Workgroup Spotlight: Vocabulary and Evidence Network

OHDSI Community Call
May 26, 2026 • 11 am ET

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Upcoming Community Calls

Date	Topic
May 26	Workgroup Spotlight: Vocabulary and Evidence Network
June 2	LLM Research Around The World, Session 1
June 9	LLM Research Around The World, Session 2
June 16	LLM Research Around The World, Session 3
June 23	CANCELLED: OHDSI Summer School at Columbia University
June 30	OMOP & OHDSI Research Spotlight

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Natthawut Adulyanukosol, Krittaphas Chaisutyakorn, Saknarong Sombutjaroan, Suchanan Kanjanapong and Prapat Suriyaphol** on the recent publication of **Efficient Drug Terminology Mapping with Bidirectional Late-Interaction Reranking and Deterministic Reordering** in *Healthcare Informatics Research*.

Original Article

Healthc Inform Res. 2026 April;32(2):156-165.
https://doi.org/10.4258/hir.2026.32.2.156
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Efficient Drug Terminology Mapping with Bidirectional Late-Interaction Reranking and Deterministic Reordering

Natthawut Adulyanukosol, Krittaphas Chaisutyakorn, Saknarong Sombutjaroan, Suchanan Kanjanapong, Prapat Suriyaphol

Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand

Objectives: Standardizing medication concepts across heterogeneous vocabularies is essential for interoperable analytics and observational research. In the Observational Medical Outcomes Partnership (OMOP) Common Data Model, local drug codes must be mapped to standardized RxNorm concepts, but automated mapping is challenging because drug strings encode clinically critical attributes, including strength, dosage form/route, release characteristics, and brand. **Methods:** We propose THIRAWAT (Terminology Harmonization using Late-Interaction Reranker With Alignment-tuned Transformers), a fine-tuned ColBERTv1 late-interaction reranker, and embed it within THIRAWAT Mapper, a retrieval–reranking pipeline with deterministic tie-breaking and stable ordering. Candidate generation used approximate nearest-neighbor retrieval with a bi-encoder (SapBERT-XLMR or BioLORD-2023). Candidates were reranked by THIRAWAT models that were fine-tuned using one-sided MaxSim and scored at inference using our adapted Bidirectional MaxSim (BiMaxSim) pooling. Finally, a deterministic tie-breaker extracted clinically salient cues, including strength, dosage form/route, release characteristics, and bracketed brand annotations, to resolve near-ties reproducibly. **Results:** We evaluated three mapping settings: Branded Drugs, Clinical Drugs, and Thai Medicines Terminology (TMT). Using SapBERT-XLMR retrieval with THIRAWAT-SapBERT reranking and deterministic tie-breaking, THIRAWAT Mapper achieved MRR@100 values of 0.954 (95% confidence interval [CI], 0.921–0.983), 0.898 (95% CI, 0.866–0.925), and 0.912 (95% CI, 0.891–0.931), outperforming a lexical term frequency–inverse document frequency baseline (0.491, 0.216, and 0.143, respectively). Hits@1 improved to 0.942 (95% CI, 0.899–0.978), 0.859 (95% CI, 0.817–0.898), and 0.868 (95% CI, 0.838–0.896), respectively. **Conclusions:** BiMaxSim and deterministic tie-breaking improved drug mapping to RxNorm while preserving an efficient runtime profile and stable ordering. Overall, THIRAWAT Mapper offers a pragmatic combination of learned semantic matching and deterministic lexical constraints. Models and code are available on Hugging Face (https://huggingface.co/collections/sidataplus/thirawat) and GitHub (https://github.com/sidataplus/THIRAWAT-mapper).

OHDSI Shoutouts!

Congratulations to the team of **Ricardo Lourenço Santos and Ricardo João Cruz-Correia** on the recent publication of **An HL7 FHIR® IG for lifestyle medicine in learning health systems: Multi-vendor wearable interoperability with documented terminology gaps** in the *International Journal of Medical Informatics*.

International Journal of Medical Informatics 217 (2026) 106465

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Original research article

An HL7 FHIR® IG for lifestyle medicine in learning health systems: Multi-vendor wearable interoperability with documented terminology gaps

Ricardo Lourenço Santos [a, b, *], Ricardo João Cruz-Correia [a, b]

[a] Faculty of Medicine, University of Porto, Porto, Portugal
[b] RISE-Health Research Network, Universidade do Porto, Porto, Portugal

HIGHLIGHTS

- IPS-compatible, EHDS-ready HL7 FHIR® IG for patient-generated wearable data.
- Multi-vendor semantic convergence buffer spanning 7 wearable vendors and 11 domains across lifestyle medicine and wearable biometrics.
- Documents critical wearable terminology gaps: RMSSD and pNN50 lack LOINC codes; all SNOMED CT HRV concepts are fetal/obstetric.
- 718 custom codes with 28 ConceptMaps providing declarative migration paths to LOINC, SNOMED CT, and OMOP via OCL governance.
- Iterative terminology audit: 1173 codes refined to 718, with 19 LOINC substitutions across recurrent audit cycles.

ARTICLE INFO

Keywords:
FHIR®
Implementation guide
Wearable devices
Patient-generated health data
Lifestyle medicine
Heart rate variability
Interoperability
Terminology gaps
LOINC
EHDS

ABSTRACT

Background: Consumer wearable devices collect clinically useful physiological data from patients, yet less than 5% of this patient-generated health data (PGHD) reaches clinical systems. The European Health Data Space (EHDS) Regulation 2025/327 mandates standardized health data exchange, creating an urgent need for validated transformation frameworks.

Objective: To develop and validate a FHIR® R4 Implementation Guide enabling multi-vendor wearable data transformation for lifestyle medicine, while documenting terminology gaps across seven vendor ecosystems.

