OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Welcome to OHDSI/ How To Get Started

OHDSI Community Call

Oct. 14, 2025 • 11 am ET

Upcoming Community Calls

Date	Topic
Oct. 14	Welcome to OHDSI
Oct. 21	Tribute to Andrew Williams/The Power of Collaboration
Oct. 28	Meet the Titans
Nov. 4	Collaborator Showcase Honorees
Nov. 11	TBA
Nov. 18	DARWIN EU 2025 Update
Nov. 25	TBA
Dec. 2	OHDSI/OMOP Research Spotlight
Dec. 9	How Did OHDSI Do This Year?
Dec. 16	Holiday Farewell To 2025

Three Stages of The Journey

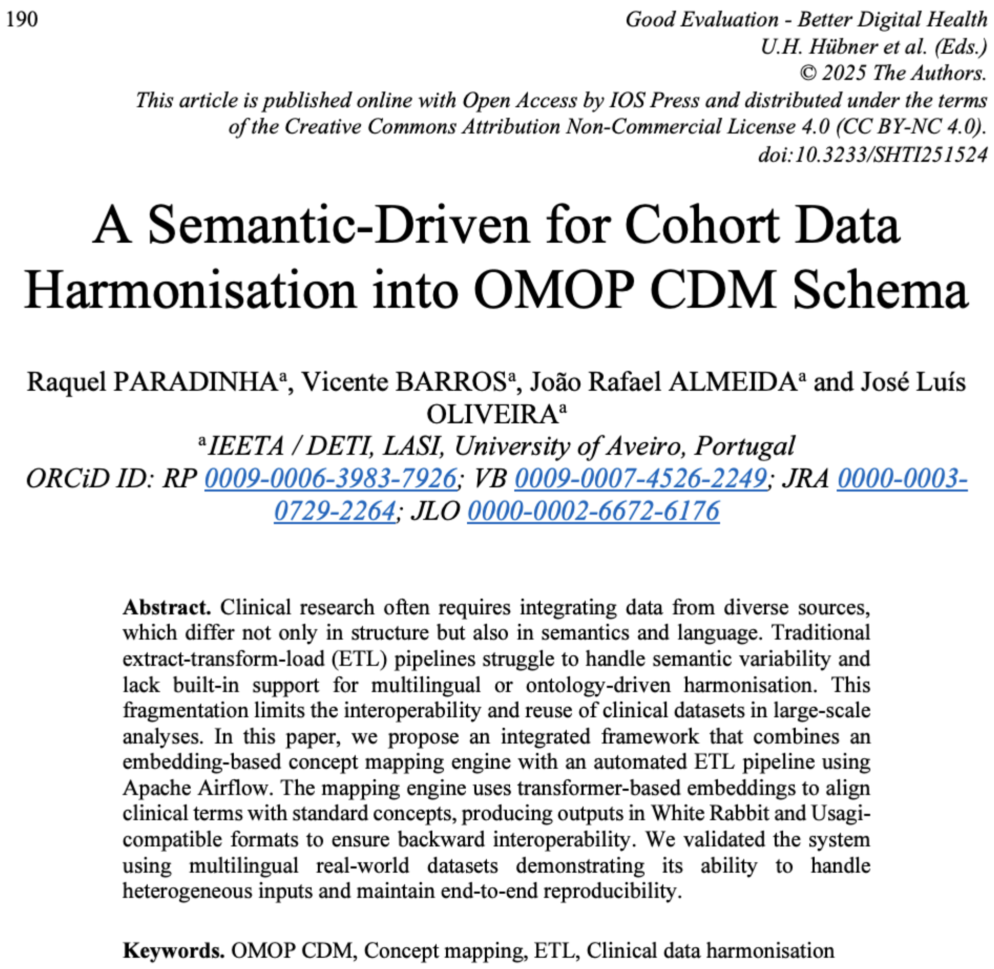
Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Raquel Paradinha, Vicente Barros, João Rafael Almeida, and José Luís Oliveira** on the publication of **A Semantic-Driven for Cohort Data Harmonisation into OMOP CDM Schema** in *Volume 332 of Studies in Health Technology and Informatics: Good Evaluation - Better Digital Health.*



190

Good Evaluation - Better Digital Health
U.H. Hübner et al. (Eds.)
© 2025 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI251524

A Semantic-Driven for Cohort Data Harmonisation into OMOP CDM Schema

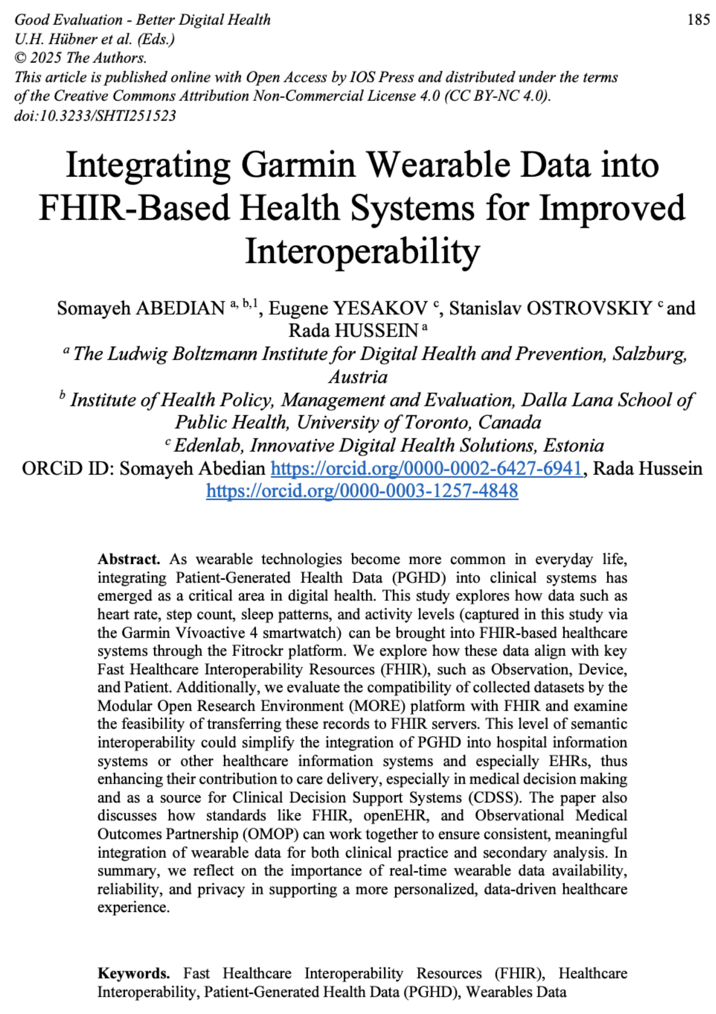
Raquel PARADINHA[a], Vicente BARROS[a], João Rafael ALMEIDA[a] and José Luís OLIVEIRA[a]
[a] *IEETA / DETI, LASI, University of Aveiro, Portugal*
ORCiD ID: RP 0009-0006-3983-7926; VB 0009-0007-4526-2249; JRA 0000-0003-0729-2264; JLO 0000-0002-6672-6176

Abstract. Clinical research often requires integrating data from diverse sources, which differ not only in structure but also in semantics and language. Traditional extract-transform-load (ETL) pipelines struggle to handle semantic variability and lack built-in support for multilingual or ontology-driven harmonisation. This fragmentation limits the interoperability and reuse of clinical datasets in large-scale analyses. In this paper, we propose an integrated framework that combines an embedding-based concept mapping engine with an automated ETL pipeline using Apache Airflow. The mapping engine uses transformer-based embeddings to align clinical terms with standard concepts, producing outputs in White Rabbit and Usagi-compatible formats to ensure backward interoperability. We validated the system using multilingual real-world datasets demonstrating its ability to handle heterogeneous inputs and maintain end-to-end reproducibility.

Keywords. OMOP CDM, Concept mapping, ETL, Clinical data harmonisation

OHDSI Shoutouts!

Congratulations to the team of **Somayeh Abedian, Eugene Yesakov, Stanislav Ostrovskiy, and Rada Hussein** on the publication of **Integrating Garmin Wearable Data into FHIR-Based Health Systems for Improved Interoperability** in *Volume 332 of Studies in Health Technology and Informatics: Good Evaluation - Better Digital Health.*



Good Evaluation - Better Digital Health 185
U.H. Hübner et al. (Eds.)

doi:10.3233/SHTI251523

Integrating Garmin Wearable Data into FHIR-Based Health Systems for Improved Interoperability

Somayeh ABEDIAN [a, b,1], Eugene YESAKOV [c], Stanislav OSTROVSKIY [c] and Rada HUSSEIN [a]
[a] *The Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*
[b] *Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Canada*
[c] *Edenlab, Innovative Digital Health Solutions, Estonia*
ORCiD ID: Somayeh Abedian https://orcid.org/0000-0002-6427-6941, Rada Hussein https://orcid.org/0000-0003-1257-4848

Abstract. As wearable technologies become more common in everyday life, integrating Patient-Generated Health Data (PGHD) into clinical systems has emerged as a critical area in digital health. This study explores how data such as heart rate, step count, sleep patterns, and activity levels (captured in this study via the Garmin Vívoactive 4 smartwatch) can be brought into FHIR-based healthcare systems through the Fitrockr platform. We explore how these data align with key Fast Healthcare Interoperability Resources (FHIR), such as Observation, Device, and Patient. Additionally, we evaluate the compatibility of collected datasets by the Modular Open Research Environment (MORE) platform with FHIR and examine the feasibility of transferring these records to FHIR servers. This level of semantic interoperability could simplify the integration of PGHD into hospital information systems or other healthcare information systems and especially EHRs, thus enhancing their contribution to care delivery, especially in medical decision making and as a source for Clinical Decision Support Systems (CDSS). The paper also discusses how standards like FHIR, openEHR, and Observational Medical Outcomes Partnership (OMOP) can work together to ensure consistent, meaningful integration of wearable data for both clinical practice and secondary analysis. In summary, we reflect on the importance of real-time wearable data availability, reliability, and privacy in supporting a more personalized, data-driven healthcare experience.

Keywords. Fast Healthcare Interoperability Resources (FHIR), Healthcare Interoperability, Patient-Generated Health Data (PGHD), Wearables Data

OHDSI Shoutouts!

Congratulations to the team of **Pawel Rajwa, Angelika Borkowetz, Thomas Abbott, Andrea Alberti, Katharina Beyer, Anders Bjartell, James T Brash, Andrew Chilelli, Eleanor Davies, Bertrand De Meulder, Tamas Fazekas, Asieh Golozar, Ayman Hijazy, Andreas Josefsson, Veeru Kasivisvanathan, Raivo Kolde, Daniel Kotik, Michael S Leapman, Marcin Miszczyk, Rossella Nicoletti, Peter Prinsen, Sebastiaan Remmers, Maria J Ribal, Juan Gómez Rivas, Lara Rodriguez-Sanchez, Monique J Roobol, Emma Smith, Robert Snijder, Carl Steinbeisser, Hein V Stroomberg, Giorgio Gandaglia, Philip Cornford, Susan Evans-Axelsson, James N'Dow, Peter-Paul M Willemse and the PIONEER Consortium** on the publication of **Observational Health Data Analysis of the Cardiovascular Adverse Events of Systemic Treatment in Patients with Metastatic Hormone-sensitive Prostate Cancer: Big Data Analytics Using the PIONEER Platform** in *European Urology Focus.*

ARTICLE IN PRESS

EUROPEAN UROLOGY FOCUS xxx (xxxx) xxx–xxx

available at www.sciencedirect.com
journal homepage: www.europeanurology.com/eufocus

