Clinical Guideline Review, Session 1

OHDSI Community Call
Jan. 21, 2025 • 11 am ET

Upcoming Community Calls

Date	Topic
Jan. 21	Clinical Guideline Review, Session I
Jan. 28	Clinical Guideline Review, Session II
Feb. 4	First Week of 2025 Workgroup OKRs/Phenotype Phebruary
Feb. 11	Second Week of 2025 Workgroup OKRs/Phenotype Phebruary
Feb. 18	Third Week of 2025 Workgroup OKRs/Phenotype Phebruary
Feb. 25	Fourth Week of 2025 Workgroup OKRs/Phenotype Phebruary

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Mitchell Conover, Yasser Albogami, Jill Hardin, Christian Reich, Anna Ostropolets, Patrick Ryan, and the OHDSI Research Network** on the publication of **Glucagon-Like Peptide 1 Receptor Agonists and Chronic Lower Respiratory Disease Among Type 2 Diabetes Patients: Replication and Reliability Assessment Across a Research Network** in *Pharmacoepidemiology & Drug Safety.*

Pharmacoepidemiology and Drug Safety

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Glucagon-Like Peptide 1 Receptor Agonists and Chronic Lower Respiratory Disease Among Type 2 Diabetes Patients: Replication and Reliability Assessment Across a Research Network

Mitchell M. Conover[1,2] | Yasser Albogami[1,3] | Jill Hardin[1,2] | Christian G. Reich[1,4] | Anna Ostropolets[1,2,5] | Patrick B. Ryan[1,2,5] | Observational Health Data Sciences and Informatics (OHDSI) Research Network

[1]Observational Health Data Science and Informatics, New York, New York, USA | [2]Observational Health Data Analytics, Johnson & Johnson, Titusville, New Jersey, USA | [3]Department of Clinical Pharmacy, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia | [4]Real World Solutions, IQVIA, Cambridge, Massachusetts, USA | [5]Department of Biomedical Informatics, Columbia University, New York, New York, USA

Correspondence: Mitchell M. Conover (mconove1@its.jnj.com)

Received: 31 January 2024 | **Revised:** 12 December 2024 | **Accepted:** 13 December 2024

Funding: This study was partially funded by Observational Health Data Sciences and Informatics (OHDSI) Research Network and Janssen Research & Development, a Johnson & Johnson Company.

Keywords: chronic lower respiratory disease | common data model | glucagon-like peptide 1 receptor agonists | pharmacoepidemiology | real world data | real world evidence | reliability | replicability | reproducibility | transparency

OHDSI Shoutouts!

Congratulations to the team of **Karamarie Fecho, Juan J. Garcia, Hong Yi, Griffin Roupe and Ashok Krishnamurthy** on the publication of **FHIR PIT: a geospatial and spatiotemporal data integration pipeline to support subject-level clinical research** in *BMC Medical Informatics and Decision Making.*

Fecho *et al.*
BMC Medical Informatics and Decision Making (2025) 25:24
https://doi.org/10.1186/s12911-024-02815-6

BMC Medical Informatics and Decision Making

SOFTWARE Open Access

FHIR PIT: a geospatial and spatiotemporal data integration pipeline to support subject-level clinical research

Karamarie Fecho[1,2*†], Juan J. Garcia[3†], Hong Yi[1†], Griffin Roupe[1,3] and Ashok Krishnamurthy[1,3]

Abstract

Background Environmental exposures such as airborne pollutant exposures and socio-economic indicators are increasingly recognized as important to consider when conducting clinical research using electronic health record (EHR) data or other sources of clinical data such as survey data. While numerous public sources of geospatial and spatiotemporal data are available to support such research, the data are challenging to work with due to inconsistencies in file formats and spatiotemporal resolutions, computational challenges with large file sizes, and a lack of tools for patient- or subject-level data integration.

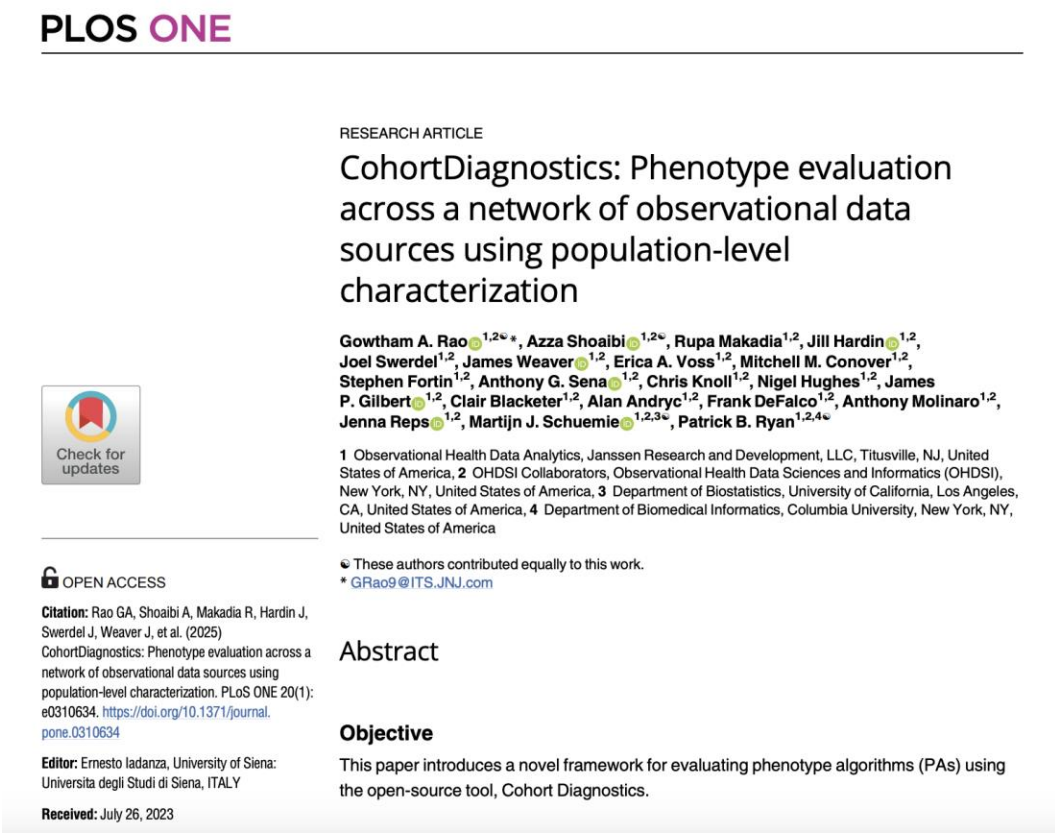
Results We developed FHIR PIT (HL7® Fast Healthcare Interoperability Resources Patient data Integration Tool) as an open-source, modular, data-integration software pipeline that consumes EHR data in FHIR® format and integrates the data at the level of the patient or subject with environmental exposures data of varying spatiotemporal resolutions and file formats. We applied FHIR PIT to generate "integrated feature tables" containing patient- or subject-level EHR data integrated with environmental exposures data on two cohorts: one on patients with asthma and related common pulmonary disorders; and a second on patients with primary ciliary dyskinesia and related rare pulmonary disorders. The data were then exposed via the open Integrated Clinical and Environmental Exposures Service, which was then queried to explore relationships between exposures to two representative airborne pollutants (particulate matter and ozone) and annual emergency department or inpatient visits for respiratory issues. We found that hospitalizations for respiratory issues were more common among patients exposed to relatively high levels of particulate matter and ozone and were higher overall among patients with primary ciliary dyskinesia than among patients with asthma.

