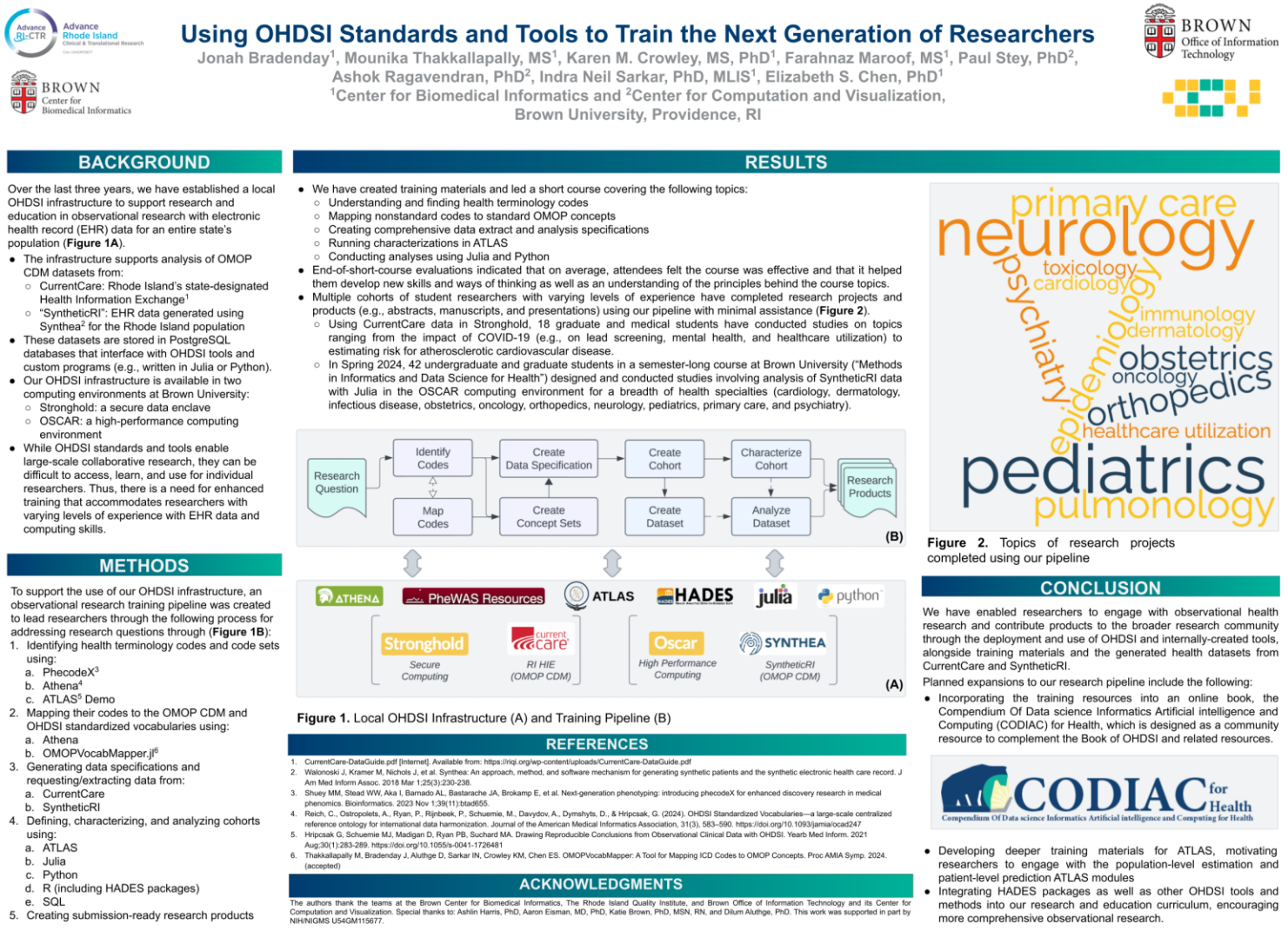
Johnson&Johnson
Global Epidemiology Organization