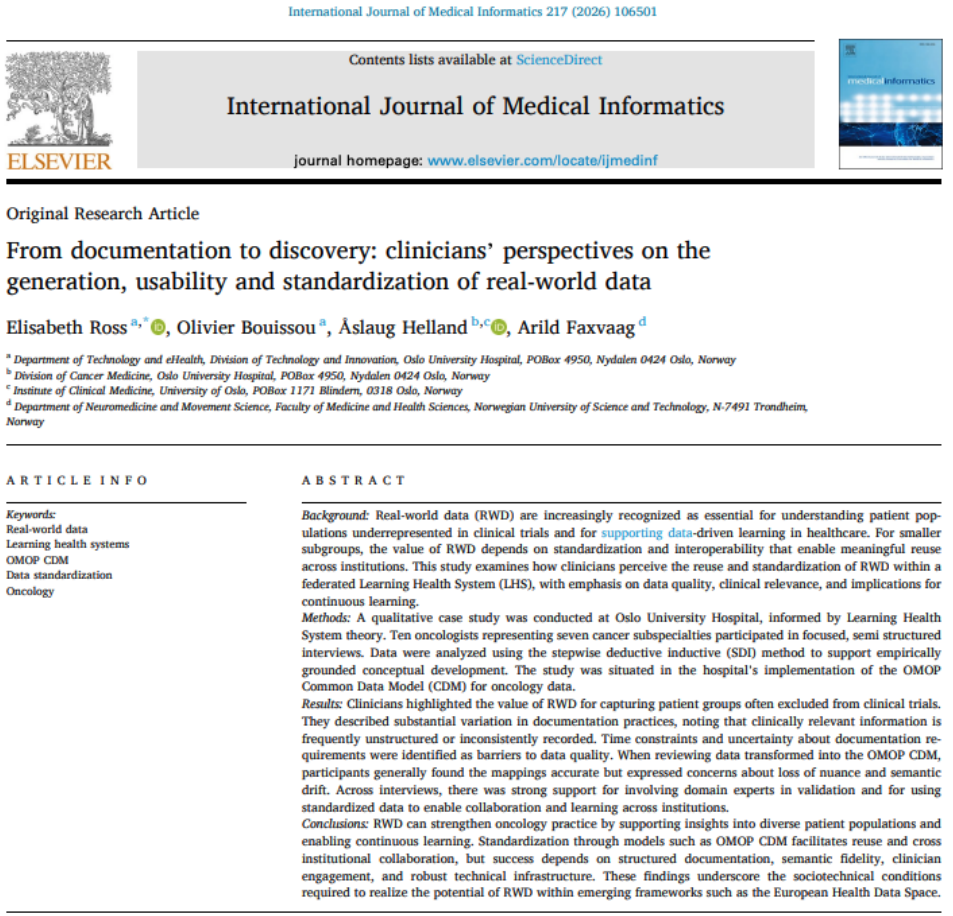
Methods: We developed FHIR® profiles, extensions, and ConceptMaps using a multi-layer architecture covering eleven domains (six lifestyle medicine pillars and five wearable-specific domains). Technical validation included HL7 IG Publisher compilation, FHIR® Validator conformance testing, and systematic terminology verification.

Results: The IG comprises 74 profiles, 50 extensions, 14 CodeSystems, 174 ValueSets, and 28 ConceptMaps (546 artifacts). A two-phase terminology audit refined 1173 custom codes to 718 without standard direct equivalents, including genuine gaps (RMSSD, pNN50), wearable-specific metrics, and granularity not yet addressed by LOINC or SNOMED CT. Build validation produces 23 errors (all IPS upstream) and 73 warnings (97% non-actionable). Of six core HRV metrics, only SDNN has a LOINC code (80404-7). The multi-layer architecture achieved 75% weighted reuse across seven vendor ecosystems.

Conclusions: This work extends the HL7 Physical Activity IG's "Temporary Codes" pattern into a semantic convergence buffer—a vendor-neutral intermediary CodeSystem with outbound ConceptMap coverage to LOINC, SNOMED CT, and OMOP enabling multi-vendor semantic convergence. The 75% weighted reuse and iterative terminology audits progressively reduce custom codes as standards mature. All profiles are vendor-agnostic and EHDS-compliant.

OHDSI Shoutouts!

Congratulations to the team of **Elisabeth Ross, Olivier Bouissou, Åslaug Helland and Arild Faxvaag** on the recent publication of **From documentation to discovery: clinicians' perspectives on the generation, usability and standardization of real-world data** in the *International Journal of Medical Informatics*.



International Journal of Medical Informatics 217 (2026) 106501

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Original Research Article

From documentation to discovery: clinicians' perspectives on the generation, usability and standardization of real-world data

Elisabeth Ross [a,*], Olivier Bouissou [a], Åslaug Helland [b,c], Arild Faxvaag [d]

[a] Department of Technology and eHealth, Division of Technology and Innovation, Oslo University Hospital, POBox 4950, Nydalen 0424 Oslo, Norway
[b] Division of Cancer Medicine, Oslo University Hospital, POBox 4950, Nydalen 0424 Oslo, Norway
[c] Institute of Clinical Medicine, University of Oslo, POBox 1171 Blindern, 0318 Oslo, Norway
[d] Department of Neuromedicine and Movement Science, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

ARTICLE INFO

Keywords:
Real-world data
Learning health systems
OMOP CDM
Data standardization
Oncology

ABSTRACT

Background: Real-world data (RWD) are increasingly recognized as essential for understanding patient populations underrepresented in clinical trials and for supporting data-driven learning in healthcare. For smaller subgroups, the value of RWD depends on standardization and interoperability that enable meaningful reuse across institutions. This study examines how clinicians perceive the reuse and standardization of RWD within a federated Learning Health System (LHS), with emphasis on data quality, clinical relevance, and implications for continuous learning.
Methods: A qualitative case study was conducted at Oslo University Hospital, informed by Learning Health System theory. Ten oncologists representing seven cancer subspecialties participated in focused, semi structured interviews. Data were analyzed using the stepwise deductive inductive (SDI) method to support empirically grounded conceptual development. The study was situated in the hospital's implementation of the OMOP Common Data Model (CDM) for oncology data.
Results: Clinicians highlighted the value of RWD for capturing patient groups often excluded from clinical trials. They described substantial variation in documentation practices, noting that clinically relevant information is frequently unstructured or inconsistently recorded. Time constraints and uncertainty about documentation requirements were identified as barriers to data quality. When reviewing data transformed into the OMOP CDM, participants generally found the mappings accurate but expressed concerns about loss of nuance and semantic drift. Across interviews, there was strong support for involving domain experts in validation and for using standardized data to enable collaboration and learning across institutions.
Conclusions: RWD can strengthen oncology practice by supporting insights into diverse patient populations and enabling continuous learning. Standardization through models such as OMOP CDM facilitates reuse and cross institutional collaboration, but success depends on structured documentation, semantic fidelity, clinician engagement, and robust technical infrastructure. These findings underscore the sociotechnical conditions required to realize the potential of RWD within emerging frameworks such as the European Health Data Space.

OHDSI Shoutouts!

Congratulations to the team of **Adnan Jouned, Laura Verbei, Florian Katsch, Marta Ferri Peradalta, Sofia Bazakou, Tanja Stamm, Georg Duftschmid and Renske Los** on the recent publication of **Integration of Patient-Reported Outcomes into the OMOP Common Data Model** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