Conclusions Our manuscript describes a major release of FHIR PIT v1.0 and includes a technical demonstration use case and a clinical application on the use of FHIR PIT to support research on environmental exposures and health outcomes related to asthma and primary ciliary dyskinesia. For application of the tool to common data models (CDMs) other than FHIR, we offer open-source conversion tools to map from the PCORnet, i2b2, and OMOP CDMs to FHIR.

Keywords Asthma, Primary ciliary dyskinesia, HL7® FHIR®, Data integration, Environmental exposures, Airborne pollutant exposures, Socioeconomic exposures, Hospital visits

OHDSI Shoutouts!

Congratulations to the team of **Gowtham Rao, Azza Shoaibi, Rupa Makadia, Jill Hardin, Joel Swerdel, James Weaver, Erica Voss, Mitchell Conover, Stephen Fortin, Anthony Sena, Chris Knoll, Nigel Hughes, James Gilbert, Clair Blacketer, Alan Andryc, Frank DeFalco, Anthony Molinaro, Jenna Reps, Martijn Schuemie, and Patrick Ryan** on the publication of **CohortDiagnostics: Phenotype evaluation across a network of observational data sources using population-level characterization** in the *PLOS One.*



PLOS ONE

RESEARCH ARTICLE

CohortDiagnostics: Phenotype evaluation across a network of observational data sources using population-level characterization

Gowtham A. Rao[1,2☯]*, Azza Shoaibi[1,2☯], Rupa Makadia[1,2], Jill Hardin[1,2], Joel Swerdel[1,2], James Weaver[1,2], Erica A. Voss[1,2], Mitchell M. Conover[1,2], Stephen Fortin[1,2], Anthony G. Sena[1,2], Chris Knoll[1,2], Nigel Hughes[1,2], James P. Gilbert[1,2], Clair Blacketer[1,2], Alan Andryc[1,2], Frank DeFalco[1,2], Anthony Molinaro[1,2], Jenna Reps[1,2], Martijn J. Schuemie[1,2,3☯], Patrick B. Ryan[1,2,4☯]

1 Observational Health Data Analytics, Janssen Research and Development, LLC, Titusville, NJ, United States of America, **2** OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, United States of America, **3** Department of Biostatistics, University of California, Los Angeles, CA, United States of America, **4** Department of Biomedical Informatics, Columbia University, New York, NY, United States of America

☯ These authors contributed equally to this work.
* GRao9@ITS.JNJ.com

Check for updates

OPEN ACCESS

Citation: Rao GA, Shoaibi A, Makadia R, Hardin J, Swerdel J, Weaver J, et al. (2025) CohortDiagnostics: Phenotype evaluation across a network of observational data sources using population-level characterization. PLoS ONE 20(1): e0310634. https://doi.org/10.1371/journal.pone.0310634

Editor: Ernesto Iadanza, University of Siena: Universita degli Studi di Siena, ITALY

Received: July 26, 2023

Abstract

Objective

This paper introduces a novel framework for evaluating phenotype algorithms (PAs) using the open-source tool, Cohort Diagnostics.

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

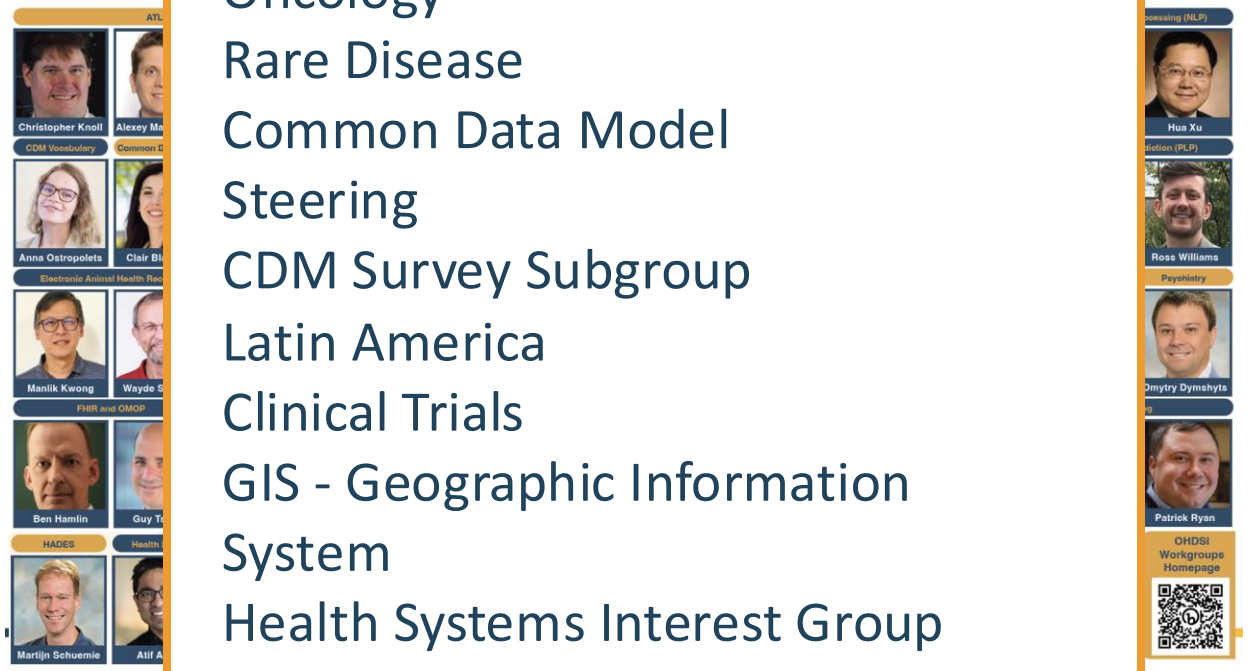
Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Tuesday	12 pm	Common Data Model Vocabulary Subgroup
Tuesday	12 pm	Atlas
Tuesday	1 pm	Common Data Model
Wednesday	7 am	Medical Imaging
Wednesday	12 pm	Latin America
Wednesday	1 pm	Perinatal & Reproductive Health
Thursday	8 am	Medical Devices
Thursday	9:30 am	Network Data Quality
Thursday	7 pm	Dentistry
Friday	9 am	Phenotype Development & Evaluation
Friday	10 am	GIS - Geographic Information System
Friday	11:30 am	Steering
Monday	9 am	Vaccine Vocabulary
Tuesday	9 am	OMOP CDM Oncology Genomic Subgroup