Prostatic Disease

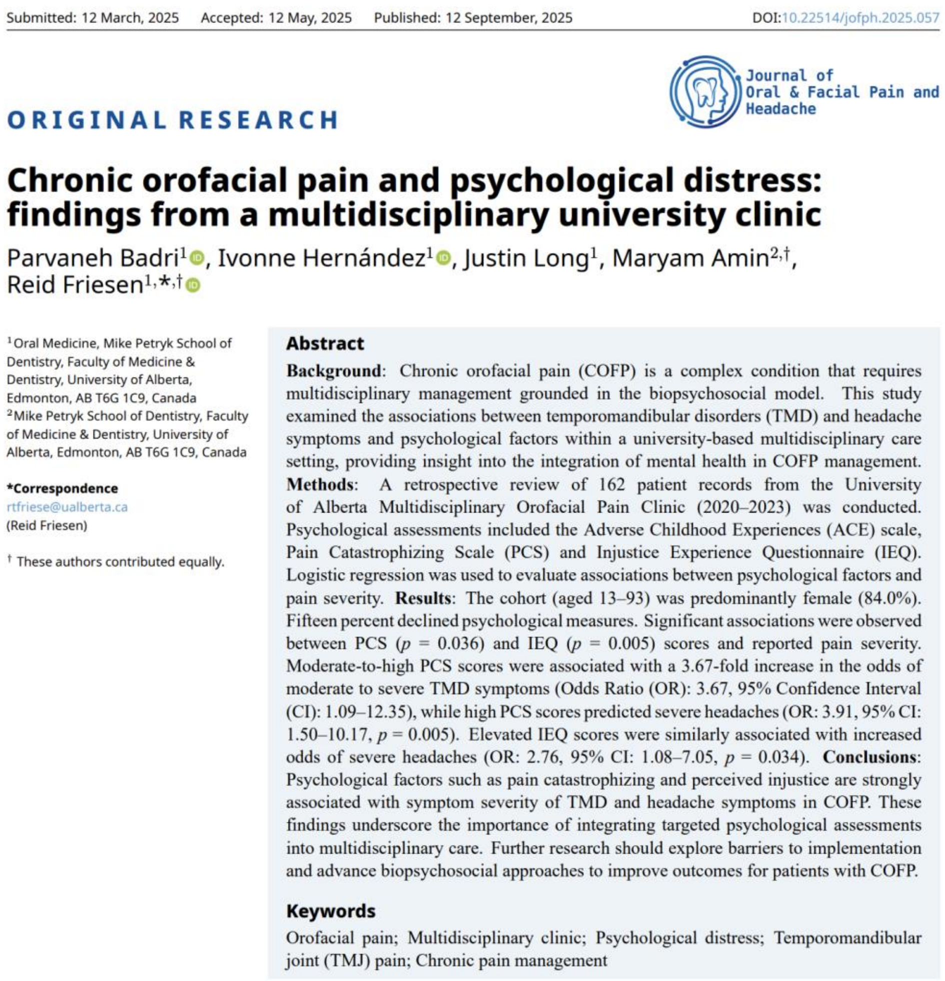
Observational Health Data Analysis of the Cardiovascular Adverse Events of Systemic Treatment in Patients with Metastatic Hormone-sensitive Prostate Cancer: Big Data Analytics Using the PIONEER Platform

Pawel Rajwa [a,b,c,], Angelika Borkowetz [d,e], Thomas Abbott [f], Andrea Alberti [g], Katharina Beyer [h], Anders Bjartell [i], James T. Brash [j], Andrew Chilelli [k], Eleanor Davies [j], Bertrand De Meulder [f,l], Tamas Fazekas [c,m], Asieh Golozar [n,o], Ayman Hijazy [l], Andreas Josefsson [p], Veeru Kasivisvanathan [a], Raivo Kolde [q], Daniel Kotik [r,s], Michael S. Leapman [t], Marcin Miszczyk [c,u], Rossella Nicoletti [g], Peter Prinsen [v], Sebastiaan Remmers [h], Maria J. Ribal [w], Juan Gómez Rivas [x], Lara Rodriguez-Sanchez [y], Monique J. Roobol [h], Emma Smith [z], Robert Snijder [k], Carl Steinbeisser [aa], Hein V. Stroomberg [bb,cc], Giorgio Gandaglia [dd], Philip Cornford [ee], Susan Evans-Axelsson [f,ff], James N'Dow [gg], Peter-Paul M. Willemse [hh], on behalf of the PIONEER Consortium*

[a] *Division of Surgery and Interventional Sciences, University College London and University College London Hospital, London, UK;* [b] *Second Department of Urology, Centre of Postgraduate Medical Education, Warsaw, Poland;* [c] *Department of Urology, Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria;* [d] *Department of Urology, University Hospital, University of Rostock, Rostock, Germany;* [e] *Department of Urology, University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany;* [f] *European Association of Urology, Nijmegen, The Netherlands;* [g] *Unit of Urological Robotic Surgery and Renal Transplantation, University of Florence, Careggi Hospital, Florence, Italy;* [h] *Department of Urology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands;* [i] *Department of Translational Medicine, Lund University, Lund, Sweden;* [j] *IQVIA, Real World Solutions, Brighton, UK;* [k] *Astellas Pharma Europe Ltd., Surrey, UK;* [l] *Association EISBM, Vourles, France;* [m] *Department of Urology, Semmelweis University, Budapest, Hungary;* [n] *Odysseus Data Services, New York, NY, USA;* [o] *OHDSI Center, Northeastern University, Boston, MA, USA;* [p] *Department of Urology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden;* [q] *Institute of Computer Science, University of Tartu, Tartu, Estonia;* [r] *Center for Advanced Systems Understanding, Görlitz, Germany;* [s] *Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany;* [t] *Department of Urology, Yale School of Medicine, New Haven, CT, USA;* [u] *Collegium Medicum - Faculty of Medicine, WSB University, Dąbrowa Górnicza, Poland;* [v] *Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands;* [w] *Uro-Oncology Unit, Hospital Clinic, University of Barcelona, Barcelona, Spain;* [x] *Department of Urology, Hospital Clinico San Carlos, Madrid, Spain;* [y] *Department of Urology, Institut Mutualiste Montsouris, Paris, France;* [z] *Guidelines Office, European Association of Urology, Arnhem, The Netherlands;* [aa] *Collaborate Project Management, Munich, Germany;* [bb] *Copenhagen Prostate Cancer Center, Department of Urology, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark;* [cc] *Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark;* [dd] *Division of Oncology/Unit of Urology, Soldera Prostate Cancer Lab, URI, IRCCS San Raffaele Scientific Institute, Milan, Italy;* [ee] *Liverpool University Hospitals NHS Trust, Liverpool, UK;* [ff] *Bayer AG, Berlin, Germany;* [gg] *Academic Urology Unit, University of Aberdeen, Aberdeen, UK;* [hh] *Department of Urology, Cancer Center, University Medical Center Utrecht, Utrecht, The Netherlands*

OHDSI Shoutouts!

Congratulations to the team of **Parvaneh Badri, Ivonne Hernández, Justin Long, Maryam Amin, and Reid Friesen** on the publication of **Chronic orofacial pain and psychological distress: findings from a multidisciplinary university clinic** in the *Journal of Oral & Facial Pain and Headache.*



Submitted: 12 March, 2025 Accepted: 12 May, 2025 Published: 12 September, 2025 DOI:10.22514/jofph.2025.057

Journal of Oral & Facial Pain and Headache

ORIGINAL RESEARCH

Chronic orofacial pain and psychological distress: findings from a multidisciplinary university clinic

Parvaneh Badri[1], Ivonne Hernández[1], Justin Long[1], Maryam Amin[2,†], Reid Friesen[1,*,†]

[1] Oral Medicine, Mike Petryk School of Dentistry, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB T6G 1C9, Canada
[2] Mike Petryk School of Dentistry, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB T6G 1C9, Canada

***Correspondence**
rtfriese@ualberta.ca
(Reid Friesen)

[†] These authors contributed equally.

Abstract

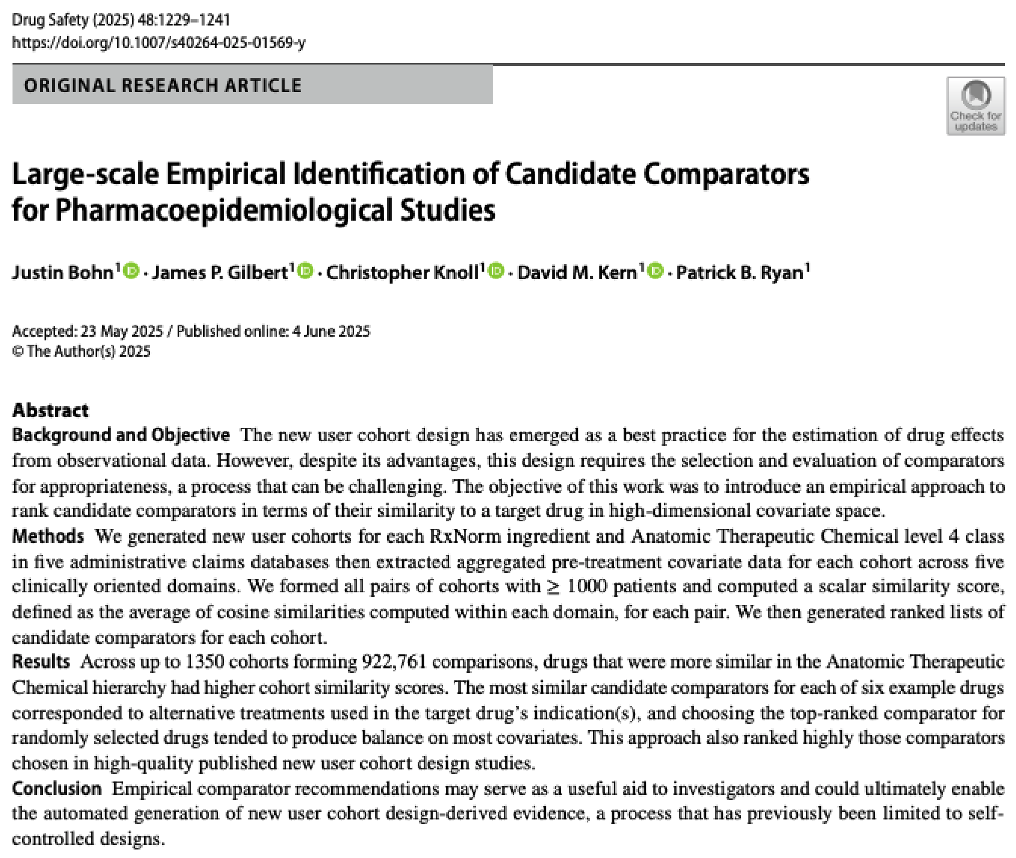
Background: Chronic orofacial pain (COFP) is a complex condition that requires multidisciplinary management grounded in the biopsychosocial model. This study examined the associations between temporomandibular disorders (TMD) and headache symptoms and psychological factors within a university-based multidisciplinary care setting, providing insight into the integration of mental health in COFP management. **Methods**: A retrospective review of 162 patient records from the University of Alberta Multidisciplinary Orofacial Pain Clinic (2020–2023) was conducted. Psychological assessments included the Adverse Childhood Experiences (ACE) scale, Pain Catastrophizing Scale (PCS) and Injustice Experience Questionnaire (IEQ). Logistic regression was used to evaluate associations between psychological factors and pain severity. **Results**: The cohort (aged 13–93) was predominantly female (84.0%). Fifteen percent declined psychological measures. Significant associations were observed between PCS (p = 0.036) and IEQ (p = 0.005) scores and reported pain severity. Moderate-to-high PCS scores were associated with a 3.67-fold increase in the odds of moderate to severe TMD symptoms (Odds Ratio (OR): 3.67, 95% Confidence Interval (CI): 1.09–12.35), while high PCS scores predicted severe headaches (OR: 3.91, 95% CI: 1.50–10.17, p = 0.005). Elevated IEQ scores were similarly associated with increased odds of severe headaches (OR: 2.76, 95% CI: 1.08–7.05, p = 0.034). **Conclusions**: Psychological factors such as pain catastrophizing and perceived injustice are strongly associated with symptom severity of TMD and headache symptoms in COFP. These findings underscore the importance of integrating targeted psychological assessments into multidisciplinary care. Further research should explore barriers to implementation and advance biopsychosocial approaches to improve outcomes for patients with COFP.

Keywords

Orofacial pain; Multidisciplinary clinic; Psychological distress; Temporomandibular joint (TMJ) pain; Chronic pain management

OHDSI Shoutouts!

Congratulations to the team of **Justin Bohn, James P. Gilbert, Christopher Knoll, David M. Kern and Patrick B. Ryan** on the publication of **Large-scale Empirical Identification of Candidate Comparators for Pharmacoepidemiological Studies** in *Drug Safety.*



Drug Safety (2025) 48:1229–1241
https://doi.org/10.1007/s40264-025-01569-y

ORIGINAL RESEARCH ARTICLE

Large-scale Empirical Identification of Candidate Comparators for Pharmacoepidemiological Studies

Justin Bohn[1] · James P. Gilbert[1] · Christopher Knoll[1] · David M. Kern[1] · Patrick B. Ryan[1]

Accepted: 23 May 2025 / Published online: 4 June 2025

Abstract

Background and Objective The new user cohort design has emerged as a best practice for the estimation of drug effects from observational data. However, despite its advantages, this design requires the selection and evaluation of comparators for appropriateness, a process that can be challenging. The objective of this work was to introduce an empirical approach to rank candidate comparators in terms of their similarity to a target drug in high-dimensional covariate space.

Methods We generated new user cohorts for each RxNorm ingredient and Anatomic Therapeutic Chemical level 4 class in five administrative claims databases then extracted aggregated pre-treatment covariate data for each cohort across five clinically oriented domains. We formed all pairs of cohorts with ≥ 1000 patients and computed a scalar similarity score, defined as the average of cosine similarities computed within each domain, for each pair. We then generated ranked lists of candidate comparators for each cohort.

Results Across up to 1350 cohorts forming 922,761 comparisons, drugs that were more similar in the Anatomic Therapeutic Chemical hierarchy had higher cohort similarity scores. The most similar candidate comparators for each of six example drugs corresponded to alternative treatments used in the target drug's indication(s), and choosing the top-ranked comparator for randomly selected drugs tended to produce balance on most covariates. This approach also ranked highly those comparators chosen in high-quality published new user cohort design studies.

Conclusion Empirical comparator recommendations may serve as a useful aid to investigators and could ultimately enable the automated generation of new user cohort design-derived evidence, a process that has previously been limited to self-controlled designs.