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

doi:10.3233/SHTI260580

Integration of Patient-Reported Outcomes into the OMOP Common Data Model

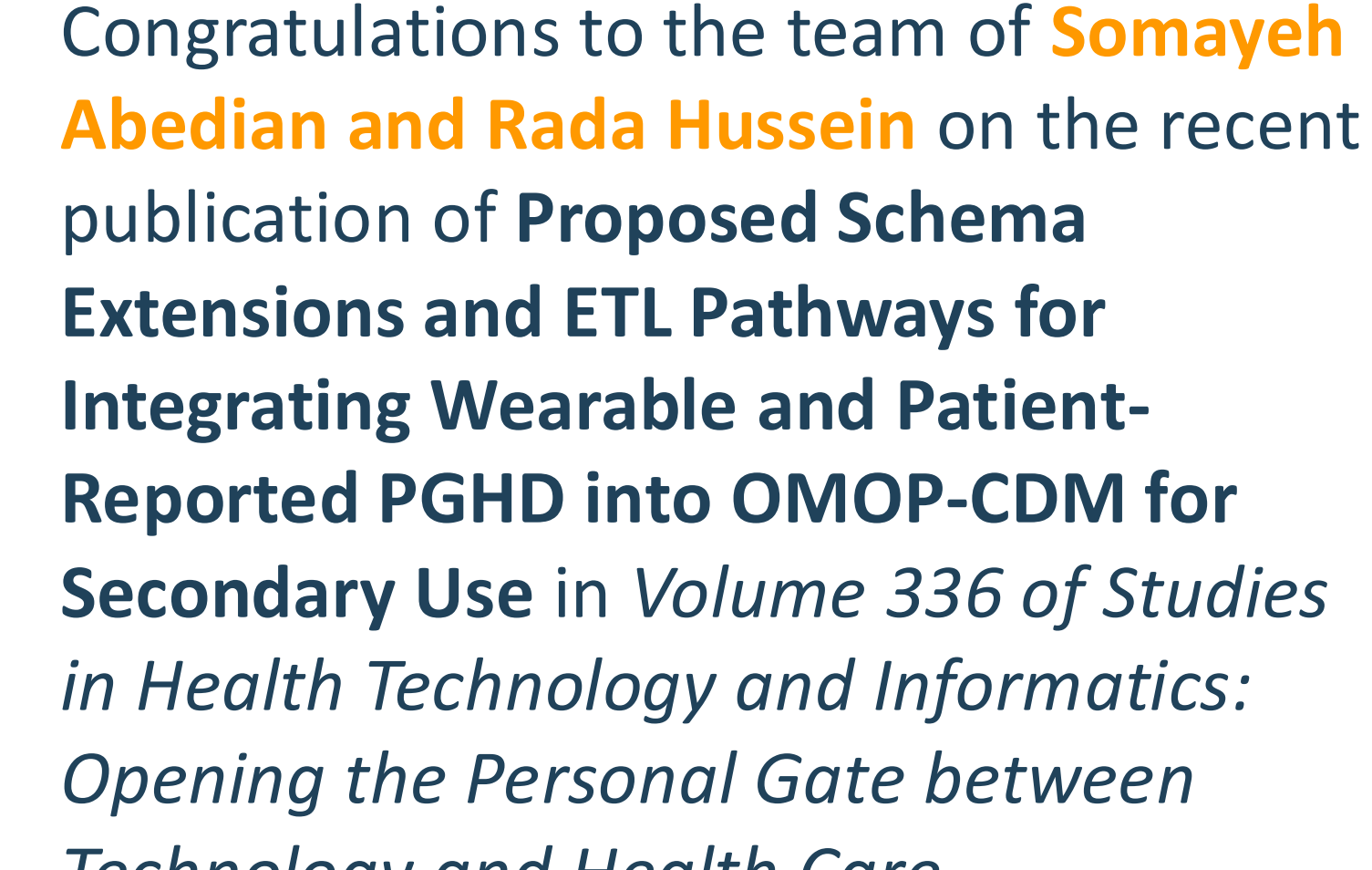
Adnan JOUNED[a,1], Laura VERBEIJ[b], Florian KATSCH[a], Marta FERRI PERADALTA[d], Sofia BAZAKOU[e], Tanja STAMM[f], Georg DUFTSCHMID[a], Renske LOS[e]
[a]*Institute of Medical Information Management, Center for Medical Data Science, Medical University of Vienna, Vienna, Austria*
[b]*Department of Quality and Patient Care, Erasmus Medical Center, University Medical Center Rotterdam, Rotterdam, The Netherlands*
[d]*Digital Transformation Area, Vall d'Hebron Institute of Research, Vall d'Hebron Barcelona Hospital Campus, Barcelona, Spain*
[e]*Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands*
[f]*Institute of Outcomes Research, Center for Medical Data Science, Medical University of Vienna & Ludwig Boltzmann Institute for Arthritis and Rehabilitation, Vienna, Austria*

Abstract. An approach for integrating Patient-Reported Outcomes into the OMOP Common Data Model is presented. Feasibility is demonstrated, although coverage remains constrained by vocabulary gaps. Information loss can be minimized by the consideration of outcome origin, negation, and temporal annotation.

Keywords. Patient-Reported Outcomes, OMOP Common Data Model, CDM

OHDSI Shoutouts!

Congratulations to the team of **Somayeh Abedian and Rada Hussein** on the recent publication of **Proposed Schema Extensions and ETL Pathways for Integrating Wearable and Patient-Reported PGHD into OMOP-CDM for Secondary Use** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.



Opening the Personal Gate between Technology and Health Care 1885
M. Gtacomini et al. (Eds.)

dot:10.3233/SHTI260565

Proposed Schema Extensions and ETL Pathways for Integrating Wearable and Patient-Reported PGHD into OMOP-CDM for Secondary Use

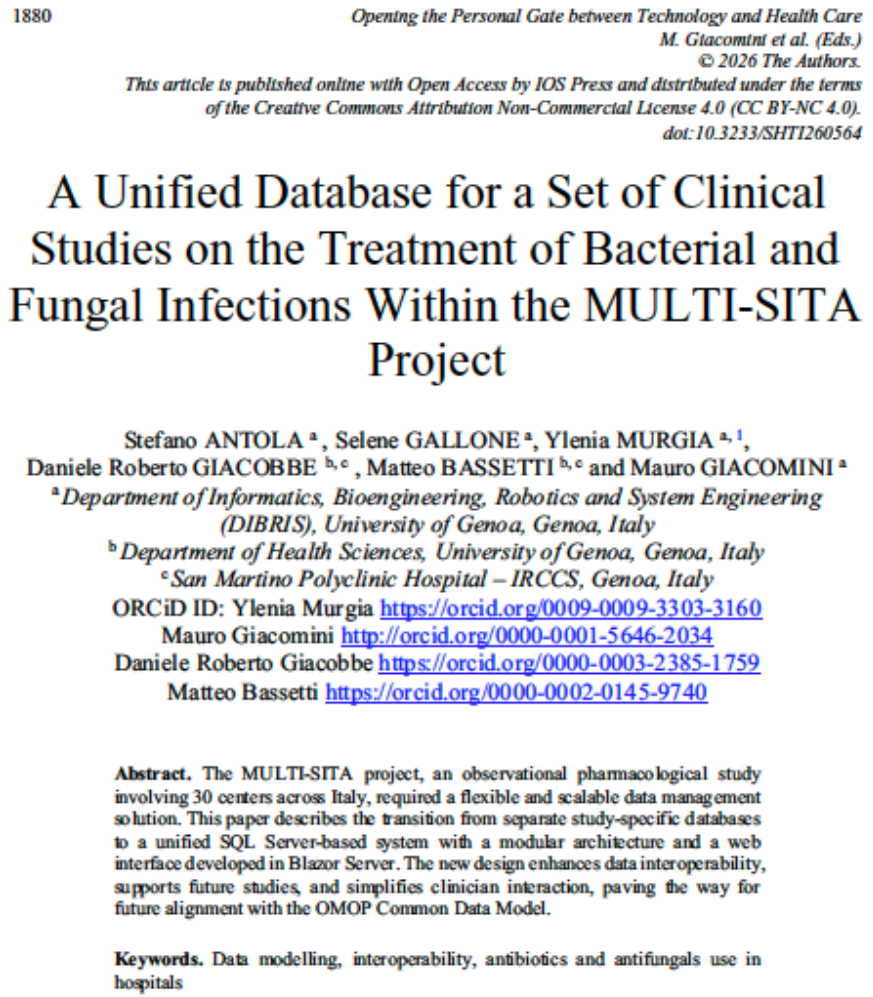
Somayeh ABEDIAN [a,b,1], and Rada HUSSEIN [a]
[a] *Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*
[b] *Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Canada*
ORCiD ID: Somayeh Abedian https://orcid.org/0000-0002-6427-6941,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