Workgroup OKRs

Each year, workgroup representatives join a February community call to present the mission, objectives and key results for their respective groups. These 2-4 minute presentations are recorded and posted on the Workgroups homepage on OHDSI.org.

Please choose a date to sign up for a February date; once a date has at least 10 workgroups, it will be closed.



Already Signed Up:

Oncology
Rare Disease
Common Data Model
Steering
CDM Survey Subgroup
Latin America
Clinical Trials
GIS - Geographic Information System
Health Systems Interest Group
Eye Care and Vision Research
Transplant
Themis
Medical Devices

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

Vipina Keloth, PhD

Associate Research Scientist in Biomedical Informatics and Data Science at Yale University School of Medicine

'Exploring the realm of large language models for information extraction in the biomedical domain'

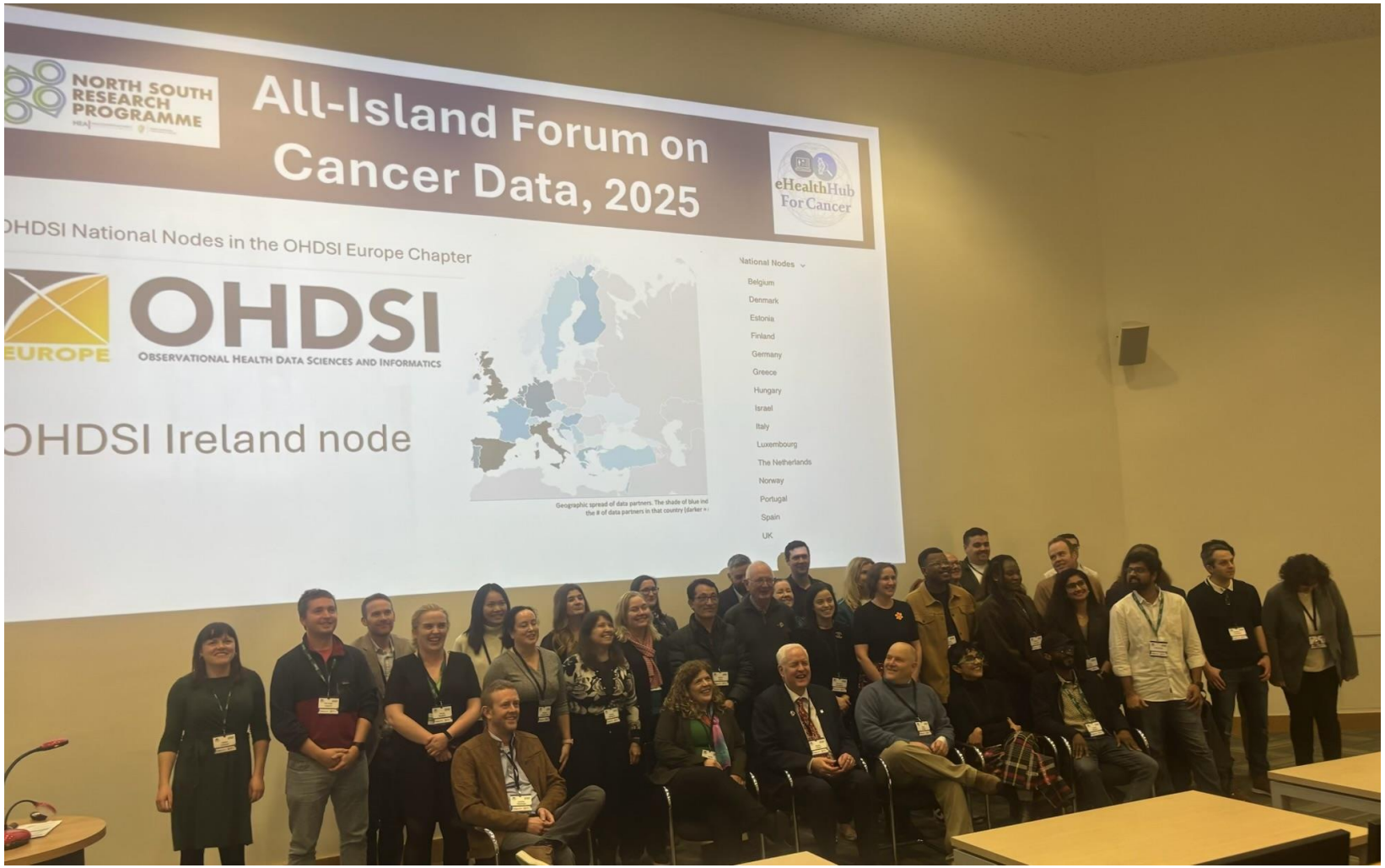
January 23, 2025, 11am-12pm EST
Virtually via Zoom

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

TuftsMedicine
Tufts Medical Center

Save The Dates!

Coming Soon: OHDSI Ireland Node



CDM Survey Subgroup Landscape Assessment

The CDM Survey Subgroup invites colleagues who have or are going to design, develop, and/or implement research surveys and use them with the OMOP CDM to share information about those efforts by completing this survey. Your completion of this 10-15 minute survey will provide information to the CDM workgroup about OMOP utilization among survey research teams. The CDM Survey subgroup is a collaborative effort, led by a team at the National Cancer Institute, to develop standardized approaches and best practices for helping research teams better integrate survey data elements into the OMOP common data model.

The deadline has been extended to mid-January.

LANDSCAPE ASSESSMENT

- **Activities**
  - Invite representatives from cohorts with experience using the CDM for survey data to share their knowledge and challenges.
  - Conduct a community survey to gather information on experiences and needs related to survey data in the CDM.
  - Review the most used Common Data Elements (CDMs) as a foundation for developing standards, tools, and best practices.
- **Key Result**
  - A comprehensive report summarizing survey CDM mapping resources, challenges, and identified development priorities (vocabulary, standards, tools, best practices) to be shared with the OHDSI community.