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/WebAPI
Tuesday	12 pm	Generative AI and Analytics
Wednesday	8 am	Psychiatry
Wednesday	11 am	Common Data Model
Wednesday	1 pm	Perinatal & Reproductive Health
Wednesday	7 pm	Medical Imaging
Thursday	8 am	India Community Call
Thursday	11 am	Themis
Thursday	12 pm	Medical Devices
Thursday	12 pm	HADES
Thursday	7 pm	Dentistry
Friday	10 am	Transplant
Friday	10 am	GIS-Geographic Information System
Friday	11:30 am	Steering
Monday	10 am	Healthcare Systems
Monday	11 am	Data Bricks User Group
Monday	2 pm	Electronic Animal Health Records
Tuesday	9 am	Data2Evidence

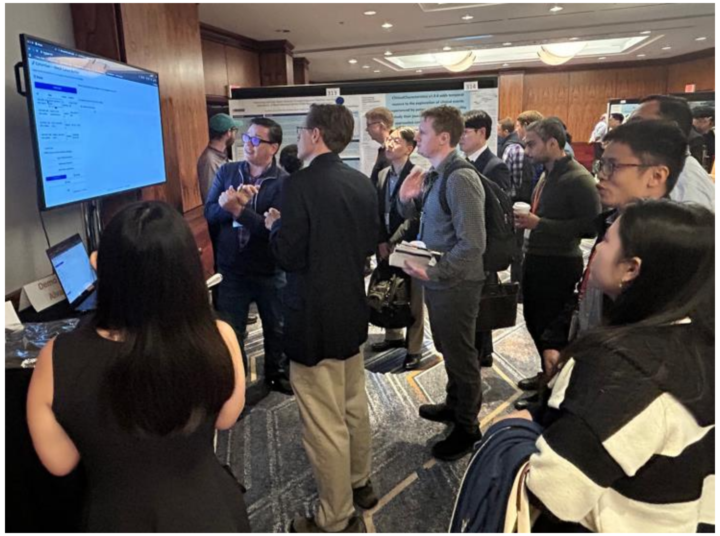
Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Tuesday	12 pm	ATLAS/WebAPI
Tuesday	12 pm	Generative AI and Analytics
Wednesday	8 am	Psychiatry
Wednesday	11 am	Common Data Model
Wednesday	1 pm	Perinatal & Reproductive Health
Wednesday	7 pm	Medical Imaging
Thursday	8 am	India Community Call
Thursday	11 am	Themis
Thursday	12 pm	Medical Devices
Thursday	12 pm	HADES
Thursday	7 pm	Dentistry
Friday	10 am	Transplant
Friday	10 am	GIS-Geographic Information System
Friday	11:30 am	Steering
Monday	10 am	Healthcare Systems
Monday	11 am	Data Bricks User Group
Monday	2 pm	Electronic Animal Health Records
Tuesday	9 am	Data2Evidence

OHDSI 2025

ohdsi.org/ohdsi2025

OHDSI 2025



219
AutoSPICT: Identifying End of Life Care Needs using an electronic implementation of the SPICT™ Questionnaire
OHDSI
Prognostic Risk Prediction using Large Language Models
Aniek Markus & Tom Seinen
LLM delivers reasonable predictive performance below simple but still pro
Aniek Markus
EUROPE

Define (indication) target, comparator, and outcome cohorts
Define Analysis specification
Leverage large-scale propensity score and run analysis
Evaluate Diagnostics
ATLAS/CapR DownloadCohorts.R
CreateAnalysisSpecification.R
StrategusCodeToRun.R
Rshiny App
HYATT REGENCY

Accumulated wisdom from 3 decades of pre- and postmarket safety review in FDA Center for Drug Evaluation and Research
OHDSI
Judy Racoosin
OHDSI

ohdsi.org/ohdsi2025

OHDSI 2025

If you have a great photo from OHDSI2025, please share it! Use the link in the chat (same link for sharing Showcase posters)

ohdsi.org/ohdsi2025

2025 Titan Awards

Data Standards

2025 Titan Awards

Methodological Research

2025 Titan Awards

Open-Source Development

2025 Titan Awards

Clinical Applications

2025 Titan Awards

Community Collaboration

2025 Titan Awards

Community Support

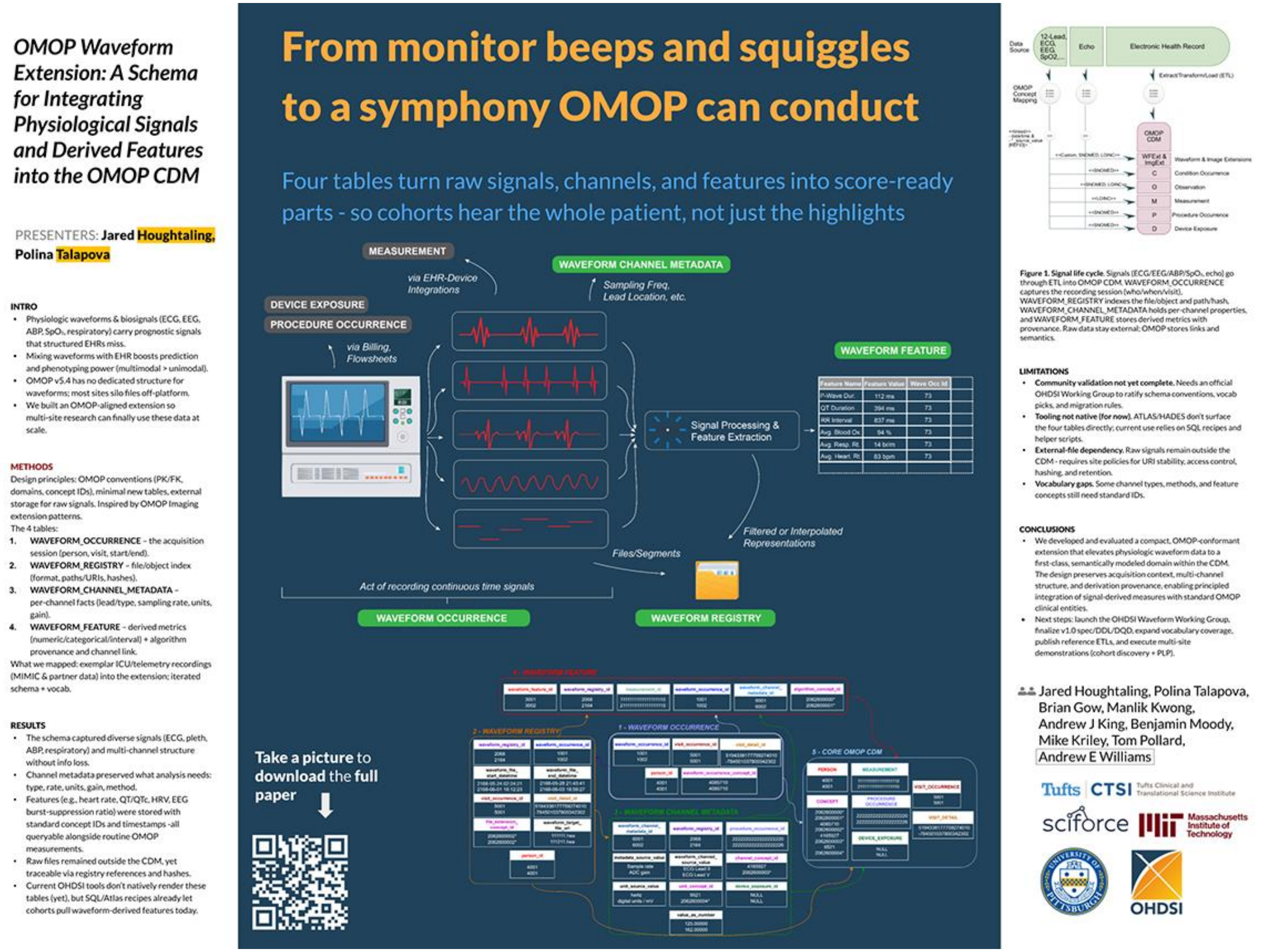
2025 Titan Awards

Community Leadership

Best Community Contribution Winners

Data Standards

Jared Houghtaling, Polina Talapova, Brian Gow, Manlik Kwong, Andrew J King, Benjamin Moody, Mike Kriley, Tom Pollard, Andrew E Williams



OMOP Waveform Extension: A Schema for Integrating Physiological Signals and Derived Features into the OMOP CDM

PRESENTERS: **Jared Houghtaling, Polina Talapova**

INTRO

- Physiologic waveforms & biosignals (ECG, EEG, ABP, SpO₂, respiratory) carry prognostic signals that structured EHRs miss.
- Mixing waveforms with EHR boosts prediction and phenotyping power (multimodal > unimodal).
- OMOP v5.4 has no dedicated structure for waveforms; most sites silo files off-platform.
- We built an OMOP-aligned extension so multi-site research can finally use these data at scale.

METHODS

Design principles: OMOP conventions (PK/FK, domains, concept IDs), minimal new tables, external storage for raw signals. Inspired by OMOP Imaging extension patterns.

The 4 tables:

1. **WAVEFORM_OCCURRENCE** – the acquisition session (person, visit, start/end).
2. **WAVEFORM_REGISTRY** – file/object index (format, paths/URIs, hashes).
3. **WAVEFORM_CHANNEL_METADATA** – per-channel facts (lead/type, sampling rate, units, gain).
4. **WAVEFORM_FEATURE** – derived metrics (numeric/categorical/interval) + algorithm provenance and channel link.

What we mapped: exemplar ICU/telemetry recordings (MIMIC & partner data) into the extension; iterated schema + vocab.

RESULTS

- The schema captured diverse signals (ECG, pleth, ABP, respiratory) and multi-channel structure without info loss.
- Channel metadata preserved what analysis needs: type, rate, units, gain, method.
- Features (e.g., heart rate, QT/QTc, HRV, EEG burst-suppression ratio) were stored with standard concept IDs and timestamps -all queryable alongside routine OMOP measurements.
- Raw files remained outside the CDM, yet traceable via registry references and hashes.
- Current OHDSI tools don't natively render these tables (yet), but SQL/Atlas recipes already let cohorts pull waveform-derived features today.

From monitor beeps and squiggles to a symphony OMOP can conduct

Four tables turn raw signals, channels, and features into score-ready parts - so cohorts hear the whole patient, not just the highlights

Feature Name	Feature Value	Wave Occ Id
P-Wave Dur.	112 ms	73
QT Duration	394 ms	73
RR Interval	837 ms	73
Avg. Blood Ox.	94 %	73
Avg. Resp. Rt.	14 br/m	73
Avg. Heart. Rt.	83 bpm	73

Take a picture to download the full paper

Figure 1. Signal life cycle. Signals (ECG/EEG/ABP/SpO₂, echo) go through ETL into OMOP CDM. WAVEFORM_OCCURRENCE captures the recording session (who/when/visit), WAVEFORM_REGISTRY indexes the file/object and path/hash, WAVEFORM_CHANNEL_METADATA holds per-channel properties, and WAVEFORM_FEATURE stores derived metrics with provenance. Raw data stay external; OMOP stores links and semantics.

LIMITATIONS

- **Community validation not yet complete.** Needs an official OHDSI Working Group to ratify schema conventions, vocab picks, and migration rules.
- **Tooling not native (for now).** ATLAS/HADES don't surface the four tables directly; current use relies on SQL recipes and helper scripts.
- **External-file dependency.** Raw signals remain outside the CDM - requires site policies for URI stability, access control, hashing, and retention.
- **Vocabulary gaps.** Some channel types, methods, and feature concepts still need standard IDs.

CONCLUSIONS

- We developed and evaluated a compact, OMOP-conformant extension that elevates physiologic waveform data to a first-class, semantically modeled domain within the CDM. The design preserves acquisition context, multi-channel structure, and derivation provenance, enabling principled integration of signal-derived measures with standard OMOP clinical entities.
- Next steps: launch the OHDSI Waveform Working Group, finalize v1.0 spec/DDL/DQD, expand vocabulary coverage, publish reference ETLs, and execute multi-site demonstrations (cohort discovery + PLP).