Abstract. The rapid expansion of wearable sensing technologies and patient-reported outcomes (PROs) has revealed a persistent gap between high-frequency, context-rich patient-generated health data (PGHD) and standardized data models used for secondary analysis. Although the OMOP Common Data Model (OMOP-CDM) supports large-scale observational research, it does not natively capture key PGHD characteristics such as device provenance, calibration parameters, temporal resolution, and contextual metadata. This paper introduces an extension layer for OMOP-CDM and a direct extract–transform–load (ETL) pathway that enhances its capacity to handle heterogeneous, time-series, and PRO data while maintaining semantic consistency. The approach preserves device traceability, enables flexible data compression, and links PRO instruments with OMOP constructs without altering the core schema, remaining fully compatible with OHDSI analytical tools and aligned with FAIR principles. The model has strong potential for community discussion and future OMOP updates on PGHD.

Keywords. Patient-Generated Health Data (PGHD), Wearables, Interoperability, OMOP Common Data Model, FAIR Principles

OHDSI Shoutouts!

Congratulations to the team of **Stefano Antola, Selene Gallone, Ylenia Murgia, Daniele Roberto Giacobbe, Matteo Bassetti and Mauro Giacomini** on the recent publication of **A Unified Database for a Set of Clinical Studies on the Treatment of Bacterial and Fungal Infections Within the MULTI-SITA Project** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.



doi:10.3233/SHTI260564

A Unified Database for a Set of Clinical Studies on the Treatment of Bacterial and Fungal Infections Within the MULTI-SITA Project

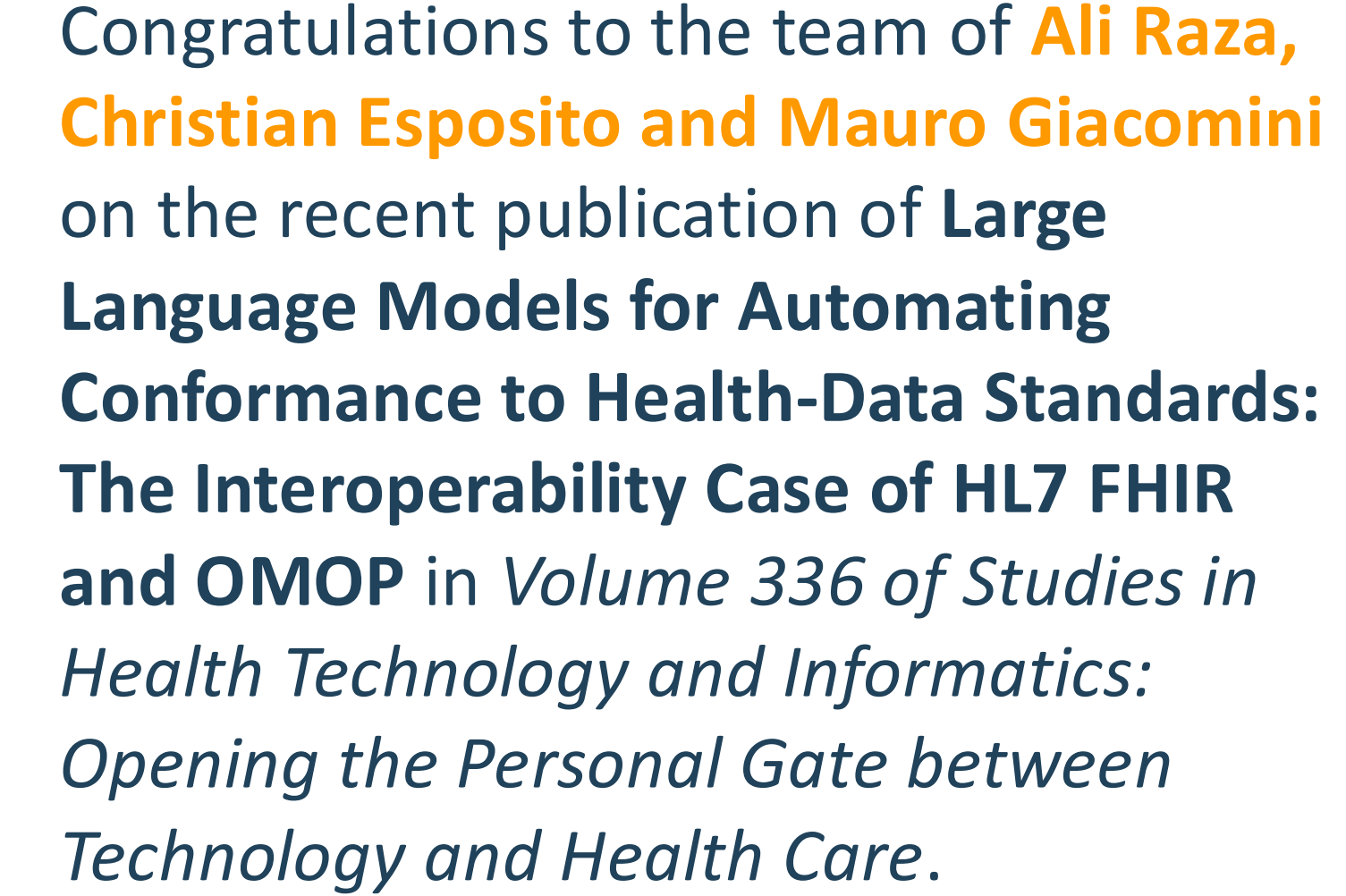
Stefano ANTOLA [a], Selene GALLONE [a], Ylenia MURGIA [a,1], Daniele Roberto GIACOBBE [b,c], Matteo BASSETTI [b,c] and Mauro GIACOMINI [a]
[a] Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Genoa, Italy
[b] Department of Health Sciences, University of Genoa, Genoa, Italy
[c] San Martino Polyclinic Hospital – IRCCS, Genoa, Italy
ORCiD ID: Ylenia Murgia https://orcid.org/0009-0009-3303-3160
Mauro Giacomini http://orcid.org/0000-0001-5646-2034
Daniele Roberto Giacobbe https://orcid.org/0000-0003-2385-1759
Matteo Bassetti https://orcid.org/0000-0002-0145-9740

Abstract. The MULTI-SITA project, an observational pharmacological study involving 30 centers across Italy, required a flexible and scalable data management solution. This paper describes the transition from separate study-specific databases to a unified SQL Server-based system with a modular architecture and a web interface developed in Blazor Server. The new design enhances data interoperability, supports future studies, and simplifies clinician interaction, paving the way for future alignment with the OMOP Common Data Model.

Keywords. Data modelling, interoperability, antibiotics and antifungals use in hospitals

OHDSI Shoutouts!