WHO SHOULD PARTICIPATE

- You have survey data and you've mapped it to the OMOP CDM
- You have survey data and you would like to map it to the OMOP CDM
- You are in the process of developing a survey(s) and plan to map to the OMOP CDM
- Multiple perspectives from the same team
- Multiple surveys from the same person

#OHDSISocialShowcase This Week

Monday

Leveraging UDI for Advanced Medical Device Tracking in OMOP-CDM

(Seojeong Shin, Yiju Park, Sujeong Eom, Kyulee Jeon, Seng Chan You)

Title: Leveraging UDI for Advanced Medical Device Safety Study

Enhancing Data Integration and Safety Analysis through UDI Incorporation in OMOP-CDM

PRESENTER: **Seojeong Shin**

INTRO

- Medical device safety research is essential for ensuring patient safety and effectively managing recalls and adverse events. While the OMOP-CDM includes a device table, international studies have not been actively conducted due to limitations in data granularity and standardization.
- For instance, the U.S. uses CPT4, while Korea maps its data to SNOMED-CT, which creates challenges in harmonizing device data across countries.
- This study aims to evaluate the feasibility of using the Unique Device Identifier (UDI) to link hospital clinical data with OMOP-CDM.

OMOP-CDM DEVICE_EXPOSURE Table

CDM Field	User Guide	Example
device_concept_id	OMOP Standard Vocabulary Concept ID	45767329
unique_device_id	UDI Device Identifier (UDI-DI)	(01)08801234512343
production_id	UDI Production Identifier (UDI-PI)	(10)110500(17)120501 (21)9G837GH234J

- By enhancing the traceability and identification of medical devices through UDI, we expect to improve the quality of medical device safety research and contribute to international collaborative studies.

METHODS & RESULTS

Medical Device Data ETL

- Among all device domain Electronic Data Interchange (EDI) codes managed by the Health Insurance Review and Assessment Service (HIRA) in Korea, 80.02% were mapped to the OMOP standard vocabulary.
- At Severance Hospital, a tertiary hospital in Korea, UDI information is mapped to medical device usage records from 2006 to 2023 and loaded into the DEVICE_EXPOSURE table.
- Among the hospital's medical device management codes, 19,503 (27.9%) were linked with UDI.

Loading UDI (Unique Device Identification) into the Device_exposure table can enhance medical device safety management.

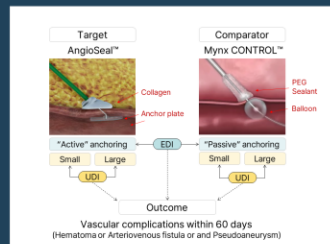


Figure 1. Feasibility study design using device codes. Comparison groups were established, including those that could be compared without the use of UDI (eg. brand name) and those that required UDI for comparison (eg. Specification such as size)

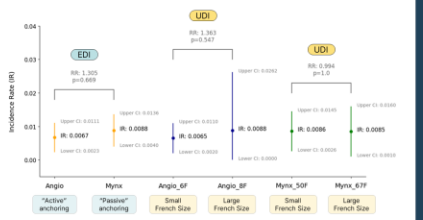


Figure 2. Incidence of vascular complications within 60 days following the use of a vascular closure. The incidence rate (IR) and 95% confidence interval (CI) were calculated for each group. The relative risk(RR) between groups was determined, and Fisher's exact test was performed to calculate the p-value.

Pilot Analysis Design

- Vascular complications are compared according to the closure method and French size of Vascular Closure Devices (VCDs) (Figure 1).
- We identified 1,336 patients using the AngioSeal VCDs and 1,479 patients using the Mynx VCDs (Table 1).

Table 1. Degree of Medical Device Information Coverage by Code

Vascular Closure Devices (VCDs)					
Brand	EDI	Model	UDI-DI	French Size	Patients
Angio-Seal	J4770066	610132	00389701011806	Small (6F)	1,293
		610133	00389701011790	Large (8F)	125
Mynx	J4770213	MX5060E	10862028000441	Small (5F)	1,018
		MX6760E	10862028000458	Large (6F/7F)	597

- The relative risks (RR) were as follows: Angio vs. Mynx (RR= 1.31, p=0.67), Angio_6F vs. Angio_8F (RR= 1.36, p=0.55), Mynx_5F vs. Mynx_6/7F (RR=0.99, p=1.00) (Figure 2).

CONCLUSION

- Incorporating the Unique Device Identifier (UDI) into the DEVICE_EXPOSURE table of OMOP-CDM has demonstrated potential in enhancing medical device data analysis.
- Comparisons between VCD brands can be conducted using claim codes (EDI) without UDI information. However, comparisons based on specific specifications are feasible only when UDI information is mapped.

Seojeong Shin [1], Yiju Park [1,2], Sujeong Eom [1,2], Kyulee Jeon [1,2], Seng Chan You [1,2]

[1]Institute for Innovation in Digital Healthcare, Yonsei University Health System, Korea
[2]Department of Biomedical Systems Informatics, Yonsei University, Korea

YONSEI UNIVERSITY HEALTH SYSTEM

OHDSI

#OHDSISocialShowcase This Week

Tuesday

OMOP on a Data Lake: Addressing the Critical Need for Scalable Solutions in Healthcare Data Management with OHDSI Tools and AWS Services

(Lance Eighme, Lisa McEwen, Simon White, Tobias Cauoette, Oliver Tucher, Anna Swigart)

OHDSI

OMOP on a Data Lake: Addressing the Critical Need for Scalable Solutions in Healthcare Data Management with OHDSI Tools and AWS Services

Lance Eighme, Lisa McEwen, Simon White, Tobias Cauoette, Oliver Tucher, Anna Swigart
Helix, Inc. 101 S Ellsworth Ave #350, San Mateo, CA 94401, United States

Background

The OHDSI community has made significant strides in standardizing healthcare data through the OMOP CDM and stack of open-source tooling for data quality and analytics. However, challenges remain in managing and analyzing vast, complex healthcare datasets:

- Efficiently process billions of records from multiple sources
- Enable rapid, large-scale observational studies

Our project leverages the scalability, performance, and governance capabilities of AWS to develop a comprehensive dataset of over 10 million patients across multiple US health systems that addresses the need for scalable solutions in healthcare data management.