#OHDSISocialShowcase This Week

Thursday

Using OHDSI Standards and Tools to Train the Next Generation of Researchers

(Jonah Bradenday, Mounika Thakkallapally, Karen M. Crowley, Farahnaz Maroof, Paul Stey, Ashok Ragavendran, Indra Neil Sarkar, Elizabeth S. Chen)



Using OHDSI Standards and Tools to Train the Next Generation of Researchers

Jonah Bradenday[1], Mounika Thakkallapally, MS[1], Karen M. Crowley, MS, PhD[1], Farahnaz Maroof, MS[1], Paul Stey, PhD[2], Ashok Ragavendran, PhD[2], Indra Neil Sarkar, PhD, MLIS[1], Elizabeth S. Chen, PhD[1]
[1]Center for Biomedical Informatics and [2]Center for Computation and Visualization, Brown University, Providence, RI

BACKGROUND

Over the last three years, we have established a local OHDSI infrastructure to support research and education in observational research with electronic health record (EHR) data for an entire state's population (**Figure 1A**).

- The infrastructure supports analysis of OMOP CDM datasets from:
  - CurrentCare: Rhode Island's state-designated Health Information Exchange[1]
  - "SyntheticRI": EHR data generated using Synthea[2] for the Rhode Island population
- These datasets are stored in PostgreSQL databases that interface with OHDSI tools and custom programs (e.g., written in Julia or Python).
- Our OHDSI infrastructure is available in two computing environments at Brown University:
  - Stronghold: a secure data enclave
  - OSCAR: a high-performance computing environment
- While OHDSI standards and tools enable large-scale collaborative research, they can be difficult to access, learn, and use for individual researchers. Thus, there is a need for enhanced training that accommodates researchers with varying levels of experience with EHR data and computing skills.

METHODS

To support the use of our OHDSI infrastructure, an observational research training pipeline was created to lead researchers through the following process for addressing research questions through (**Figure 1B**):

1. Identifying health terminology codes and code sets using:
   a. PhecodeX[3]
   b. Athena[4]
   c. ATLAS[5] Demo
2. Mapping their codes to the OMOP CDM and OHDSI standardized vocabularies using:
   a. Athena
   b. OMOPVocabMapper.jl[6]
3. Generating data specifications and requesting/extracting data from:
   a. CurrentCare
   b. SyntheticRI
4. Defining, characterizing, and analyzing cohorts using:
   a. ATLAS
   b. Julia
   c. Python
   d. R (including HADES packages)
   e. SQL
5. Creating submission-ready research products

RESULTS

- We have created training materials and led a short course covering the following topics:
  - Understanding and finding health terminology codes
  - Mapping nonstandard codes to standard OMOP concepts
  - Creating comprehensive data extract and analysis specifications
  - Running characterizations in ATLAS
  - Conducting analyses using Julia and Python
- End-of-short-course evaluations indicated that on average, attendees felt the course was effective and that it helped them develop new skills and ways of thinking as well as an understanding of the principles behind the course topics.
- Multiple cohorts of student researchers with varying levels of experience have completed research projects and products (e.g., abstracts, manuscripts, and presentations) using our pipeline with minimal assistance (**Figure 2**).
  - Using CurrentCare data in Stronghold, 18 graduate and medical students have conducted studies on topics ranging from the impact of COVID-19 (e.g., on lead screening, mental health, and healthcare utilization) to estimating risk for atherosclerotic cardiovascular disease.
  - In Spring 2024, 42 undergraduate and graduate students in a semester-long course at Brown University ("Methods in Informatics and Data Science for Health") designed and conducted studies involving analysis of SyntheticRI data with Julia in the OSCAR computing environment for a breadth of health specialties (cardiology, dermatology, infectious disease, obstetrics, oncology, orthopedics, neurology, pediatrics, primary care, and psychiatry).

Figure 1. Local OHDSI Infrastructure (A) and Training Pipeline (B)

Figure 2. Topics of research projects completed using our pipeline

CONCLUSION

We have enabled researchers to engage with observational health research and contribute products to the broader research community through the deployment and use of OHDSI and internally-created tools, alongside training materials and the generated health datasets from CurrentCare and SyntheticRI.