Congratulations to the team of **Ali Raza, Christian Esposito and Mauro Giacomini** on the recent publication of **Large Language Models for Automating Conformance to Health-Data Standards: The Interoperability Case of HL7 FHIR and OMOP** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.



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

doi:10.3233/SHTI260562

Large Language Models for Automating Conformance to Health-Data Standards: The Interoperability Case of HL7 FHIR and OMOP

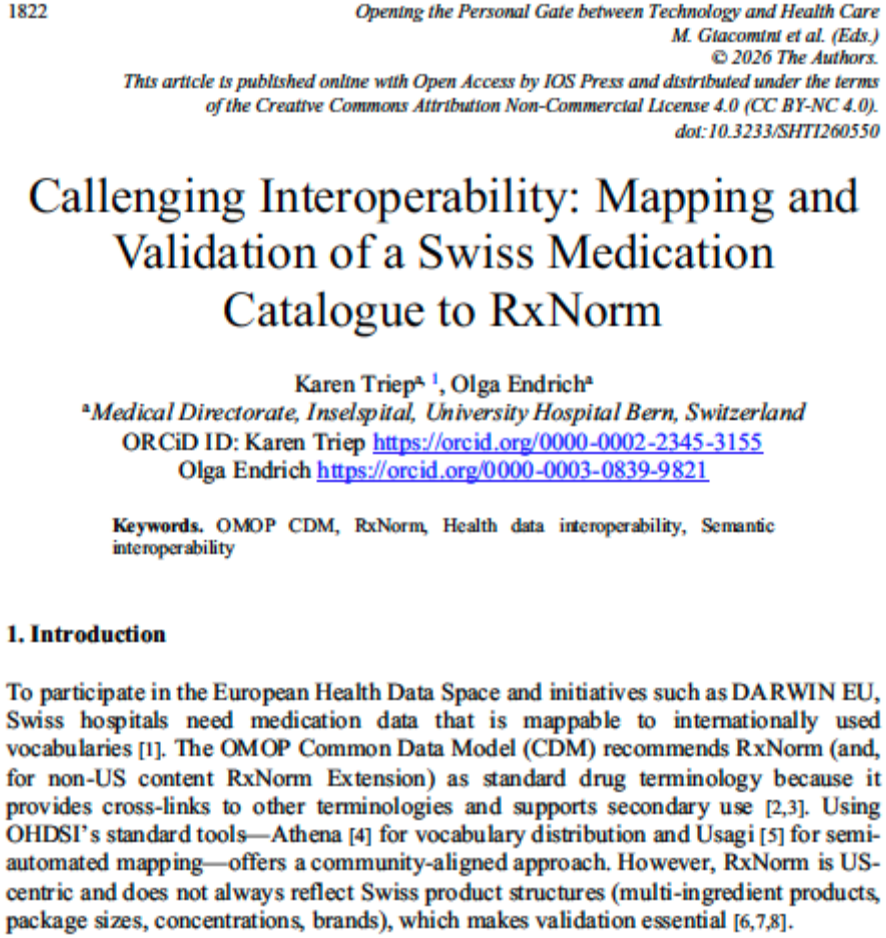
Ali RAZA[a], Christian ESPOSITO[a] and Mauro GIACOMINI[b]
[a] *University of Salerno, Italy*
[b] *University of Genoa, Italy*
ORCiD ID: Christian ESPOSITO https://orcid.org/0000-0002-0085-0748
Mauro GIACOMINI https://orcid.org/0000-0001-5646-2034

Abstract. Interoperability standards enable different healthcare systems to exchange electronic health information seamlessly, and support large-scale analytics needed for current Secondary uses of healthcare data. Transforming diverse clinical inputs into standards-compliant, analyzable data remains labor-intensive and error-prone. Moreover, the learning curve for their proper exploitation can be steep due to extensive documentation, specific technical requirements, and the need for both semantic and technical integration, making them not fully accessible to practitioners in the healthcare domain unless they undergo extensive training. Our study investigated whether large language models can automate conformance while preserving syntactic validity, semantic correctness, and reproducibility, thereby simplifying the adoption of these standards. Using diabetes measurements as a probe domain, we validated extraction fidelity, coding accuracy against controlled vocabularies, and conformance to the domain's main standards. Our prototype emits FHIR Observations and OMOP MEASUREMENT rows.

Keywords. HL7; OMOP; large language models; data standardization; LOINC; SNOMED CT; diabetes; OGTT.

OHDSI Shoutouts!

Congratulations to the team of **Karen Triep and Olga Endrich** on the recent publication of **Callenging Interoperability: Mapping and Validation of a Swiss Medication Catalogue to RxNorm** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.



doi:10.3233/SHTI260550

Callenging Interoperability: Mapping and Validation of a Swiss Medication Catalogue to RxNorm

Karen Triep[a, 1], Olga Endrich[a]
[a]Medical Directorate, Inselspital, University Hospital Bern, Switzerland
ORCiD ID: Karen Triep https://orcid.org/0000-0002-2345-3155
Olga Endrich https://orcid.org/0000-0003-0839-9821

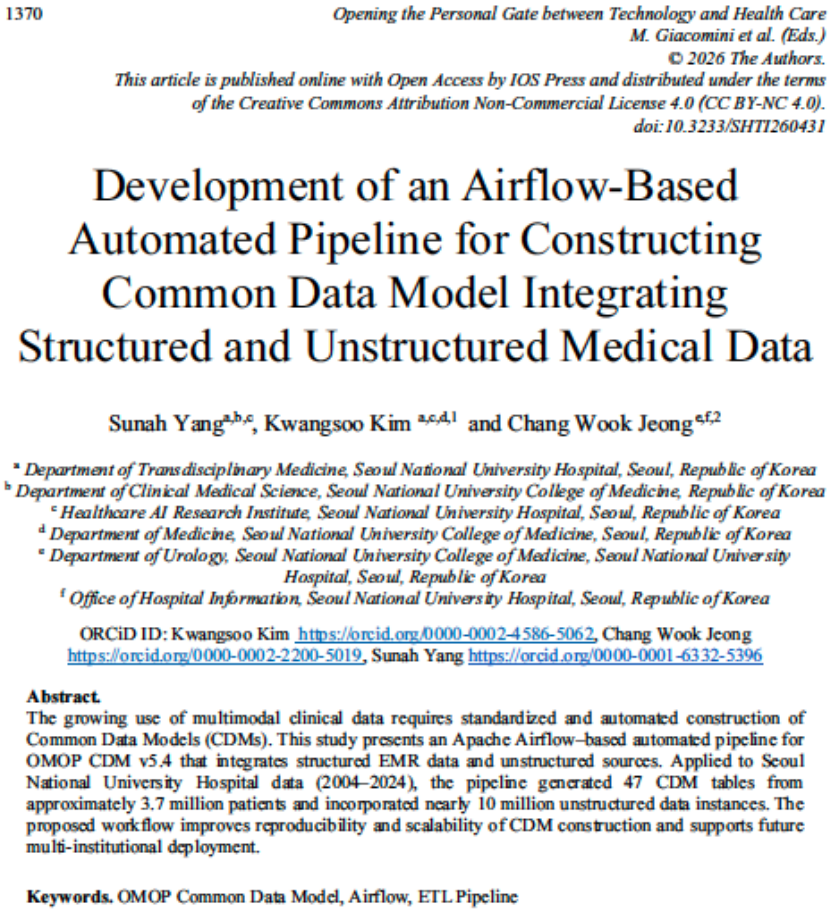
Keywords. OMOP CDM, RxNorm, Health data interoperability, Semantic interoperability

1. Introduction

To participate in the European Health Data Space and initiatives such as DARWIN EU, Swiss hospitals need medication data that is mappable to internationally used vocabularies [1]. The OMOP Common Data Model (CDM) recommends RxNorm (and, for non-US content RxNorm Extension) as standard drug terminology because it provides cross-links to other terminologies and supports secondary use [2,3]. Using OHDSI's standard tools—Athena [4] for vocabulary distribution and Usagi [5] for semi-automated mapping—offers a community-aligned approach. However, RxNorm is US-centric and does not always reflect Swiss product structures (multi-ingredient products, package sizes, concentrations, brands), which makes validation essential [6,7,8].