Methods

We implemented a scalable data lake environment combining OHDSI tools with AWS services:

- **Data Ingestion:**
  - Multiple health systems' data ingested into Apache Iceberg tables
  - Optimized for high-performance big data handling and access[1]
- **Data Quality:**
  - Employed AWS Data Quality Definition Language (DQDL)[2]
  - Automated data quality control for OMOP CDM (referential integrity, data types, sql based checks)
- **ETL Processing:**
  - Integrated AWS Glue jobs[3] leveraging Apache Spark for distributed data processing
- **Data Governance:**
  - Implemented AWS Lake Formation[4] for streamlined data lake management
  - Ensured robust security and governance for internal and external partner access
- **Analytics:**
  - Utilized Amazon Redshift Spectrum[5] for queries and analytics workflows avoiding data duplication
  - Integrated with OHDSI tools (DataQualityDashboard[6], Achilles[7], ATLAS[8])

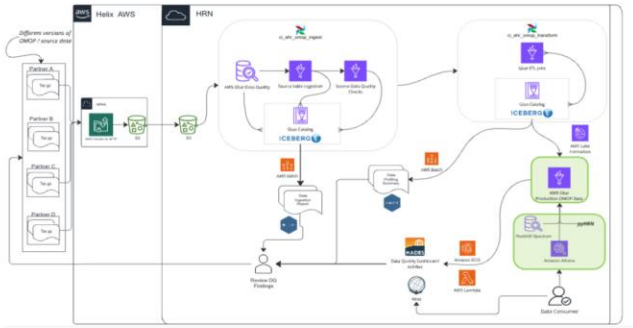
References

1. Ryan Blue, Fokkema S, Vander V, et al. Apache Iceberg: A High-Performance Table Format for Huge Analytic Datasets. Apache Software Foundation. Available from: https://iceberg.apache.org/.
2. AWS Data Quality Definition Language. Amazon Web Services (AWS). Available from: https://docs.aws.amazon.com/glue/latest/dg/dqdl.html.
3. AWS Glue. Amazon Web Services (AWS). Available from: https://docs.aws.amazon.com/glue/
4. AWS Lake Formation. Amazon Web Services (AWS). Available from: https://aws.amazon.com/lake-formation/
5. AWS Redshift Spectrum. Amazon Web Services (AWS). Available from: https://docs.aws.amazon.com/redshift/latest/dg/c-getting-started-using-spectrum.html
6. Blacketer C, Schuemie FJ, Ryan PB, Rijnbeek P (2021). "Increasing trust in real-world evidence through evaluation of observational data quality." Journal of the American Medical Informatics Association, 28(10), 2251-2257. Version 2.6.0.
7. DeFalco F, Ryan P, Schuemie M, Huser V, Knoll C, Londhe A, Abdul-Basser T, Molinaro A (2023). Achilles: Achilles Data Source Characterization. R package version 1.7.2.
8. OHDSI ATLAS. Observational Health Data Sciences and Informatics (OHDSI). Available from: https://atlas.ohdsi.org/.

Results

Deploying OMOP on a scalable data lake architecture with AWS has significantly enhanced the efficiency and scalability of healthcare data management and analytics. With our new pipeline across we have achieved the following:

- Improved data governance and security with AWS Lake Formation and Iceberg tables
- Established scalable and automated data quality infrastructure using AWS Glue DQDL
- Leveraged OHDSI tools, such as DataQualityDashboard, using AWS batch jobs to seamlessly integrate into our Airflow managed pipelines allowing for the identification of data quality issues across the CDM
- Enabled faster and more complex analytical queries using AWS Redshift Spectrum
- Reduced processing time from hours to an average of 5 minutes per OMOP table which has drastically improved our efficiency as compared to using an AWS RDS postgres database
- Allowed for concurrent table and data source runs by using AWS Glue and Iceberg for efficient handling of multiple source datasets.



Conclusions

- We have deployed our OMOP pipelines on a scalable data lake architecture using AWS services significantly improving the efficiency of our data processing times by over 10X.
- Furthermore, with these scalable infrastructure services available within AWS, analysts and researchers are able to conduct timely and in-depth inquiries into these data, pushing the boundaries of observational health data sciences and fostering new insights into healthcare outcomes.
- This approach offers a robust and modern solution to the OHDSI community, addressing critical needs in large-scale data management and analysis.

Contact: contact@ohdsi.org

#OHDSISocialShowcase This Week

Wednesday

Generalizable Approaches for Medical Term Normalization

(Jacob Berkowitz, Yasaman Fatapour, Nicholas Tatonetti)

Generalizable Approaches for Medical Term Normalization

Jacob Berkowitz, Yasaman Fatapour, Nicholas P. Tatonetti
Department of Computational Biomedicine – Cedars-Sinai

Introduction

Approximately 80% of electronic health record (EHR) data consists of unstructured text, complicating the extraction of potentially life-saving medical insights. The complexity of medical language within EHRs presents challenges for downstream analysis. Large language models (LLMs) can offer promising solutions to these challenges by normalizing unstructured text to standardized medical terminologies. Here, we develop and evaluate generalizable approaches for medical text normalization using OpenAI's GPT-4. We selected GPT-4 for its wide availability and ease of use, as the computation is handled remotely, not requiring extensive local resources.

Evaluation

We evaluate our frameworks on their ability to map medical term synonyms to Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) IDs using two datasets: one oncology-specific and one covering a broad range of medical conditions. We generate these datasets using GPT-4 to produce ten synonyms for each term.

Oncology-Specific Dataset: we extracted terms related to "Malignant neoplastic disease" with over 1,000 uses in our institution's OMOP database.

Cross-Domain Dataset: we randomly selected terms from institutional billing codes, ensuring a diverse representation of medical conditions.

To assess performance, we look at the proportion of correct terms identified by the approach.

Approaches

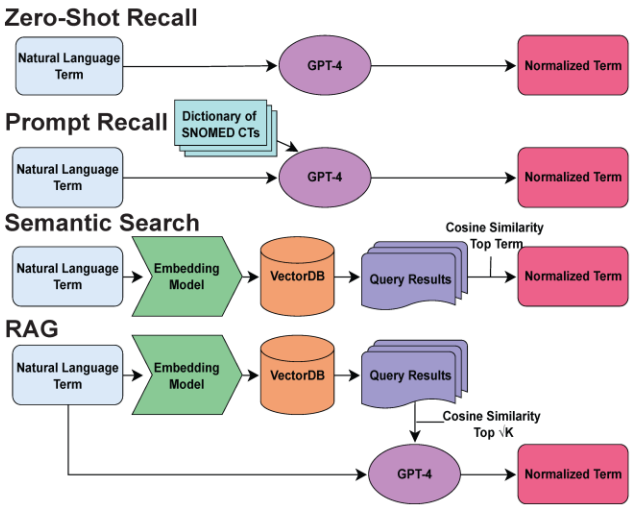
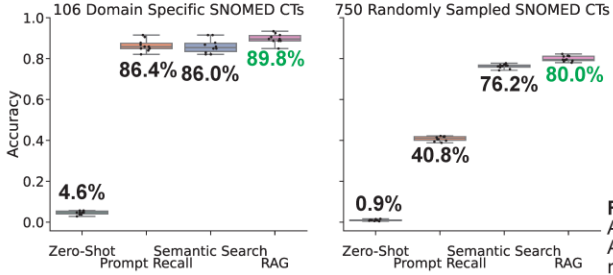


Figure 1: Methodology Flowchart. Step-by-step approach for four different normalization methods.