Jared Houghtaling, Polina Talapova, Brian Gow, Manlik Kwong, Andrew J King, Benjamin Moody, Mike Kriley, Tom Pollard, Andrew E Williams

Tufts | CTSI Tufts Clinical and Translational Science Institute

sciforce

Massachusetts Institute of Technology

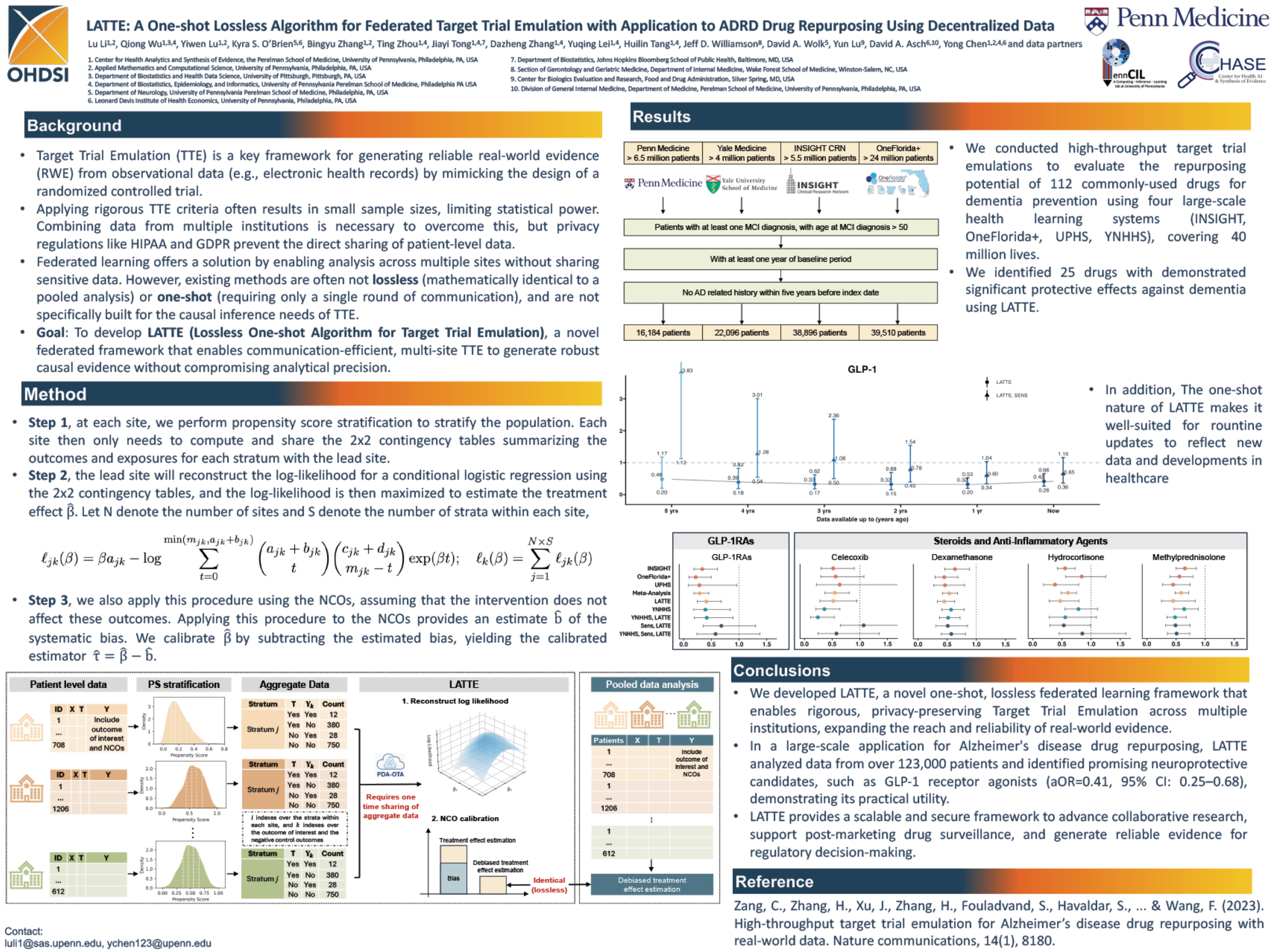
OHDSI

ohdsi.org/2025-global-collaborator-showcase

Best Community Contribution Winners

Methods Research

Lu Li, Qiong Wu, Yiwen Lu, Kyra S. O'Brien, Bingyu Zhang, Ting Zhou, Jiayi Tong, Dazheng Zhang, Yuqing Lei, Huilin Tang, Yun Lu, David Asch, Yong Chen



LATTE: A One-shot Lossless Algorithm for Federated Target Trial Emulation with Application to ADRD Drug Repurposing Using Decentralized Data

Lu Li[1,2], Qiong Wu[1,3,4], Yiwen Lu[1,2], Kyra S. O'Brien[5,6], Bingyu Zhang[1,2], Ting Zhou[1,4], Jiayi Tong[1,4,7], Dazheng Zhang[1,4], Yuqing Lei[1,4], Huilin Tang[1,4], Jeff D. Williamson[8], David A. Wolk[5], Yun Lu[9], David A. Asch[6,10], Yong Chen[1,2,4,6] and data partners

1. Center for Health Analytics and Synthesis of Evidence, the Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
2. Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, PA, USA
3. Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, PA, USA
4. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia PA USA
5. Department of Neurology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
6. Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, USA
7. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
8. Section of Gerontology and Geriatric Medicine, Department of Internal Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA
9. Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA
10. Division of General Internal Medicine, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Background

- Target Trial Emulation (TTE) is a key framework for generating reliable real-world evidence (RWE) from observational data (e.g., electronic health records) by mimicking the design of a randomized controlled trial.
- Applying rigorous TTE criteria often results in small sample sizes, limiting statistical power. Combining data from multiple institutions is necessary to overcome this, but privacy regulations like HIPAA and GDPR prevent the direct sharing of patient-level data.
- Federated learning offers a solution by enabling analysis across multiple sites without sharing sensitive data. However, existing methods are often not **lossless** (mathematically identical to a pooled analysis) or **one-shot** (requiring only a single round of communication), and are not specifically built for the causal inference needs of TTE.
- **Goal**: To develop **LATTE (Lossless One-shot Algorithm for Target Trial Emulation)**, a novel federated framework that enables communication-efficient, multi-site TTE to generate robust causal evidence without compromising analytical precision.

Method

- **Step 1**, at each site, we perform propensity score stratification to stratify the population. Each site then only needs to compute and share the 2x2 contingency tables summarizing the outcomes and exposures for each stratum with the lead site.
- **Step 2**, the lead site will reconstruct the log-likelihood for a conditional logistic regression using the 2x2 contingency tables, and the log-likelihood is then maximized to estimate the treatment effect $\hat{\beta}$. Let N denote the number of sites and S denote the number of strata within each site,

$$\ell_{jk}(\beta) = \beta a_{jk} - \log \sum_{t=0}^{\min(m_{jk}, a_{jk}+b_{jk})} \binom{a_{jk}+b_{jk}}{t}\binom{c_{jk}+d_{jk}}{m_{jk}-t} \exp(\beta t); \quad \ell_k(\beta) = \sum_{j=1}^{N\times S} \ell_{jk}(\beta)$$

- **Step 3**, we also apply this procedure using the NCOs, assuming that the intervention does not affect these outcomes. Applying this procedure to the NCOs provides an estimate $\hat{b}$ of the systematic bias. We calibrate $\hat{\beta}$ by subtracting the estimated bias, yielding the calibrated estimator $\hat{\tau} = \hat{\beta} - \hat{b}$.

Patient level data | PS stratification | Aggregate Data | LATTE | Pooled data analysis

ID	X	T	Y
1			Include outcome of interest and NCOs
...			
708			

ID	X	T	Y
1			
...			
1206			

ID	X	T	Y
1			
...			
612			

Stratum	T	Y_k	Count
Stratum *j*	Yes	Yes	12
	Yes	No	380
	No	Yes	28
	No	No	750

Stratum	T	Y_k	Count
Stratum *j*	Yes	Yes	12
	Yes	No	380
	No	Yes	28
	No	No	750

Stratum	T	Y_k	Count
Stratum *j*	Yes	Yes	12
	Yes	No	380
	No	Yes	28
	No	No	750

l indexes over the strata within each site, and *k* indexes over the outcome of interest and the negative control outcomes

PDA-OTA

Requires one time sharing of aggregate data

1. Reconstruct log likelihood

2. NCO calibration

Treatment effect estimation

Bias

Debiased treatment effect estimation

Identical (lossless)

Patients	X	T	Y
1			Include outcome of interest and NCOs
...			
708			
1			
...			
1206			
1			
...			
612			

Debiased treatment effect estimation

Results

Penn Medicine > 6.5 million patients	Yale Medicine > 4 million patients	INSIGHT CRN > 5.5 million patients	OneFlorida+ > 24 million patients

Patients with at least one MCI diagnosis, with age at MCI diagnosis > 50

With at least one year of baseline period

No AD related history within five years before index date

16,184 patients	22,096 patients	38,896 patients	39,510 patients

- We conducted high-throughput target trial emulations to evaluate the repurposing potential of 112 commonly-used drugs for dementia prevention using four large-scale health learning systems (INSIGHT, OneFlorida+, UPHS, YNHHS), covering 40 million lives.
- We identified 25 drugs with demonstrated significant protective effects against dementia using LATTE.

GLP-1

- In addition, The one-shot nature of LATTE makes it well-suited for rountine updates to reflect new data and developments in healthcare

GLP-1RAs | Steroids and Anti-Inflammatory Agents

Conclusions

- We developed LATTE, a novel one-shot, lossless federated learning framework that enables rigorous, privacy-preserving Target Trial Emulation across multiple institutions, expanding the reach and reliability of real-world evidence.
- In a large-scale application for Alzheimer's disease drug repurposing, LATTE analyzed data from over 123,000 patients and identified promising neuroprotective candidates, such as GLP-1 receptor agonists (aOR=0.41, 95% CI: 0.25–0.68), demonstrating its practical utility.
- LATTE provides a scalable and secure framework to advance collaborative research, support post-marketing drug surveillance, and generate reliable evidence for regulatory decision-making.

Reference

Zang, C., Zhang, H., Xu, J., Zhang, H., Fouladvand, S., Havaldar, S., ... & Wang, F. (2023). High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data. Nature communications, 14(1), 8180.

Contact:
luli1@sas.upenn.edu, ychen123@upenn.edu

ohdsi.org/2025-global-collaborator-showcase

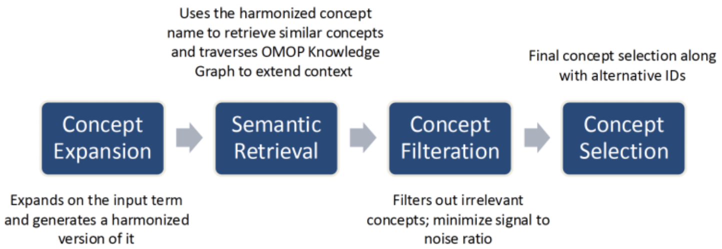
Best Community Contribution Winners

Open-Source Community

Adil Ahmed, Selvin Soby, Boudewijn Aasman, Parsa Mirhaji

Summary

A LLM-workflow that maps clinical terminologies to standard OMOP concepts. The pipeline consists of 4 stages:



Uses the harmonized concept name to retrieve similar concepts and traverses OMOP Knowledge Graph to extend context

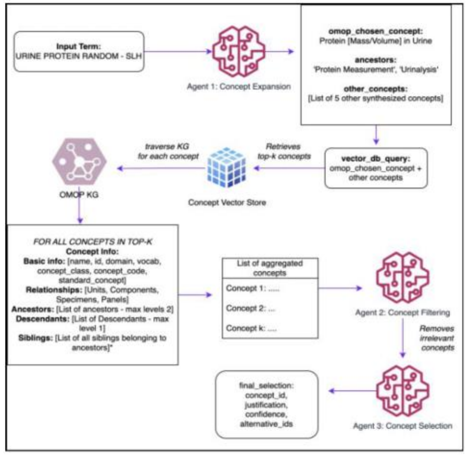
Final concept selection along with alternative IDs

Concept Expansion → Semantic Retrieval → Concept Filtration → Concept Selection

Expands on the input term and generates a harmonized version of it

Filters out irrelevant concepts; minimize signal to noise ratio

Pipeline Workflow


Input Term:
URINE PROTEIN RANDOM - SLH
Agent 1: Concept Expansion
omop_chosen_concept:
Protein [Mass/Volume] in Urine
ancestors:
'Protein Measurement', 'Urinalysis'
other_concepts:
[List of 5 other synthesized concepts]
vector_db_query:
omop_chosen_concept + other concepts
Retrieves top-k concepts
Concept Vector Store
traverse KG for each concept
OMOP KG
FOR ALL CONCEPTS IN TOP-K
Concept Info:
Basic info: [name, id, domain, vocab, concept_class, concept_code, standard_concept]
Relationships: [Units, Components, Specimens, Panels]
Ancestors: [List of ancestors - max levels 2]
Descendants: [List of Descendants - max level 1]
Siblings: [List of all siblings belonging to ancestors]*
List of aggregated concepts
Concept 1: .....
Concept 2: ....
Concept k: .....
Agent 2: Concept Filtering
Removes irrelevant concepts
Agent 3: Concept Selection
final_selection:
concept_id,
justification,
confidence,
alternative_ids

ohdsi.org/2025-global-collaborator-showcase

Best Community Contribution Winners

Clinical Applications

Hsin Yi “Cindy” Chen, Thomas Falconer, Anna Ostropolets, Tara V. Anand, Xinzhuo Jiang, David Dávila-García, Linying Zhang, Ruochong Fan, Hannah Morgan-Cooper, George Hripcsak

OHDSI

Heterogeneity of Treatment Effects Across Nine Glucose-Lowering Drug Classes in Type 2 Diabetes: Extension of the LEGEND-T2DM Network Study

Hsin Yi Chen BS[1], Thomas Falconer MS[1], Anna Ostropolets MD, PhD[1,2], Tara V. Anand MA, MPhil[1], Xinzhuo Jiang MS[1], David Dávila-García MA[3], Linying Zhang PhD[3], Ruochong Fan MA[3], Hannah Morgan-Cooper MSci[4], George Hripcsak MD, MS[1]

[1]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, [2]Janssen Research & Development, Titusville, NJ, [3]Institute for Informatics, Data Science and Biostatistics, Washington University in St. Louis, St. Louis, MO, [4] Stanford School of Medicine and Stanford Health Care, Palo Alto, CA

Background

- Type 2 diabetes mellitus (T2DM) is a major cause of morbidity and mortality, affecting more than 525 million people globally.
- OHDSI's LEGEND-T2DM study [1, 2] investigated the relative effects of different antihyperglycemic agents on cardiovascular risk and patient-centered safety outcomes
- However, T2DM patients are a heterogeneous group, varying in terms of demographics and comorbidities, which may modify the benefits and risks associated with different drugs.
- Here, we extend the LEGEND-T2DM study by studying the comparative effectiveness and safety of T2DM drugs and whether they differ significantly based on patient characteristics.