OHDSI Shoutouts!

Congratulations to the team of **Sunah Yang, Kwangsoo Kim and Chang Wook Jeong** on the recent publication of **Development of an Airflow-Based Automated Pipeline for Constructing Common Data Model Integrating Structured and Unstructured Medical Data** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.



doi:10.3233/SHTI260431

Development of an Airflow-Based Automated Pipeline for Constructing Common Data Model Integrating Structured and Unstructured Medical Data

Sunah Yang[a,b,c], Kwangsoo Kim [a,c,d,1] and Chang Wook Jeong [e,f,2]

[a] Department of Transdisciplinary Medicine, Seoul National University Hospital, Seoul, Republic of Korea
[b] Department of Clinical Medical Science, Seoul National University College of Medicine, Republic of Korea
[c] Healthcare AI Research Institute, Seoul National University Hospital, Seoul, Republic of Korea
[d] Department of Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea
[e] Department of Urology, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Republic of Korea
[f] Office of Hospital Information, Seoul National University Hospital, Seoul, Republic of Korea

ORCiD ID: Kwangsoo Kim https://orcid.org/0000-0002-4586-5062, Chang Wook Jeong https://orcid.org/0000-0002-2200-5019, Sunah Yang https://orcid.org/0000-0001-6332-5396

Abstract.
The growing use of multimodal clinical data requires standardized and automated construction of Common Data Models (CDMs). This study presents an Apache Airflow–based automated pipeline for OMOP CDM v5.4 that integrates structured EMR data and unstructured sources. Applied to Seoul National University Hospital data (2004–2024), the pipeline generated 47 CDM tables from approximately 3.7 million patients and incorporated nearly 10 million unstructured data instances. The proposed workflow improves reproducibility and scalability of CDM construction and supports future multi-institutional deployment.

Keywords. OMOP Common Data Model, Airflow, ETL Pipeline

OHDSI Shoutouts!

Congratulations to the team of **Maria Parra Rodriguez-Armijo, Celia Alvarez-Romero, Jan van den Brand, Boris Delange and Carlos Luis Parra-Calderón** on the recent publication of **Design of a Privacy-Preserving ETL Dataflow for Federated ICU Data Reuse in INDICATE** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

doi:10.3233/SHTI260424

Design of a Privacy-Preserving ETL Dataflow for Federated ICU Data Reuse in INDICATE

Maria PARRA RODRIGUEZ-ARMIJO[a,1], Celia ALVAREZ-ROMERO[a], Jan VAN DEN BRAND[b], Boris DELANGE[c], Carlos Luis PARRA-CALDERÓN[a]
[a]*Computational Health Informatics Group. Institute of Biomedicine of Seville, IBIS/Virgen del Rocio University Hospital/CSIC/University of Seville*
[b]*Erasmus MC University Medical Center, Department of Adult Intensive Care, Rotterdam, The Netherlands*
[c]*Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, Rennes, France*

ORCiD ID: Maria Parra Rodriguez-Armijo https://orcid.org/0009-0001-5356-3777, Celia Alvarez-Romero https://orcid.org/0000-0001-8647-9515, Carlos Luis Parra-Calderón https://orcid.org/0000-0003-2609-575X, Jan VAN DEN BRAND https://orcid.org/0000-0002-1685-9886, Boris DELANGE https://orcid.org/0009-0002-6055-6935

Abstract. The INDICATE project, funded by the European Union under the Digital Europe Programme, develops a federated framework for the secure reuse of intensive care unit (ICU) data within the EHDS. This work presents the design of a privacy-preserving ETL dataflow that standardizes and validates ICU data locally. The process enables harmonization of ICU data into the OMOP Common Data Model, based on HL7 FHIR as well, ensuring semantic interoperability, data quality, and regulatory compliance. Only aggregated study results are uploaded to the INDICATE Portal, ensuring that no patient-level data leave the hospital domain.

Keywords. ETL dataflow, ICU data, federated analytics, OMOP CDM, EHDS.

OHDSI Shoutouts!

Congratulations to the team of **Philip Stampfer, Hendrik Lef, Sai Pavan Kumar Veeranki, Birgit Fürst, Antonia Schelnast, Martina Kroissenbrunner, Elisabeth Mayrhuber, Elske Ammenwerth, Katharina Lichtenegger and Franz Feichtner** on the recent publication of **Cross-Institutional Data Harmonization for AI in Nursing Care Using the OMOP CDM** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.