Zero-Shot Recall: Uses a single question to elicit the correct term from the model without any prior examples or fine-tuning. **Prompt Recall:** Feeds the model a comprehensive list of terms, prompting it to select the most appropriate one based on the input context. **Semantic Search:** Matches input terms with the closest semantic equivalents using a precomputed vector space of embeddings. **Retrieval-Augmented Generation (RAG):** First retrieves most semantically relevant terms and then uses the generative decoder to choose the best matching term.

Results



Conclusion

While all approaches have their merit and may be optimal in specific use-cases, the RAG approach demonstrates the most promise in text normalization to SNOMED CT. Zero-Shot Recall's poor performance may be attributed to lack of specific knowledge, however it correctly identified commonly used SNOMED CT such as "primary malignant neoplasm of female breast" and "primary malignant neoplasm of prostate." Despite Prompt Recall ensuring the LLM has access to the correct term, the increase in irrelevant terminology overwhelms the model and reduces performance. Narrowing the candidate list down through semantic search saves on time and cost, while demonstrating greater performance. This study highlights the potential of LLMs in improving the accuracy and efficiency of EHR data management, which could lead to enhanced patient care and outcomes. Further research is needed to refine these techniques and their implementation in healthcare environments.

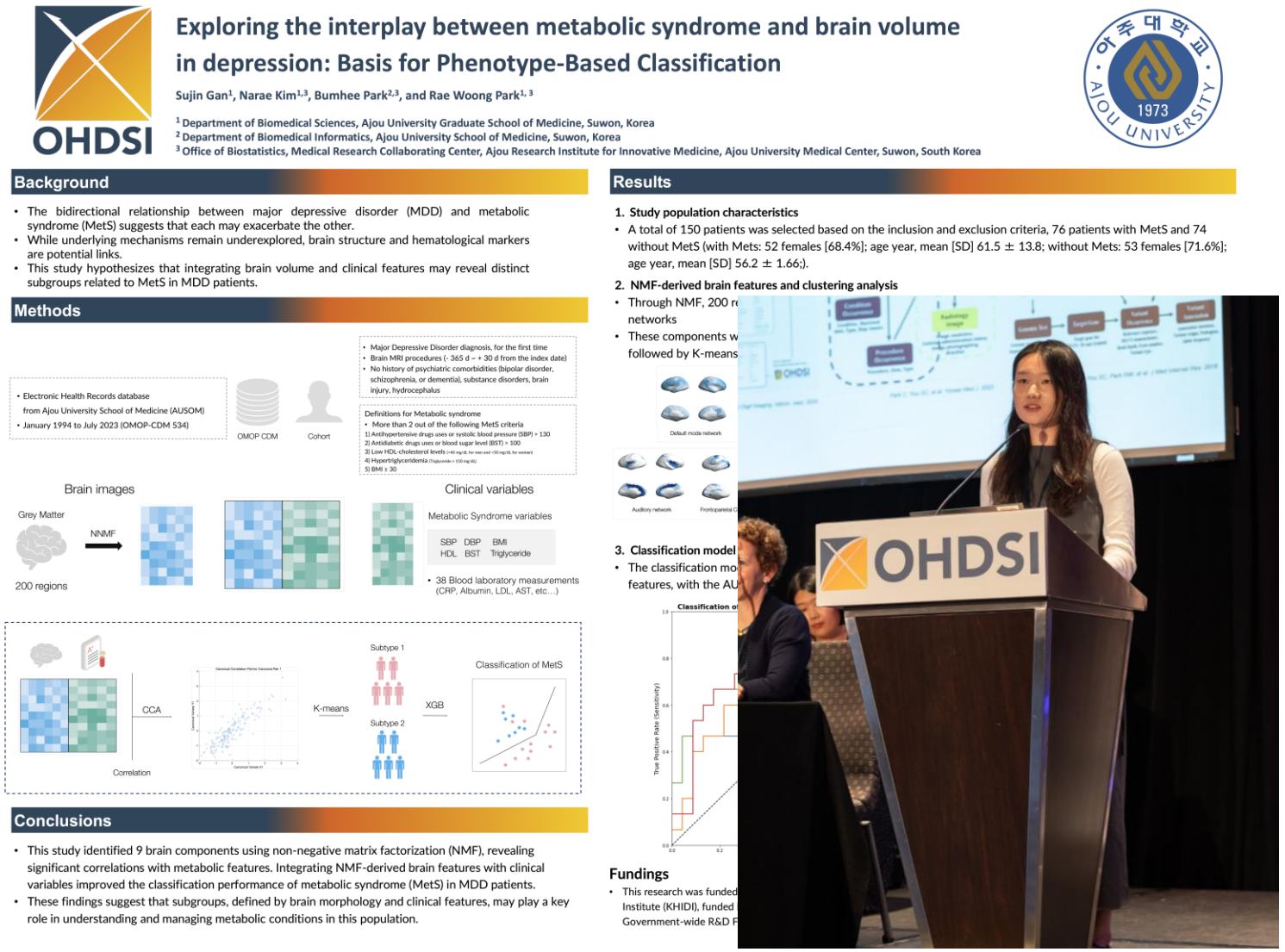
Figure 2: Boxplots of Approach Performance. Accuracy of four different normalization methods

#OHDSISocialShowcase This Week

Thursday

Exploring the interplay between metabolic syndrome and brain volume in depression: Basis for Phenotype-Based Classification

(Sujin Gan (presenter), Narae Kim, Bumhee Park, and Rae Woong Park)



Exploring the interplay between metabolic syndrome and brain volume in depression: Basis for Phenotype-Based Classification

Sujin Gan[1], Narae Kim[1,3], Bumhee Park[2,3], and Rae Woong Park[1, 3]

[1] Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea
[2] Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea
[3] Office of Biostatistics, Medical Research Collaborating Center, Ajou Research Institute for Innovative Medicine, Ajou University Medical Center, Suwon, South Korea

Background

- The bidirectional relationship between major depressive disorder (MDD) and metabolic syndrome (MetS) suggests that each may exacerbate the other.
- While underlying mechanisms remain underexplored, brain structure and hematological markers are potential links.
- This study hypothesizes that integrating brain volume and clinical features may reveal distinct subgroups related to MetS in MDD patients.