Methods

- We conducted a large-scale, multinational, real-world comparative effectiveness and safety study, extending the LEGEND-T2DM study
- **Study Design:**
  - Target Cohorts: Adults (≥18 years of age) diagnosed with T2DM who initiated treatment with a drug agent from one of the nine specified glucose-lowering drug classes: (1) Alpha-Glucosidase Inhibitors, (2) Biguanides, (3) Dipeptidyl Peptidase-4 Inhibitors (DPP-4i), (4) dual Glucose-dependent Insulinotropic Polypeptide (GIP) and Glucagon-Like Peptide-1 Receptor Agonists (GLP-1RA), (5) GLP-1RA, (6) Meglitinides, (7) Sodium-Glucose Cotransporter-2 Inhibitors (SGLT2-i), (8) Sulfonylureas (SU), and (9) Thiazolidinediones.
  - Outcomes of interest: Acute myocardial infarction, acute renal failure, hospitalization for heart failure, stroke, abnormal weight gain, acute pancreatitis, diabetic ketoacidosis, diarrhea, hypoglycemia, vomiting, and hepatic failure.
  - We calculated heterogeneity of treatment effects (HTE) across 10 pre-defined subgroups, which can be seen in Table 1.
- To calculate HTE, we (1) calculated calibrated hazard ratios (HR) for each target-comparator-outcome-subgroup combination, then (2) calculated the difference between the log transformed HRs, $\Delta\ln(HR) = \ln\left(HR_{\text{subgroup1}}\right) - \ln\left(HR_{\text{subgroup2}}\right)$, between subgroups within the same target-comparator-outcome comparison.

Subgroup	Comparisons
Age	Pairwise comparisons between <21 years, 21-60 years, and >60 years
Biological Sex	Comparison between Female and Male (as recorded in database based on SNOMED codes)
Renal Impairment	Stratified into three categories based on diagnosis codes for chronic kidney disease (CKD) and end-stage renal disease (ESRD), dialysis procedures, and relevant laboratory measures (e.g., estimated glomerular filtration rate (eGFR), serum creatinine, urine albumin). We compare No renal impairment (NRI) vs Renal impairment without dialysis (RI-ND), and NRI vs Renal impairment on dialysis (RI-D)
Obesity	Presence of obesity, defined as a Body Mass Index (BMI) >30 kg/m², body weight >120 kg, or a diagnosis code for obesity. Compared against the non-obese subgroup.
Poorly Controlled Diabetes	Defined as an HbA1c >8% (64 mmol/mol) or a diagnosis code indicating uncontrolled type 2 diabetes or poor diabetes control. Compared against those not meeting these criteria.
Diabetic Ketoacidosis (DKA)	History of DKA based on diagnosis codes. Compared against those with no history of DKA.
Diabetic Retinopathy	History of diabetic retinopathy based on diagnosis codes. Compared against those with no history of diabetic retinopathy.
Essential Hypertension	History of essential hypertension based on diagnosis codes. Compared against those with no history of essential hypertension.
Hyperlipidemia	History of hyperlipidemia based on diagnosis codes. Compared against those with no history of hyperlipidemia.
Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD)	History of MASLD (including non-alcoholic fatty liver disease or non-alcoholic steatohepatitis) based on diagnosis codes. Compared against those with no history of MASLD.

Table 1. Subgroups under comparison.

Results

- This study was run on a total of 6 databases, 5 of which had results that passed diagnostics.
- Some subgroup comparisons were excluded due to failed diagnostics. Of the subgroup comparisons that passed diagnostics (Figure 1), there were signals of HTE in the hyperlipidemia, obesity, hypertension, and gender subgroups.
- Our findings reflected well-documented pharmacologic patterns but also identified potential areas subgroup heterogeneity, ex:
  - Lower risk of heart failure hospitalization with biguanide (vs. DPP-4i) and lower risk of stroke with GLP-1 RA (vs. SGLT-2i) for obese patients
  - Higher risk of diarrhea on GLP-1 RA (vs. DPP-4i) for females, consistent with literature showing that women experience GI side effects with GLP-1 RAs at roughly twice the rate of men [3], which could potentially be attributed to gender differences in adverse event reporting [4].

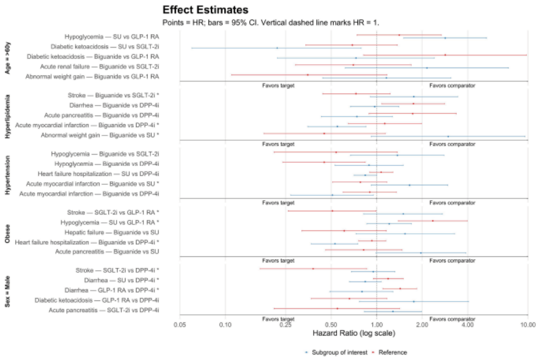


Figure 1. HR estimates of the “most significant” (5 lowest p-values) comparisons in each subgroup, and their associated 95% confidence intervals. Asterisks denote that $\ln(HR)$ is significantly different between the subgroups.

Conclusions

- This hypothesis-generating study identified several potential signals where there may exist treatment effect heterogeneity for several classes of T2DM drugs.
- While many findings did not meet the significance threshold, this preliminary study highlight the potential for personalized T2DM treatment recommendations based on patient characteristics.

References

[1] R. Khera *et al.*, "Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (LEGEND-T2DM): a protocol for a series of multinational, real-world comparative cardiovascular effectiveness and safety studies," *BMJ Open*, vol. 12, no. 6, p. e057977, June 2022.

[2] M. D. Rohan Khera *et al.*, "Comparative Effectiveness of Second-Line Antihyperglycemic Agents for Cardiovascular Outcomes: A Multinational, Federated Analysis of LEGEND-T2DM," *Journal of the American College of Cardiology*, Sept. 2024.

[3] K.-I. Joung, G.-W. Jung, H.-H. Park, H. Lee, S.-H. Park, and J.-Y. Shin, "Gender differences in adverse event reports associated with antidiabetic drugs," *Sci Rep*, vol. 10, no. 1, p. 17545, Oct. 2020.

[4] S. Watson, O. Caster, P. A. Rochon, and H. den Ruijter, "Reported adverse drug reactions in women and men: Aggregated evidence from globally collected individual case reports during half a century," *EClinicalMedicine*, vol. 17, p. 100188, Dec. 2019.

Contact: hc3292@cumc.columbia.edu

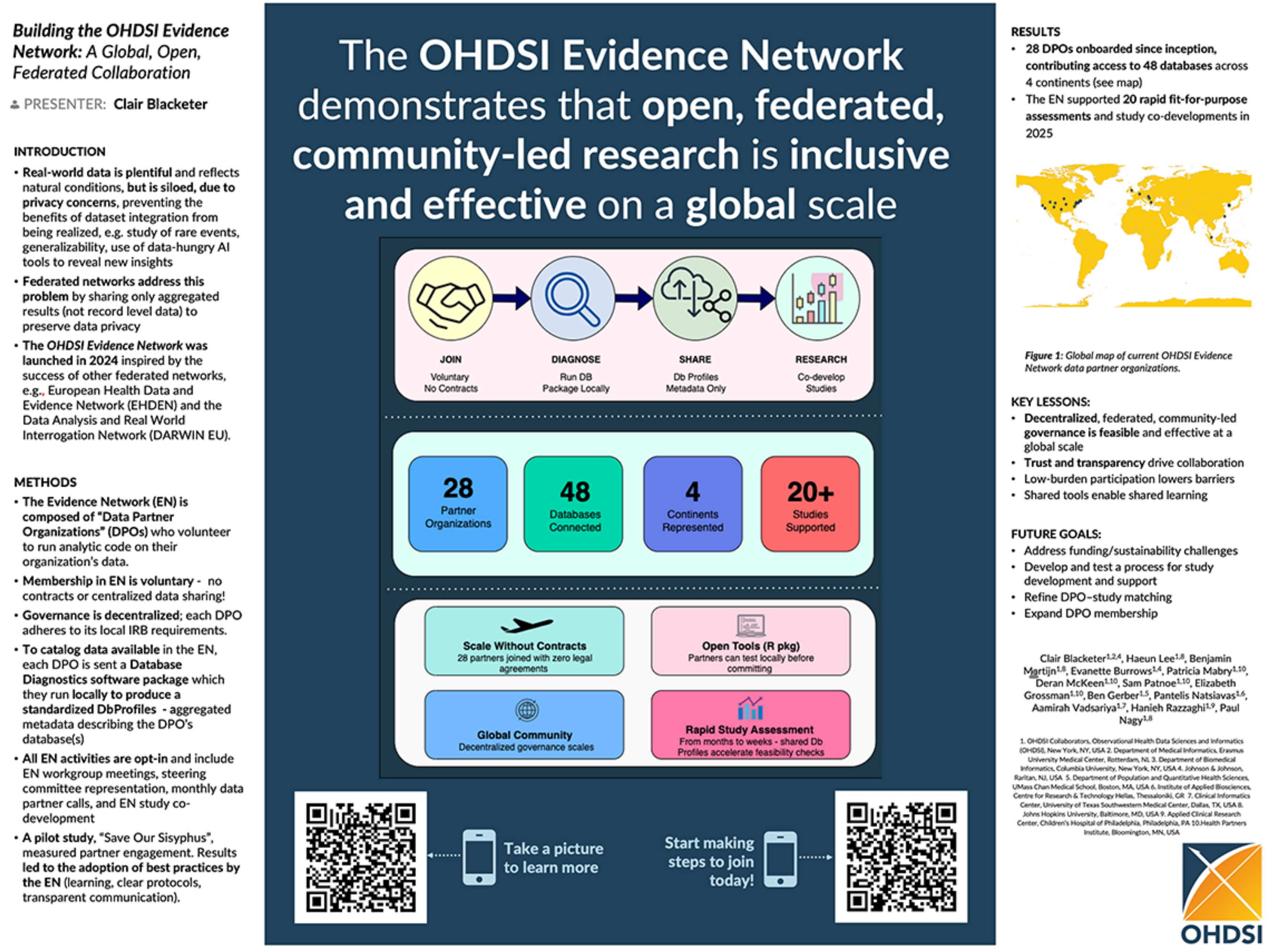
COLUMBIA | Columbia University Department of Biomedical Informatics

ohdsi.org/2025-global-collaborator-showcase

Best Community Contribution Winners

Community

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



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

PRESENTER: **Clair Blacketer**

INTRODUCTION

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

METHODS

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

The **OHDSI Evidence Network** demonstrates that **open, federated, community-led research** is **inclusive and effective** on a **global** scale

JOIN
Voluntary
No Contracts

DIAGNOSE
Run DB
Package Locally

SHARE
Db Profiles
Metadata Only

RESEARCH
Co-develop
Studies

28
Partner Organizations

48
Databases Connected

4
Continents Represented

20+
Studies Supported

Scale Without Contracts
28 partners joined with zero legal agreements

Open Tools (R pkg)
Partners can test locally before committing

Global Community
Decentralized governance scales

Rapid Study Assessment
From months to weeks - shared Db Profiles accelerate feasibility checks

Take a picture to learn more

Start making steps to join today!

RESULTS

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

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

KEY LESSONS:

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

FUTURE GOALS:

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

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

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

OHDSI

ohdsi.org/2025-global-collaborator-showcase

Africa Symposium: Nov. 10-12

The first-ever OHDSI Africa Symposium will be held Nov. 10-12 in Kampala, Uganda, at the Joint Clinical Research Centre (JCRC) and Mestil Hotel. The event will begin with a dedicated one-day training course at JCRC, followed by a two-day main conference at the Mestil Hotel.

OHDSI AFRICA Symposium

Hosted by:
Joint Clinical Research Centre
KAMPALA, UGANDA

Abstract Submission Deadline
August 25, 2025

Save The Date
NOV 10th - 12th 2025

About the Symposium:
Discover how interoperable healthcare data and federated networks can unlock real-world insights while safeguarding patient privacy. This innovative approach – bringing analyses to the data – empowers large-scale, global collaboration without compromising data ownership. A key highlight will be on advancing data-driven healthcare solutions, with a special focus on HIV.