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

doi:10.3233/SHTI260419

Cross-Institutional Data Harmonization for AI in Nursing Care Using the OMOP CDM

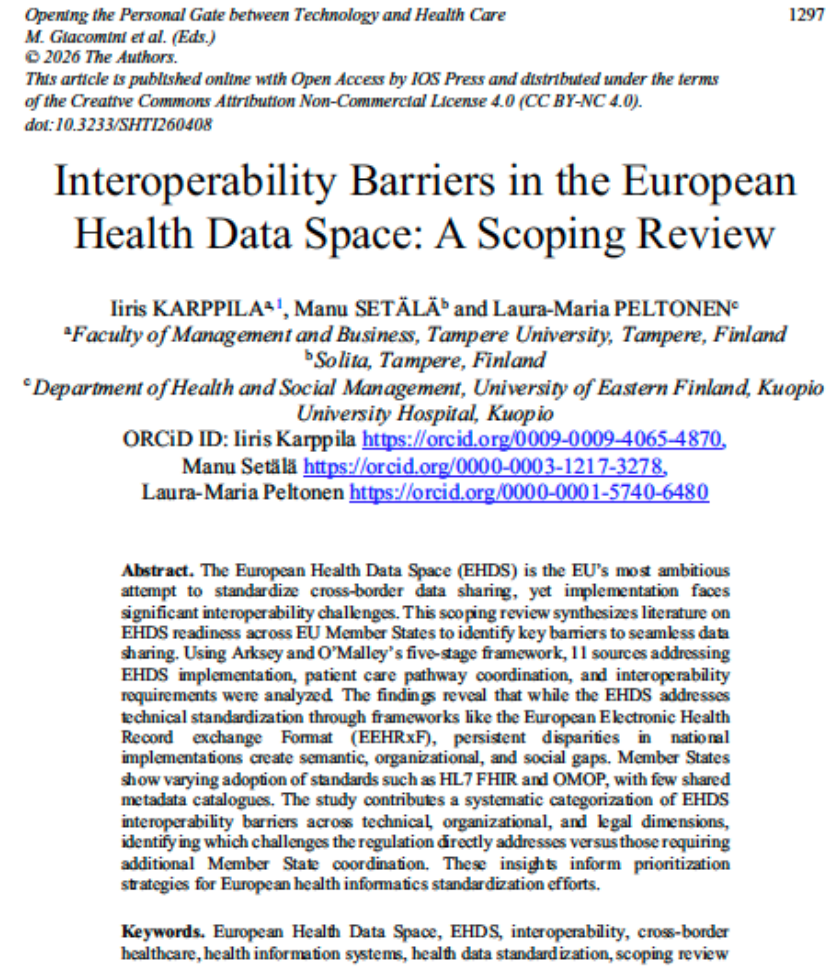
Philip STAMPFER[a,1], Hendrik LEF[b], Sai Pavan Kumar VEERANKI[c], Birgit FÜRST[c], Antonia SCHELNAST[c], Martina KROISSENBRUNNER[d], Elisabeth MAYRHUBER[e], Elske AMMENWERTH[b], Katharina LICHTENEGGER[f] and Franz FEICHTNER[a]
[a]*Joanneum Research Forschungsgesellschaft mbH, Graz, Austria*
[b]*UMIT TIROL, Institute of Medical Informatics, Hall in Tirol, Austria*
[c]*Steiermärkische Krankenanstaltengesellschaft mbH, Graz, Austria*
[d]*Althea Pflegebetriebs GmbH, Dobl, Austria*
[e]*University of Applied Sciences Upper Austria, Hagenberg, Austria*
[f]*Clinical Dep. of Endocrinology and Diabetology, Medical University of Graz, Austria*
ORCiD ID: Philip Stampfer https://orcid.org/0000-0002-0019-2140

Abstract. Artificial Intelligence (AI) offers potential to support and empower nurses, yet its development depends on the availability of high-quality, standardized data. Nursing data are often fragmented, unstructured, and semantically inconsistent, hindering their secondary use. This work aims to harmonize heterogeneous nursing data from hospitals and nursing homes to create an AI-ready data foundation. Following a generic harmonization process, example datasets from two institutions and software systems were extracted and mapped to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM). Using open-source tools, we performed dataset specification, vocabulary identification, coverage analysis, semantic and structural mapping, and initial ETL implementation. A core dataset was defined, covering key data elements such as demographics, vital signs, and medication. Initial test transfers demonstrated the feasibility of mapping and integrating nursing data, though complex nursing constructs such as care plans and assessments remain challenging. This study establishes a structured methodological approach for cross-institutional nursing data harmonization and lays the groundwork for the development of future AI applications in nursing.

Keywords. Nursing, Data Harmonization, Artificial Intelligence, Real World Data

OHDSI Shoutouts!

Congratulations to the team of **Iiris Karppila, Manu Setälä and Laura-Maria Peltonen** on the recent publication of **Interoperability Barriers in the European Health Data Space: A Scoping Review** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.



Opening the Personal Gate between Technology and Health Care 1297
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260408

Interoperability Barriers in the European Health Data Space: A Scoping Review

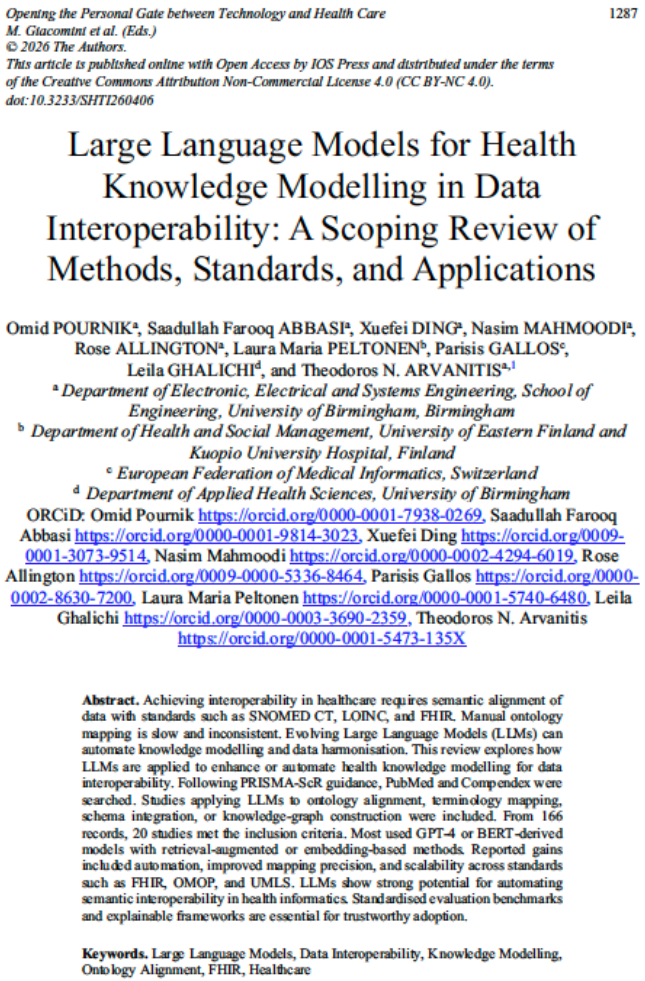
Iiris KARPPILA[a,1], Manu SETÄLÄ[b] and Laura-Maria PELTONEN[c]
[a]*Faculty of Management and Business, Tampere University, Tampere, Finland*
[b]*Solita, Tampere, Finland*
[c]*Department of Health and Social Management, University of Eastern Finland, Kuopio University Hospital, Kuopio*
ORCiD ID: Iiris Karppila https://orcid.org/0009-0009-4065-4870,
Manu Setälä https://orcid.org/0000-0003-1217-3278,
Laura-Maria Peltonen https://orcid.org/0000-0001-5740-6480

Abstract. The European Health Data Space (EHDS) is the EU's most ambitious attempt to standardize cross-border data sharing, yet implementation faces significant interoperability challenges. This scoping review synthesizes literature on EHDS readiness across EU Member States to identify key barriers to seamless data sharing. Using Arksey and O'Malley's five-stage framework, 11 sources addressing EHDS implementation, patient care pathway coordination, and interoperability requirements were analyzed. The findings reveal that while the EHDS addresses technical standardization through frameworks like the European Electronic Health Record exchange Format (EEHRxF), persistent disparities in national implementations create semantic, organizational, and social gaps. Member States show varying adoption of standards such as HL7 FHIR and OMOP, with few shared metadata catalogues. The study contributes a systematic categorization of EHDS interoperability barriers across technical, organizational, and legal dimensions, identifying which challenges the regulation directly addresses versus those requiring additional Member State coordination. These insights inform prioritization strategies for European health informatics standardization efforts.

Keywords. European Health Data Space, EHDS, interoperability, cross-border healthcare, health information systems, health data standardization, scoping review

OHDSI Shoutouts!