Methods

- Electronic Health Records database from Ajou University School of Medicine (AUSOM)
- January 1994 to July 2023 (OMOP-CDM 534)

OMOP CDM — Cohort

- Major Depressive Disorder diagnosis, for the first time
- Brain MRI procedures (- 365 d ~ + 30 d from the index date)
- No history of psychiatric comorbidities (bipolar disorder, schizophrenia, or dementia), substance disorders, brain injury, hydrocephalus

Definitions for Metabolic syndrome
- More than 2 out of the following MetS criteria
1) Antihypertensive drugs uses or systolic blood pressure (SBP) > 130
2) Antidiabetic drugs uses or blood sugar level (BST) > 100
3) Low HDL-cholesterol levels
4) Hypertriglyceridemia
5) BMI ≥ 30

Brain images: Grey Matter → NNMF → 200 regions

Clinical variables: Metabolic Syndrome variables — SBP DBP BMI HDL BST Triglyceride

- 38 Blood laboratory measurements (CRP, Albumin, LDL, AST, etc…)

CCA → Correlation

K-means → Subtype 1 / Subtype 2

XGB → Classification of MetS

Conclusions

- This study identified 9 brain components using non-negative matrix factorization (NMF), revealing significant correlations with metabolic features. Integrating NMF-derived brain features with clinical variables improved the classification performance of metabolic syndrome (MetS) in MDD patients.
- These findings suggest that subgroups, defined by brain morphology and clinical features, may play a key role in understanding and managing metabolic conditions in this population.

Results

1. **Study population characteristics**
- A total of 150 patients was selected based on the inclusion and exclusion criteria, 76 patients with MetS and 74 without MetS (with Mets: 52 females [68.4%]; age year, mean [SD] 61.5 ± 13.8; without Mets: 53 females [71.6%]; age year, mean [SD] 56.2 ± 1.66;).
2. **NMF-derived brain features and clustering analysis**
- Through NMF, 200 r
networks
- These components w
followed by K-means

Default mode network

Auditory network

3. **Classification model**
- The classification mo
features, with the AU

Classification o

True Positive Rate (Sensitivity)

Fundings

- This research was funded
Institute (KHIDI), funded
Government-wide R&D F

#OHDSISocialShowcase This Week

Friday

Quantifying the opioid use disorder crisis: PULSNAR finds nearly 3/4 undiagnosed

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

Quantifying the opioid use disorder crisis: PULSNAR finds nearly 3/4 undiagnosed

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

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

Abstract

The opioid crisis remains a major health concern, with 107,941 drug overdose deaths in the U.S. in 2022, 75.8% of which were opioid-related. The economic burden of opioid use disorder (OUD) in the U.S. surged to nearly $1.5 trillion in 2020. Despite these significant impacts, OUD is often underdiagnosed and undercoded in electronic health records (EHRs), affecting accurate prevalence estimation and impeding intervention efforts. This study applied a novel machine learning approach, "Positive Unlabeled Learning Selected Not At Random (PULSNAR)," to address these challenges and estimate the proportion of undiagnosed OUD cases. SHapley Additive exPlanations (SHAP) were employed to identify and analyze key risk factors and predictors of OUD. Our analysis included 3.34 million individuals exposed to opioids, of whom only 45,019 were diagnosed with OUD. Covariates such as age, sex, medical conditions, and drug exposures were considered. The PULSNAR method identified an additional 124,723 undiagnosed OUD cases, raising the overall OUD prevalence to 5.08%. To our knowledge, this is the first study to demonstrate the potential of Positive and Unlabeled (PU) learning in detecting undiagnosed OUD cases.

Background

The opioid crisis continues to be a significant global public health challenge.[1] In the US, 107,941 drug overdose deaths occurred in 2022, with opioids contributing to 81,806 (75.8%) of these fatalities.[2] The economic burden associated with OUD and fatal opioid overdoses in the US was estimated at $1.02 trillion in 2017 (5.25% of the GDP),[3] escalating to nearly $1.5 trillion in 2020 (7.12% of the GDP).[4]

Accurate estimation and diagnosis of OUD is essential for identifying individuals at risk, assessing treatment needs, monitoring prevention and intervention efforts, and recruiting treatment-naive participants for clinical trials. However, OUD is substantially underdiagnosed and undercoded in EHRs and claims data.[5] This poses a significant challenge in estimating the prevalence of OUD, and in applying cutting-edge machine learning (ML) techniques to model patient outcomes.

To address the issues of underdiagnosis and undercoding of OUD, our study employed a novel Positive and Unlabeled (PU) machine learning (ML) approach, termed "*Positive Unlabeled Learning Selected Not At Random (PULSNAR)*",[6] to estimate the proportion (α) of OUD among undetected individuals. Furthermore, we utilized SHAP[7] values to analyze the relationships between important features and outcomes to understand the underlying risk factors and potential predictors of OUD. To the best of our knowledge, this is the first study to apply PU learning to opioid-related data to estimate the prevalence of undercoding and predict OUD.

Materials and Methods

Figure. 1. Steps to estimate the proportion of uncoded OUD using PULSNAR. PP: Probable positives identified by PULSNAR; PN: Probable negatives identified by PULSNAR; CP: coded positives

- K = floor(|Unlabeled|/|Positive|) = 73
- Covariate classes: Age, sex, condition, and drug ingredient.
- For the PULSNAR method, we used XGBoost[8] as a classifier to estimate the predictions for positive and unlabeled examples.
- We examined the SHAP plot for the top 15 features identified by XGBoost, selected based on their average gain score across 73 balanced dataset models.
- The SHAP plot provided insights into the relationships between the covariates and the predicted outcomes, thereby enhancing the model's interpretability.
- Final α for the full dataset:

$$\alpha = \frac{|U_k|\sum_{i=1}^{k}\alpha_i}{k|U_k|} = \frac{\sum_{i=1}^{k}\alpha_i}{k}$$

Opioid list: *alfentanil, alphaprodine, buprenorphine, butorphanol, codeine, dextromoramide, dezocine, dihydrocodeine, diphenoxylate, ethylmorphine, fentanyl, heroin, hydrocodone, hydromorphone, levomethadyl, levorphanol, meperidine, meptazinol, methadone, methadyl acetate, morphine, nalbuphine, normethadone, opium, oxycodone, oxymorphone, papaveretum, pentazocine, phenazocine, phenoperidine, pirinitramide, propoxyphene, remifentanil, sufentanil, tapentadol, tilidine, tramadol*

Results

- PULSNAR estimated 124,723 additional cases of undiagnosed OUD, representing 3.78% of undiagnosed individuals (95% CI: [3.76%, 3.80%]).
- The cumulative prevalence of OUD among patients who received opioid medication over an average of 3.39 years of observation, was 5.08% across all age groups and sexes. This estimate combines both diagnosed and imputed undiagnosed cases, with 73.5% of the cases being imputed.
- Out of the 94,668 covariates available in our dataset, only 10,190 (10.76%) were utilized by the XGBoost classifier to learn from the data, comprising coded positives, as well as probable positives and negatives identified by the PULSNAR method.