Organised by JCRC in collaboration with:

www.jcrc.org.ug JCRC

OHDSI Uganda

ohdsi.org/africa2025

APAC Symposium: Dec. 6-7

The 2025 OHDSI APAC Symposium will be held Dec. 6-7 in Shanghai, China at the Shanghai Jiao Tong University. It will feature a 1-day tutorial and a 1-day main conference.

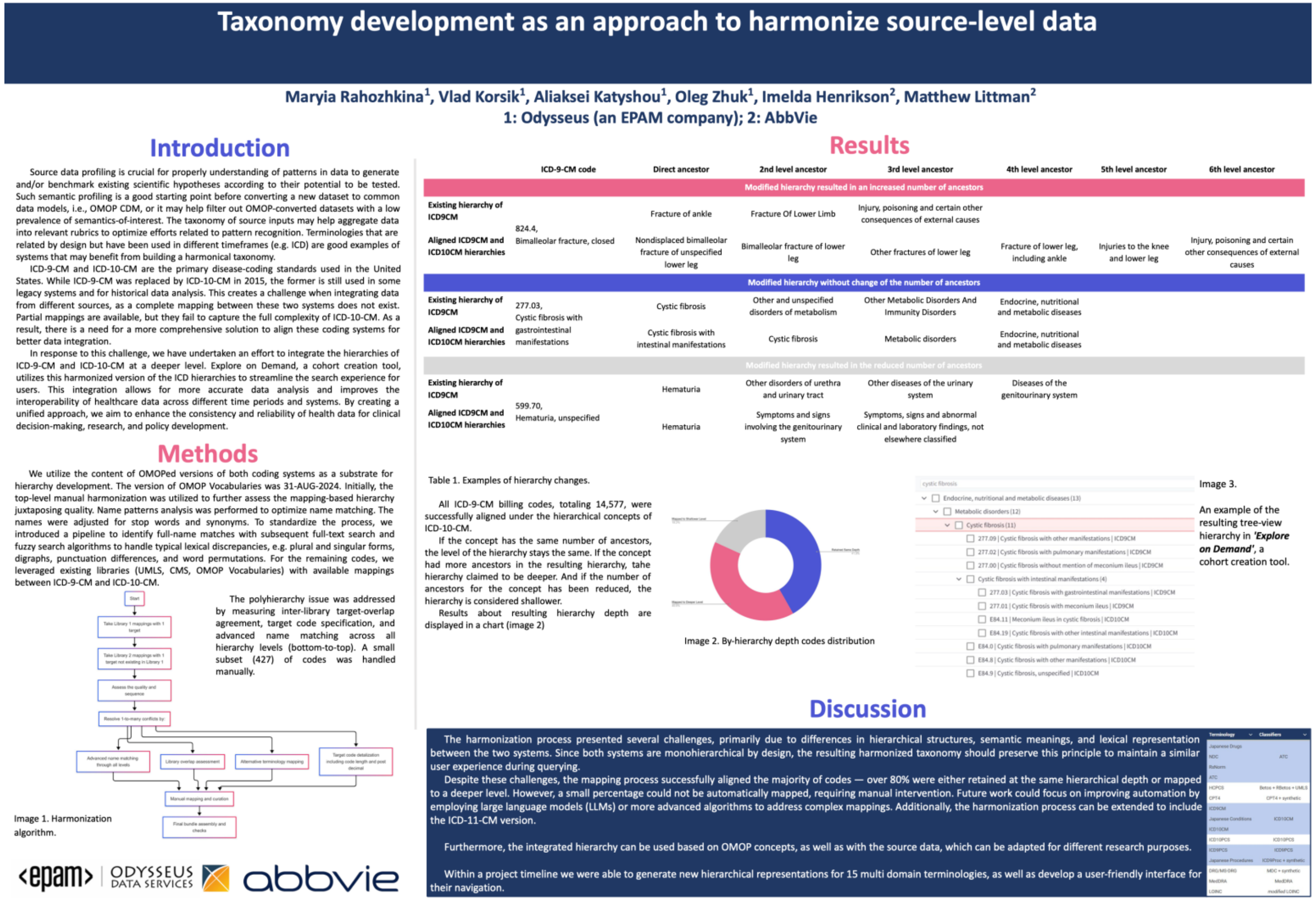
ohdsi.org/apac2025

#OHDSISocialShowcase This Week

Monday

Taxonomy development as an approach to harmonize source-level data

(Maryia Rahozhkina, Vlad Korsik, Aliaksei Katyshou, Oleg Zhuk, Imelda Henrikson, Matthew Littman)



Taxonomy development as an approach to harmonize source-level data

Maryia Rahozhkina[1], Vlad Korsik[1], Aliaksei Katyshou[1], Oleg Zhuk[1], Imelda Henrikson[2], Matthew Littman[2]
1: Odysseus (an EPAM company); 2: AbbVie

Introduction

Source data profiling is crucial for properly understanding of patterns in data to generate and/or benchmark existing scientific hypotheses according to their potential to be tested. Such semantic profiling is a good starting point before converting a new dataset to common data models, i.e., OMOP CDM, or it may help filter out OMOP-converted datasets with a low prevalence of semantics-of-interest. The taxonomy of source inputs may help aggregate data into relevant rubrics to optimize efforts related to pattern recognition. Terminologies that are related by design but have been used in different timeframes (e.g. ICD) are good examples of systems that may benefit from building a harmonical taxonomy.

ICD-9-CM and ICD-10-CM are the primary disease-coding standards used in the United States. While ICD-9-CM was replaced by ICD-10-CM in 2015, the former is still used in some legacy systems and for historical data analysis. This creates a challenge when integrating data from different sources, as a complete mapping between these two systems does not exist. Partial mappings are available, but they fail to capture the full complexity of ICD-10-CM. As a result, there is a need for a more comprehensive solution to align these coding systems for better data integration.

In response to this challenge, we have undertaken an effort to integrate the hierarchies of ICD-9-CM and ICD-10-CM at a deeper level. Explore on Demand, a cohort creation tool, utilizes this harmonized version of the ICD hierarchies to streamline the search experience for users. This integration allows for more accurate data analysis and improves the interoperability of healthcare data across different time periods and systems. By creating a unified approach, we aim to enhance the consistency and reliability of health data for clinical decision-making, research, and policy development.

Methods

We utilize the content of OMOPed versions of both coding systems as a substrate for hierarchy development. The version of OMOP Vocabularies was 31-AUG-2024. Initially, the top-level manual harmonization was utilized to further assess the mapping-based hierarchy juxtaposing quality. Name patterns analysis was performed to optimize name matching. The names were adjusted for stop words and synonyms. To standardize the process, we introduced a pipeline to identify full-name matches with subsequent full-text search and fuzzy search algorithms to handle typical lexical discrepancies, e.g. plural and singular forms, digraphs, punctuation differences, and word permutations. For the remaining codes, we leveraged existing libraries (UMLS, CMS, OMOP Vocabularies) with available mappings between ICD-9-CM and ICD-10-CM.

The polyhierarchy issue was addressed by measuring inter-library target-overlap agreement, target code specification, and advanced name matching across all hierarchy levels (bottom-to-top). A small subset (427) of codes was handled manually.

Start

Take Library 1 mappings with 1 target

Take Library 2 mappings with 1 target not existing in Library 1

Assess the quality and sequence

Resolve 1-to-many conflicts by:

Advanced name matching through all levels

Library overlap assessment

Alternative terminology mapping

Target code detalization including code length and post decimal

Manual mapping and curation

Final bundle assembly and checks

Image 1. Harmonization algorithm.

Results

	ICD-9-CM code	Direct ancestor	2nd level ancestor	3rd level ancestor	4th level ancestor	5th level ancestor	6th level ancestor
Modified hierarchy resulted in an increased number of ancestors							
Existing hierarchy of ICD9CM	824.4, Bimalleolar fracture, closed	Fracture of ankle	Fracture Of Lower Limb	Injury, poisoning and certain other consequences of external causes			
Aligned ICD9CM and ICD10CM hierarchies		Nondisplaced bimalleolar fracture of unspecified lower leg	Bimalleolar fracture of lower leg	Other fractures of lower leg	Fracture of lower leg, including ankle	Injuries to the knee and lower leg	Injury, poisoning and certain other consequences of external causes
Modified hierarchy without change of the number of ancestors							
Existing hierarchy of ICD9CM	277.03, Cystic fibrosis with gastrointestinal manifestations	Cystic fibrosis	Other and unspecified disorders of metabolism	Other Metabolic Disorders And Immunity Disorders	Endocrine, nutritional and metabolic diseases		
Aligned ICD9CM and ICD10CM hierarchies		Cystic fibrosis with intestinal manifestations	Cystic fibrosis	Metabolic disorders	Endocrine, nutritional and metabolic diseases		
Modified hierarchy resulted in the reduced number of ancestors							
Existing hierarchy of ICD9CM	599.70, Hematuria, unspecified	Hematuria	Other disorders of urethra and urinary tract	Other diseases of the urinary system	Diseases of the genitourinary system		
Aligned ICD9CM and ICD10CM hierarchies		Hematuria	Symptoms and signs involving the genitourinary system	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified			

Table 1. Examples of hierarchy changes.

All ICD-9-CM billing codes, totaling 14,577, were successfully aligned under the hierarchical concepts of ICD-10-CM.

If the concept has the same number of ancestors, the level of the hierarchy stays the same. If the concept had more ancestors in the resulting hierarchy, tahe hierarchy claimed to be deeper. And if the number of ancestors for the concept has been reduced, the hierarchy is considered shallower.

Results about resulting hierarchy depth are displayed in a chart (image 2)

Image 2. By-hierarchy depth codes distribution

cystic fibrosis

- ☐ Endocrine, nutritional and metabolic diseases (13)
  - ☐ Metabolic disorders (12)
    - ☐ Cystic fibrosis (11)
      - ☐ 277.09 | Cystic fibrosis with other manifestations | ICD9CM
      - ☐ 277.02 | Cystic fibrosis with pulmonary manifestations | ICD9CM
      - ☐ 277.00 | Cystic fibrosis without mention of meconium ileus | ICD9CM
      - ☐ Cystic fibrosis with intestinal manifestations (4)
        - ☐ 277.03 | Cystic fibrosis with gastrointestinal manifestations | ICD9CM
        - ☐ 277.01 | Cystic fibrosis with meconium ileus | ICD9CM
        - ☐ E84.11 | Meconium ileus in cystic fibrosis | ICD10CM
        - ☐ E84.19 | Cystic fibrosis with other intestinal manifestations | ICD10CM
      - ☐ E84.0 | Cystic fibrosis with pulmonary manifestations | ICD10CM
      - ☐ E84.8 | Cystic fibrosis with other manifestations | ICD10CM
      - ☐ E84.9 | Cystic fibrosis, unspecified | ICD10CM

Image 3.

An example of the resulting tree-view hierarchy in ***'Explore on Demand'***, a cohort creation tool.

Discussion

The harmonization process presented several challenges, primarily due to differences in hierarchical structures, semantic meanings, and lexical representation between the two systems. Since both systems are monohierarchical by design, the resulting harmonized taxonomy should preserve this principle to maintain a similar user experience during querying.

Despite these challenges, the mapping process successfully aligned the majority of codes — over 80% were either retained at the same hierarchical depth or mapped to a deeper level. However, a small percentage could not be automatically mapped, requiring manual intervention. Future work could focus on improving automation by employing large language models (LLMs) or more advanced algorithms to address complex mappings. Additionally, the harmonization process can be extended to include the ICD-11-CM version.

Furthermore, the integrated hierarchy can be used based on OMOP concepts, as well as with the source data, which can be adapted for different research purposes.

Within a project timeline we were able to generate new hierarchical representations for 15 multi domain terminologies, as well as develop a user-friendly interface for their navigation.