Planned expansions to our research pipeline include the following:

- Incorporating the training resources into an online book, the Compendium Of Data science Informatics Artificial intelligence and Computing (CODIAC) for Health, which is designed as a community resource to complement the Book of OHDSI and related resources.
- Developing deeper training materials for ATLAS, motivating researchers to engage with the population-level estimation and patient-level prediction ATLAS modules
- Integrating HADES packages as well as other OHDSI tools and methods into our research and education curriculum, encouraging more comprehensive observational research.

REFERENCES

1. CurrentCare-DataGuide.pdf [Internet]. Available from: https://riqi.org/wp-content/uploads/CurrentCare-DataGuide.pdf
2. Walonoski J, Kramer M, Nichols J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018 Mar 1;25(3):230-238.
3. Shuey MM, Stead WW, Aka I, Barnado AL, Bastarache JA, Brokamp E, et al. Next-generation phenotyping: introducing phecodeX for enhanced discovery research in medical phenomics. Bioinformatics. 2023 Nov 1;39(11):btad655.
4. Reich, C., Ostropolets, A., Ryan, P., Rijnbeek, P., Schuemie, M., Davydov, A., Dymshyts, D., & Hripcsak, G. (2024). OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. Journal of the American Medical Informatics Association, 31(3), 583–590. https://doi.org/10.1093/jamia/ocad247
5. Hripcsak G, Schuemie MJ, Madigan D, Ryan PB, Suchard MA. Drawing Reproducible Conclusions from Observational Clinical Data with OHDSI. Yearb Med Inform. 2021 Aug;30(1):283-289. https://doi.org/10.1055/s-0041-1726481
6. Thakkallapally M, Bradenday J, Aluthge D, Sarkar IN, Crowley KM, Chen ES. OMOPVocabMapper: A Tool for Mapping ICD Codes to OMOP Concepts. Proc AMIA Symp. 2024. (accepted)

ACKNOWLEDGMENTS

The authors thank the teams at the Brown Center for Biomedical Informatics, The Rhode Island Quality Institute, and Brown Office of Information Technology and its Center for Computation and Visualization. Special thanks to: Ashlin Harris, PhD, Aaron Eisman, MD, PhD, Katie Brown, PhD, MSN, RN, and Dilum Aluthge, PhD. This work was supported in part by NIH/NIGMS U54GM115677.

#OHDSISocialShowcase This Week

Friday

Comparing probabilistic and rule-based phenotype algorithms for hypotension and angioedema to the experience observed in randomized clinical trials

(Joel Swerdel, Martijn Schuemie, Judy Racoosin, Patrick Ryan)

Comparing probabilistic and rule-based phenotype algorithms for hypotension and angioedema to the experience observed in randomized clinical trials.

Presenter: Joel Swerdel

BACKGROUND

- Rule-based phenotype algorithms (PAs) are the standard for identifying outcomes in observational data.
- However, the performance characteristics of the PAs, such as sensitivity and positive predictive value (PPV), are estimated to be low for many phenotypes.
- Probabilistic PAs, e.g., PAs based on logistic regression models, offer an alternative to the rule-based method. Prior efforts have demonstrated the potential for probabilistic phenotyping as an alternative to rule-based PAs.[1-3]

METHODS

Developing probabilistic phenotypes:

1. Use noisy labeled positive and negative controls to develop a supervised learning probabilistic model using LASSO regularized regression.
2. Apply model at each appropriate time point during the time-at-risk for each subject in the cohort of interest.
3. Select the highest probability among the different time points within the time-at-risk for each subject.
4. Use a designated probability cut-point, e.g., 70%, to determine those with the outcome.

Evaluating the models:

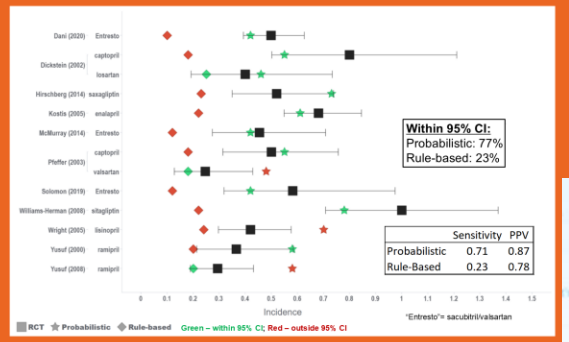
- Analysis conducted in 5 administrative claims datasets.
- Rule-based algorithm used an occurrence of a diagnosis code for hypotension or angioedema.
- Developed probabilistic phenotypes and examined the results using probability cut-points of 0.50, 0.60, 0.67, 0.70, 0.75, 0.80, and 0.90.
- Estimated incidence of angioedema, while on-treatment, for 7 anti-hypertensive and 2 anti-diabetic (DPP-4 inhibitors) drugs and of hypotension, while on-treatment, for 6 anti-hypertensive drugs.
- Using both the rule-based and probabilistic phenotypes, we performed the analysis on 9 new user drug cohorts from 2010 to 2023.

Metrics:

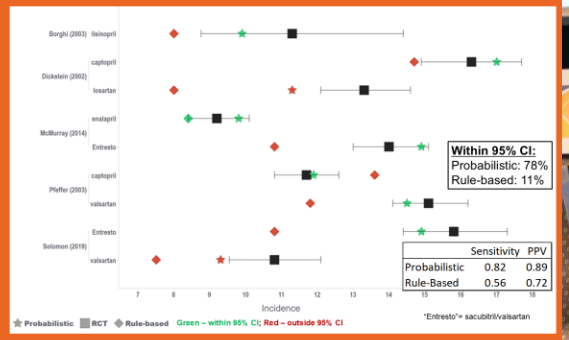
- Extracted incidence estimates and 95% confidence intervals (CI) from randomized clinical trials (RCTs) as a basis for comparison.
- Computed the proportion of incidence estimates for the rule-based and probabilistic algorithms that fell within the 95% CI of the incidence estimates from the clinical trials.
- Assessed the performance characteristics, e.g., PPV and sensitivity, of the rule-based and probabilistic phenotypes using the OHDSI tool PheValuator.

Probabilistic phenotype algorithms for angioedema and hypotension estimated incidence closer to the results from RCTs than rule-based phenotype algorithms.

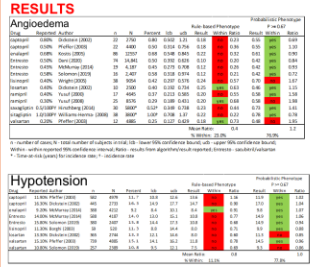
Angioedema



Hypotension



RESULTS



CONCLUSIONS

- Probabilistic phenotype algorithms (PA) for angioedema and hypotension estimated incidence closer to the results from RCTs than rule-based PAs.
- The performance of probabilistic PAs was superior to rule-based PAs on PPV and sensitivity.
- Future research is needed to evaluate the performance of probabilistic PAs and to determine how they could potentially be used to estimate the incidence of drug adverse effects.

REFERENCES

1. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc. 2016;23(6):1166-73.
2. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci

Where Are We Going?

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

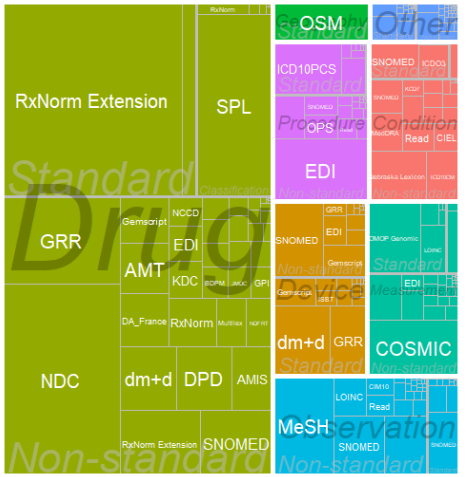
Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

March 4: Vocabulary Refresh



Anna Ostropolets

Associate Director, Johnson & Johnson Innovative Medicine
Adjunct Assistant Professor, Columbia University

Oleg Zhuk

Manager, Data Analytics Consulting, EPAM Systems

Maria Khitrun

Senior Scientific Curation Specialist, EPAM Systems

This session will also include a Phenotype Phebruary review from the members of our leadership team:

Anna Ostropolets
Gowtham Rao
Azza Shoaibi

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

Everybody is invited!

Links are sent out weekly and available at:
ohdsi.org/community-calls-2025