Figure. 2. Distribution of α estimates by PULSNAR method. Each iteration had 73 α estimates, each corresponding to one of the 73 balanced datasets. Red line: mean α value.

Figure. 3. Gain scores for the top 15 features distinguishing patients with and without OUD. The gain score represents the mean gain across 73 balanced datasets and 40 iterations.

Figure. 4. SHAP plot for the top 15 features identified by the XGBoost model across all 73 balanced datasets. The plot shows the effects of these features on XGBoost's OUD prediction for individuals.

Patient Characteristics (n=3,342,044)	Coded for OUD (n=45,019)	Uncoded for OUD (n=3,297,025)
Male	22,858 (51%)	1,415,363 (43%)
Female	22,161 (49%)	1,881,662 (57%)
Age, yr	42 (±13)	38 (±15)
Age, n(%)		
0-19	1,896 (4.21%)	459,981 (13.95%)
20-29	6,606 (14.67%)	534,020 (16.20%)
30-39	10,590 (23.52%)	671,288 (20.36%)
40-49	11,008 (24.45%)	688,924 (20.90%)
50-59	12,249 (27.21%)	761,847 (23.11%)
>=60	2,670 (5.93%)	180,965 (5.49%)
Comorbidities		
Chronic pain syndrome + Chronic pain, not elsewhere classified	18,294 (40.64%)	444,052 (13.47%)
Alcohol related disorders	4,107 (9.12%)	89,620 (2.72%)
Mental and behavioral disorders	27,182 (60.38%)	1,581,777 (47.98%)
Other psychoactive substance related disorders	2,767 (6.15%)	17,753 (0.54%)
Alcohol dependence	2,926 (6.50%)	41,913 (1.27%)
Bipolar disorder	2,823 (6.27%)	65,425 (1.98%)
Cannabis related disorders	2,077 (4.61%)	44,174 (1.34%)
Other stimulant related disorders	981 (2.18%)	8,025 (0.24%)

Table 1: Characteristics of patients with and without coded OUD. These comorbidities are from the list of top important features selected by XGBoost to learn models.

Discussion and Conclusions

- Accurate detection of OUD is crucial for identifying individuals at risk, improving responses to the opioid crisis, expanding access to treatment, guiding public health strategies, enhancing health outcomes, addressing co-occurring conditions, and ultimately saving lives.
- PU learning shows potential in detecting undercoded or undiagnosed OUD cases. The PULSNAR method provides a calibrated probability for each patient being an OUD case, which can be utilized for both screening and probabilistic phenotyping.
- In our study cohort, only 1 in 73 individuals were initially coded for OUD. However, using the PULSNAR method, we estimated that about 1 in 20 individuals exposed to opioids actually have OUD across all age and sex groups. This estimation is consistent with the prevalence ranges reported in other studies, justifying the applicability of PULSNAR.[9,10]
- Treatments for OUD, such as buprenorphine and naloxone were highly predictive of OUD, as well as the presence of chronic pain and treatments for pain (e.g., acetaminophen).
- The lower incidence of OUD among individuals who received a COVID-19 vaccine or tested positive may not indicate a direct causal link, but rather may represent a temporal bias warranting further investigation. Additionally, these individuals may have better healthcare access and management of health conditions, potentially contributing to lower OUD incidence.
- The predictive power of other substance use disorders, including alcoholism, suggests that common causes may underlie multiple substance use conditions.

References

1. Degenhardt L, Grebely J, Stone J, Hickman M, Vickerman P, Marshall BDL, et al. Global patterns of opioid use and dependence: harms to populations, interventions, and future action. Lancet. 2019;394(10208):1560–79.
2. Drug Overdose Death Rates https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates Accessed 5/28/2024
3. Florence C, Luo F, Rice K. The economic burden of opioid use disorder and fatal opioid overdose in the United States, 2017. Drug Alcohol Depend. 2021 Jan 1;218:108350. doi: 10.1016/j.drugalcdep.2020.108350. Epub 2020 Oct 27. PMID: 33121867; PMCID: PMC8091480.
4. The Economic Toll of the Opioid Crisis Reached Nearly $1.5 Trillion in 2020 https://www.jec.senate.gov/public/_cache/files/87bced99-6232-40ea-9263-f035b0280c567/jec-cost-of-opioids-issue-brief.pdf Accessed 5/28/2024
5. Haight SC, Ko JY, Tong VT, Bohm MK, Callaghan WM. Opioid use disorder documented at delivery hospitalization—United States, 1999–2014. Morbidity and Mortality Weekly Report. 2018 Aug 3;67(31):845.
6. Kumar P, Lambert CG. PULSNAR– Positive unlabeled learning selected not at random: class proportion estimation when the SCAR assumption does not hold. arXiv preprint arXiv:2303.08268. 2023 Mar 14.
7. An introduction to explainable AI with Shapley values https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html Accessed 6/20/2024
8. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 (pp. 785-794).
9. Keyes KM, Rutherford C, Hamilton A, Barocas JA, Gerberg MH, Mueller PP, Feaster DJ, El-Bassel N, Cerdá M. What is the prevalence of and trends in opioid use disorder in the United States from 2010 to 2019? Using multiplier approaches to estimate prevalence for an unknown population size. Drug and alcohol dependence reports. 2022 Jun 1;3:100052.
10. Lindner SR, Hart R, Manibusan R, McCarty D, McConnell KJ. State-and county-level geographic variation in opioid use disorder, medication treatment, and opioid-related overdose among medicaid enrollees. In JAMA Health Forum 2023 Jun 2 (Vol. 4, No. 6, pp. e231574-e231574). American Medical Association.

Acknowledgment: This research was supported by the National Institute of Mental Health of the National Institutes of Health under award numbers R01MH129764 and R56MH120826.

Where Are We Going?

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

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

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

Everybody is invited!

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