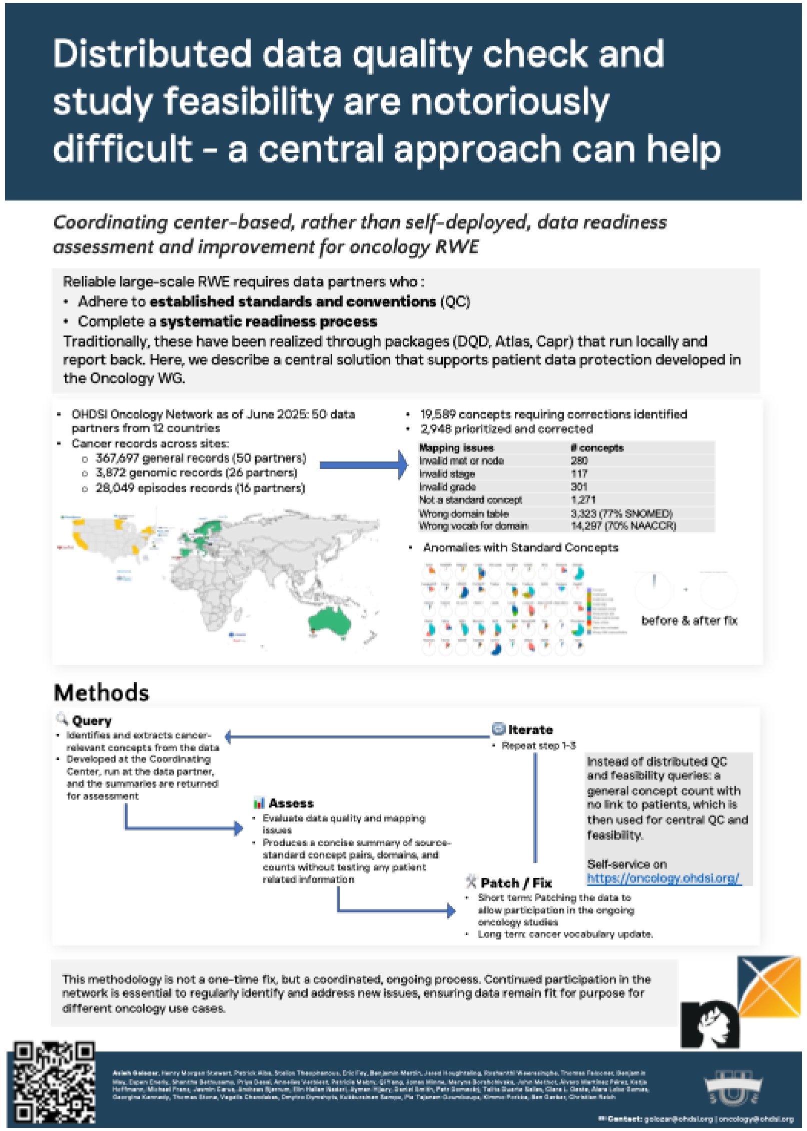
Terminology	Classifiers
Japanese Drugs	
NDC	ATC
RxNorm	
ATC	
HCPCS	Betos + RBetos + UMLS
CPT4	CPT4 + synthetic
ICD9CM	
Japanese Conditions	ICD10CM
ICD10CM	
ICD10PCS	ICD10PCS
ICD9PCS	ICD9PCS
Japanese Procedures	ICD9Proc + synthetic
DRG/MS-DRG	MDC + synthetic
MedDRA	MedDRA
LOINC	modified LOINC

#OHDSISocialShowcase This Week

Tuesday

Coordinating center-based, rather than self-deployed, data readiness assessment and improvement for oncology RWE

(Asieh Golozar, Henry Morgan Stewart, Patrick Alba, Stelios Theophanous, Eric Fey, Benjamin Martin, Jared Houghtaling, Roshanthi Weerasinghe, Thomas Falconer, Benjamin May, Espen Enerly, Shantha Bethusamy, Priya Desai, Annelies Verbiest, Patricia Mabry, Qi Yang, Jonas Minne, Maryna Borshchivska, John Methot, Alvaro Andres Alvarez Peralta, Katja Hoffmann, Michael Franz, Jasmin Carus, Andreas Bjerrum, Elin Hallan Naderi, Ayman Hijazy, Daniel Smith, Petr Domecký, Talita Duarte Salles, Clara L. Oeste Aiara Lobo Gomes, Georgina Kennedy, Thomas Stone, Vagelis Chandakas, Dmytro Dymshyts, Kukkurainen Sampo, Pia Tajanen-Doumbouya, Kimmo Porkka, Ben Gerber, Christian Reich)



Distributed data quality check and study feasibility are notoriously difficult - a central approach can help

Coordinating center-based, rather than self-deployed, data readiness assessment and improvement for oncology RWE

Reliable large-scale RWE requires data partners who :

- Adhere to **established standards and conventions** (QC)
- Complete a **systematic readiness process**

Traditionally, these have been realized through packages (DQD, Atlas, Capr) that run locally and report back. Here, we describe a central solution that supports patient data protection developed in the Oncology WG.

- OHDSI Oncology Network as of June 2025: 50 data partners from 12 countries
- Cancer records across sites:
  - 367,697 general records (50 partners)
  - 3,872 genomic records (26 partners)
  - 28,049 episodes records (16 partners)

- 19,589 concepts requiring corrections identified
- 2,948 prioritized and corrected

Mapping issues	# concepts
Invalid met or node	280
Invalid stage	117
Invalid grade	301
Not a standard concept	1,271
Wrong domain table	3,323 (77% SNOMED)
Wrong vocab for domain	14,297 (70% NAACCR)

- Anomalies with Standard Concepts

before & after fix

Methods

Query

- Identifies and extracts cancer-relevant concepts from the data
- Developed at the Coordinating Center, run at the data partner, and the summaries are returned for assessment

Assess

- Evaluate data quality and mapping issues
- Produces a concise summary of source-standard concept pairs, domains, and counts without testing any patient related information

Patch / Fix

- Short term: Patching the data to allow participation in the ongoing oncology studies
- Long term: cancer vocabulary update.

Iterate

- Repeat step 1-3

Instead of distributed QC and feasibility queries: a general concept count with no link to patients, which is then used for central QC and feasibility.

Self-service on https://oncology.ohdsi.org/

This methodology is not a one-time fix, but a coordinated, ongoing process. Continued participation in the network is essential to regularly identify and address new issues, ensuring data remain fit for purpose for different oncology use cases.

Asieh Golozar, Henry Morgan Stewart, Patrick Alba, Stelios Theophanous, Eric Fey, Benjamin Martin, Jared Houghtaling, Roshanthi Weerasinghe, Thomas Falconer, Benjamin May, Espen Enerly, Shantha Bethusamy, Priya Desai, Annelies Verbiest, Patricia Mabry, Qi Yang, Jonas Minne, Maryna Borshchivska, John Methot, Álvaro Martínez Pérez, Katja Hoffmann, Michael Franz, Jasmin Carus, Andreas Bjerrum, Elin Hallan Naderi, Ayman Hijazy, Daniel Smith, Petr Domecký, Talita Duarte Salles, Clara L. Oeste, Aiara Lobo Gomes, Georgina Kennedy, Thomas Stone, Vagelis Chandakas, Dmytro Dymshyts, Kukkurainen Sampo, Pia Tajanen-Doumbouya, Kimmo Porkka, Ben Gerber, Christian Reich

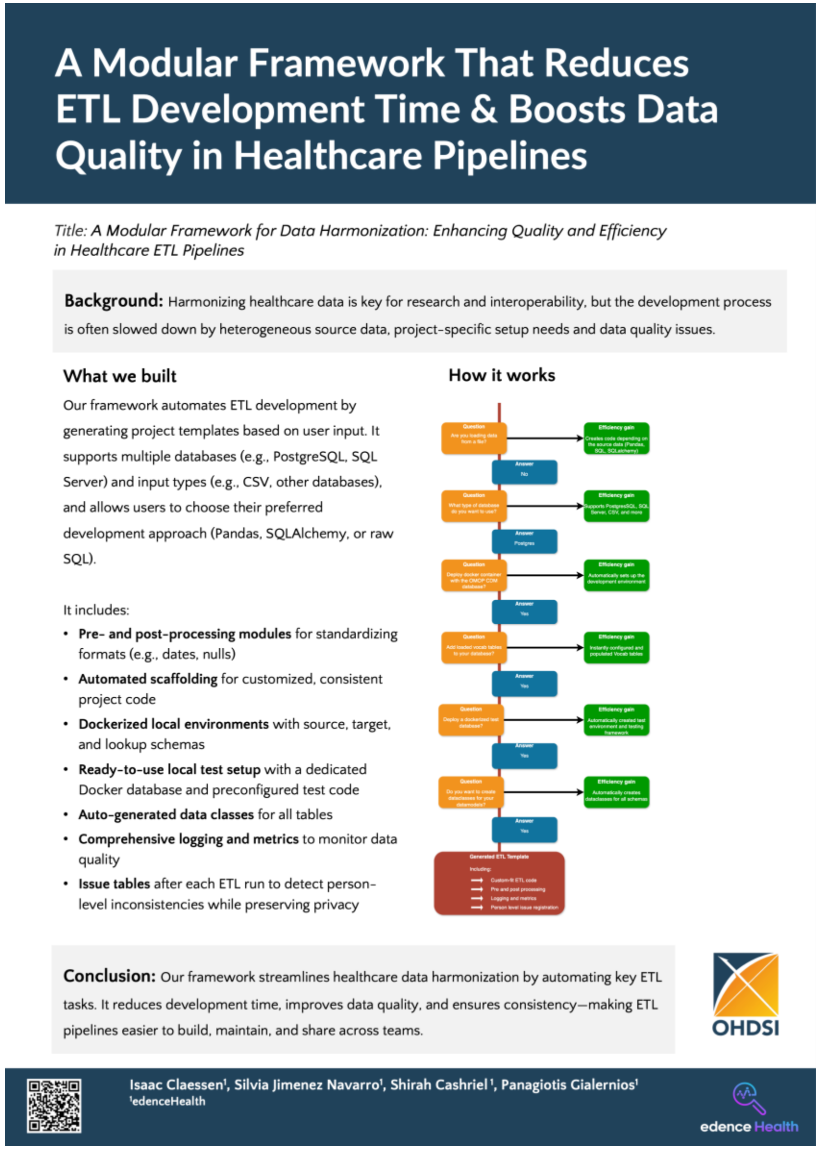
Contact: golozar@ohdsi.org | oncology@ohdsi.org

#OHDSISocialShowcase This Week

Wednesday

A Modular Framework for Data Harmonization: Enhancing Quality and Efficiency in Healthcare ETL Pipelines

(Isaac Claessen, Silvia Jimenez Navarro, Shirah Cashriel, Panagiotis Gialernios)



A Modular Framework That Reduces ETL Development Time & Boosts Data Quality in Healthcare Pipelines

Title: A Modular Framework for Data Harmonization: Enhancing Quality and Efficiency in Healthcare ETL Pipelines

Background: Harmonizing healthcare data is key for research and interoperability, but the development process is often slowed down by heterogeneous source data, project-specific setup needs and data quality issues.

What we built

Our framework automates ETL development by generating project templates based on user input. It supports multiple databases (e.g., PostgreSQL, SQL Server) and input types (e.g., CSV, other databases), and allows users to choose their preferred development approach (Pandas, SQLAlchemy, or raw SQL).

It includes:

- **Pre- and post-processing modules** for standardizing formats (e.g., dates, nulls)
- **Automated scaffolding** for customized, consistent project code
- **Dockerized local environments** with source, target, and lookup schemas
- **Ready-to-use local test setup** with a dedicated Docker database and preconfigured test code
- **Auto-generated data classes** for all tables
- **Comprehensive logging and metrics** to monitor data quality
- **Issue tables** after each ETL run to detect person-level inconsistencies while preserving privacy

How it works

Question
Are you loading data from a file?
Efficiency gain
Creates code depending on the source data (Pandas, SQL, SQLalchemy)
Answer
No
Question
What type of database do you want to use?
Efficiency gain
Supports PostgresSQL, SQL Server, CSV, and more
Answer
Postgres
Question
Deploy docker container with the OMOP CDM database?
Efficiency gain
Automatically sets up the development environment
Answer
Yes
Question
Add loaded vocab tables to your database?
Efficiency gain
Instantly configured and populated Vocab tables
Answer
Yes
Question
Deploy a dockerized test database?
Efficiency gain
Automatically created test environment and testing framework
Answer
Yes
Question
Do you want to create dataclasses for your datamodels?
Efficiency gain
Automatically creates dataclasses for all schemas
Answer
Yes
Generated ETL Template
Including:
Custom-fit ETL code
Pre and post processing
Logging and metrics
Person level issue registration

Conclusion: Our framework streamlines healthcare data harmonization by automating key ETL tasks. It reduces development time, improves data quality, and ensures consistency—making ETL pipelines easier to build, maintain, and share across teams.

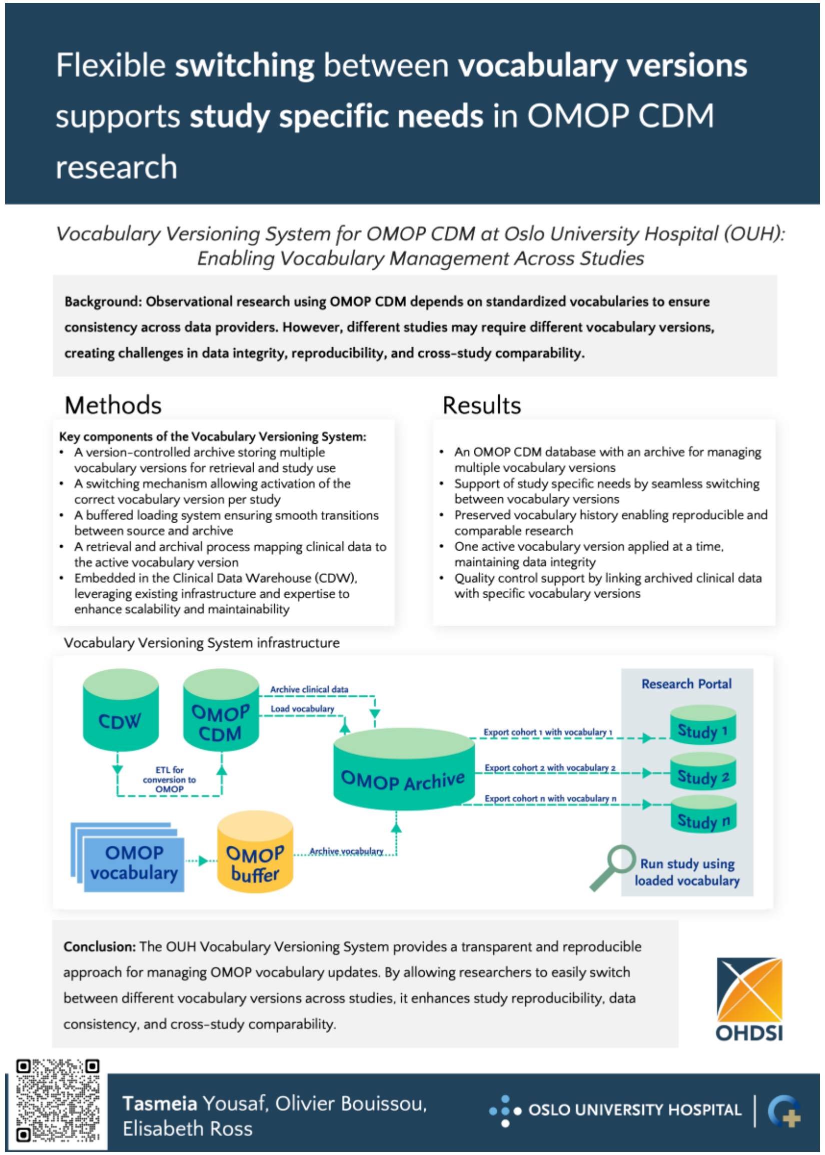
Isaac Claessen[1], Silvia Jimenez Navarro[1], Shirah Cashriel[1], Panagiotis Gialernios[1]
[1]edenceHealth

#OHDSISocialShowcase This Week

Thursday

Vocabulary Versioning System for OMOP-CDM: Enabling Vocabulary Management Across Studies

(Tasmeia Yousaf, Olivier Bouissou, Elisabeth Ross)



Flexible **switching** between **vocabulary versions** supports **study specific needs** in OMOP CDM research

Vocabulary Versioning System for OMOP CDM at Oslo University Hospital (OUH): Enabling Vocabulary Management Across Studies

Background: Observational research using OMOP CDM depends on standardized vocabularies to ensure consistency across data providers. However, different studies may require different vocabulary versions, creating challenges in data integrity, reproducibility, and cross-study comparability.

Methods

Key components of the Vocabulary Versioning System:

- A version-controlled archive storing multiple vocabulary versions for retrieval and study use
- A switching mechanism allowing activation of the correct vocabulary version per study
- A buffered loading system ensuring smooth transitions between source and archive
- A retrieval and archival process mapping clinical data to the active vocabulary version
- Embedded in the Clinical Data Warehouse (CDW), leveraging existing infrastructure and expertise to enhance scalability and maintainability

Results

- An OMOP CDM database with an archive for managing multiple vocabulary versions
- Support of study specific needs by seamless switching between vocabulary versions
- Preserved vocabulary history enabling reproducible and comparable research
- One active vocabulary version applied at a time, maintaining data integrity
- Quality control support by linking archived clinical data with specific vocabulary versions

Vocabulary Versioning System infrastructure

CDW
OMOP CDM
Archive clinical data
Load vocabulary
ETL for conversion to OMOP
OMOP Archive
Research Portal
Export cohort 1 with vocabulary 1
Export cohort 2 with vocabulary 2
Export cohort n with vocabulary n
Study 1
Study 2
Study n
OMOP vocabulary
OMOP buffer
Archive vocabulary
Run study using loaded vocabulary
OHDSI

Conclusion: The OUH Vocabulary Versioning System provides a transparent and reproducible approach for managing OMOP vocabulary updates. By allowing researchers to easily switch between different vocabulary versions across studies, it enhances study reproducibility, data consistency, and cross-study comparability.

Tasmeia Yousaf, Olivier Bouissou, Elisabeth Ross

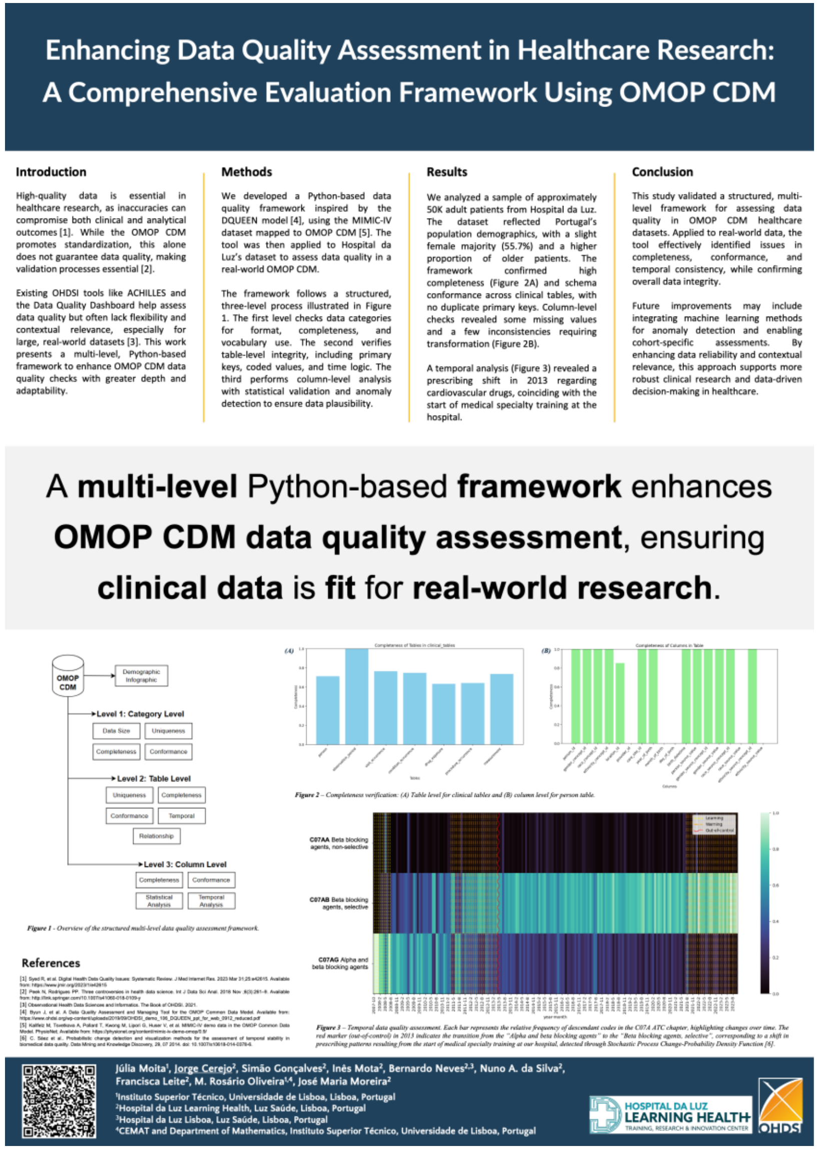
OSLO UNIVERSITY HOSPITAL

#OHDSISocialShowcase This Week

Friday

Enhancing Data Quality Assessment in Healthcare Research: A Comprehensive Evaluation Framework Using OMOP CDM

(**Júlia Moita**, Jorge Cerejo, Inês Mota, Simão Gonçalves, Bernardo Neves, Nuno André da Silva, Francisca Leite, Maria Rosário Oliveira, José Maria Moreira)



Enhancing Data Quality Assessment in Healthcare Research: A Comprehensive Evaluation Framework Using OMOP CDM

Introduction

High-quality data is essential in healthcare research, as inaccuracies can compromise both clinical and analytical outcomes [1]. While the OMOP CDM promotes standardization, this alone does not guarantee data quality, making validation processes essential [2].

Existing OHDSI tools like ACHILLES and the Data Quality Dashboard help assess data quality but often lack flexibility and contextual relevance, especially for large, real-world datasets [3]. This work presents a multi-level, Python-based framework to enhance OMOP CDM data quality checks with greater depth and adaptability.

Methods

We developed a Python-based data quality framework inspired by the DQUEEN model [4], using the MIMIC-IV dataset mapped to OMOP CDM [5]. The tool was then applied to Hospital da Luz's dataset to assess data quality in a real-world OMOP CDM.

The framework follows a structured, three-level process illustrated in Figure 1. The first level checks data categories for format, completeness, and vocabulary use. The second verifies table-level integrity, including primary keys, coded values, and time logic. The third performs column-level analysis with statistical validation and anomaly detection to ensure data plausibility.

Results

We analyzed a sample of approximately 50K adult patients from Hospital da Luz. The dataset reflected Portugal's population demographics, with a slight female majority (55.7%) and a higher proportion of older patients. The framework confirmed high completeness (Figure 2A) and schema conformance across clinical tables, with no duplicate primary keys. Column-level checks revealed some missing values and a few inconsistencies requiring transformation (Figure 2B).

A temporal analysis (Figure 3) revealed a prescribing shift in 2013 regarding cardiovascular drugs, coinciding with the start of medical specialty training at the hospital.

Conclusion

This study validated a structured, multi-level framework for assessing data quality in OMOP CDM healthcare datasets. Applied to real-world data, the tool effectively identified issues in completeness, conformance, and temporal consistency, while confirming overall data integrity.

Future improvements may include integrating machine learning methods for anomaly detection and enabling cohort-specific assessments. By enhancing data reliability and contextual relevance, this approach supports more robust clinical research and data-driven decision-making in healthcare.

A **multi-level** Python-based **framework** enhances **OMOP CDM data quality assessment**, ensuring **clinical data** is **fit** for **real-world research**.

Figure 1 - Overview of the structured multi-level data quality assessment framework.

Figure 2 – Completeness verification: (A) Table level for clinical tables and (B) column level for person table.

C07AA Beta blocking agents, non-selective
C07AB Beta blocking agents, selective
C07AG Alpha and beta blocking agents
Learning
Warning
Out-of-control
year-month

Figure 3 – Temporal data quality assessment. Each bar represents the relative frequency of descendant codes in the C07A ATC chapter, highlighting changes over time. The red marker (out-of-control) in 2013 indicates the transition from the "Alpha and beta blocking agents" to the "Beta blocking agents, selective", corresponding to a shift in prescribing patterns resulting from the start of medical specialty training at our hospital, detected through Stochastic Process Change-Probability Density Function [6].

References

[1] Syed R, et al. Digital Health Data Quality Issues: Systematic Review. J Med Internet Res. 2023 Mar 31;25:e42615. Available from: https://www.jmir.org/2023/1/e42615

[2] Peek N, Rodrigues PP. Three controversies in health data science. Int J Data Sci Anal. 2018 Nov ;6(3):261–9. Available from: http://link.springer.com/10.1007/s41060-018-0109-y

[3] Observational Health Data Sciences and Informatics. The Book of OHDSI. 2021.

[4] Byun J, et al. A Data Quality Assessment and Managing Tool for the OMOP Common Data Model. Available from: https://www.ohdsi.org/wp-content/uploads/2019/09/OHDSI_demo_106_DQUEEN_ppt_for_web_0912_reduced.pdf

[5] Kallfelz M, Tsvetkova A, Pollard T, Kwong M, Lipori G, Huser V, et al. MIMIC-IV demo data in the OMOP Common Data Model. PhysioNet; Available from: https://physionet.org/content/mimic-iv-demo-omop/0.9/

[6] C. Sáez et al.. Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. Data Mining and Knowledge Discovery, 29, 07 2014. doi: 10.1007/s10618-014-0378-6.

Júlia Moita[1], Jorge Cerejo[2], Simão Gonçalves[2], Inês Mota[2], Bernardo Neves[2,3], Nuno A. da Silva[2], Francisca Leite[2], M. Rosário Oliveira[1,4], José Maria Moreira[2]

[1]Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
[2]Hospital da Luz Learning Health, Luz Saúde, Lisboa, Portugal
[3]Hospital da Luz Lisboa, Luz Saúde, Lisboa, Portugal
[4]CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Portugal

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

OHDSI

Where Are We Going?

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

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Mad Minutes

Dmytro Dymshyts (148): Evaluating the OHDSI Phenotype library concept sets using Large Language Models
Qingrui (Carrie) Wang (115): Automated Anatomical Identification and Standardization for Medical Images
Gabriel Salvador (403): Replicating Alzheimer's Research using standardized phenotyping with the OMOP common data model imaging extension
Melanie Philofsky (141): Maximizing EHR Semantic Meaning for Rare Diseases Utilizing a Direct Mapping Strategy
Erik Benton (507): OMOP Annotator: A Database agnostic tool for reviewing and augmenting the patient record
Niko Möller-Grell (310): Agentic conversation on OMOP CDM: the OMCP-A2A foundation library
Jared Houghtaling (602): OMOP Waveform Extension: A Schema for Integrating Physiological Signals and Derived Features into the OMOP CDM
Jen Park (113): Real-World Implementation of the Medical Imaging CDM: An Alzheimer's Disease Use Case
Robert Barrett (603): Improving VSAC to OMOP Mapping Using LLM Assisted Curation
Christelle Xiong (205): AgentDose: Towards Accurate and Scalable Steroid Dose Extraction in OMOP Using NLP Parsers and LLM Agents

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