Congratulations to the team of **Omid Pournik, Saadullah Farooq Abbasi, Xuefei Ding, Nasim Mahmoodi, Rose Allington, Laura Maria Peltonen, Parisis Gallos, Leila Ghalichi and Theodoros N. Arvanitis** on the recent publication of **Large Language Models for Health Knowledge Modelling in Data Interoperability: A Scoping Review of Methods, Standards, and Applications** in *Volume 336 of Studies in Health Technology and Informatics: Opening the Personal Gate between Technology and Health Care*.



Opening the Personal Gate between Technology and Health Care 1287
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260406

Large Language Models for Health Knowledge Modelling in Data Interoperability: A Scoping Review of Methods, Standards, and Applications

Omid POURNIK[a], Saadullah Farooq ABBASI[a], Xuefei DING[a], Nasim MAHMOODI[a], Rose ALLINGTON[a], Laura Maria PELTONEN[b], Parisis GALLOS[c], Leila GHALICHI[d], and Theodoros N. ARVANITIS[a,1]
[a] *Department of Electronic, Electrical and Systems Engineering, School of Engineering, University of Birmingham, Birmingham*
[b] *Department of Health and Social Management, University of Eastern Finland and Kuopio University Hospital, Finland*
[c] *European Federation of Medical Informatics, Switzerland*
[d] *Department of Applied Health Sciences, University of Birmingham*
ORCiD: Omid Pournik https://orcid.org/0000-0001-7938-0269, Saadullah Farooq Abbasi https://orcid.org/0000-0001-9814-3023, Xuefei Ding https://orcid.org/0009-0001-3073-9514, Nasim Mahmoodi https://orcid.org/0000-0002-4294-6019, Rose Allington https://orcid.org/0009-0000-5336-8464, Parisis Gallos https://orcid.org/0000-0002-8630-7200, Laura Maria Peltonen https://orcid.org/0000-0001-5740-6480, Leila Ghalichi https://orcid.org/0000-0003-3690-2359, Theodoros N. Arvanitis https://orcid.org/0000-0001-5473-135X

Abstract. Achieving interoperability in healthcare requires semantic alignment of data with standards such as SNOMED CT, LOINC, and FHIR. Manual ontology mapping is slow and inconsistent. Evolving Large Language Models (LLMs) can automate knowledge modelling and data harmonisation. This review explores how LLMs are applied to enhance or automate health knowledge modelling for data interoperability. Following PRISMA-ScR guidance, PubMed and Compendex were searched. Studies applying LLMs to ontology alignment, terminology mapping, schema integration, or knowledge-graph construction were included. From 166 records, 20 studies met the inclusion criteria. Most used GPT-4 or BERT-derived models with retrieval-augmented or embedding-based methods. Reported gains included automation, improved mapping precision, and scalability across standards such as FHIR, OMOP, and UMLS. LLMs show strong potential for automating semantic interoperability in health informatics. Standardised evaluation benchmarks and explainable frameworks are essential for trustworthy adoption.

Keywords. Large Language Models, Data Interoperability, Knowledge Modelling, Ontology Alignment, FHIR, Healthcare

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

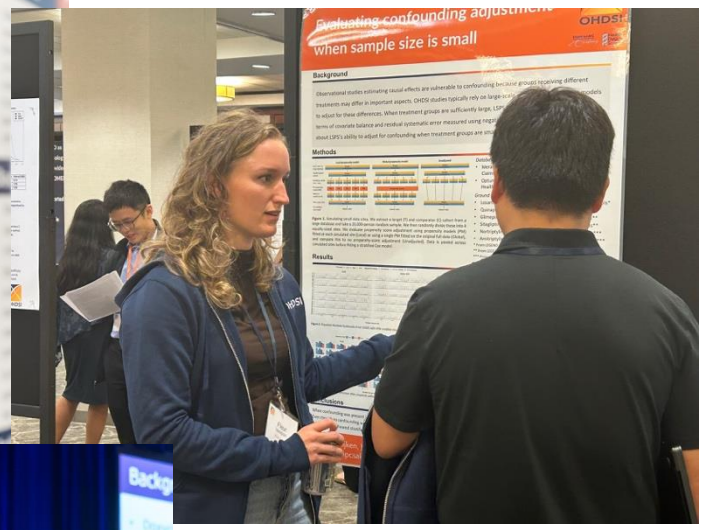
Upcoming Workgroup Calls

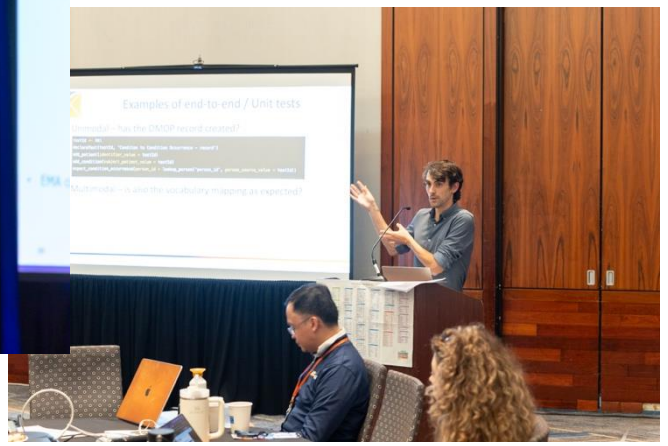
Date	Time (ET)	Meeting
Wednesday	8 am	Psychiatry
Wednesday	10 am	Surgery and Perioperative Medicine
Wednesday	10 am	Women of OHDSI
Wednesday	10 am	Common Data Model
Wednesday	1 pm	Latin America Chapter
Wednesday	7 pm	Medical Imaging
Thursday	6:30 am	India Community Call
Thursday	9 am	Phenotype Development and Evaluation
Thursday	10 am	GIS – Geographic Information System
Thursday	10:30 am	Evidence Network
Thursday	11 am	Perinatal & Reproductive Health
Thursday	7 pm	Dentistry
Friday	9 am	Waveform
Friday	10 am	Transplant
Friday	11:30 am	Steering Group
Monday	10 am	Healthcare Systems Interest Group
Tuesday	9 am	Data2Evidence

2026 OHDSI Global Symposium

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

The submission deadline is June 5 at 8 pm ET.





10 Days Remaining!

ohdsi.org/OHDSI2026

2026 OHDSI Global Symposium

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

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

ohdsi.org/OHDSI2026

Opening: Data Science Engineer

Position Summary

The Department of Biomedical Informatics at Columbia University is seeking a highly motivated data science engineer to support large-scale observational research within the OHDSI (Observational Health Data Sciences and Informatics) network. This role will focus on the design, implementation, and execution of distributed network studies using electronic health record (EHR) and administrative claims data to generate real-world evidence.

The successful candidate will contribute to characterization, population-level estimation (causal inference), and patient-level prediction analyses across multi-institutional data networks. This position offers a unique opportunity to work at the intersection of biomedical informatics, data science, and clinical research within a leading academic medical center.

This position is a full-time two-year position with a possibility of an extension, contingent on available funding.

Responsibilities

Key Responsibilities

- Design and implement observational network studies using distributed EHR and administrative claims data
- Conduct large-scale characterization, comparative effectiveness and safety estimation, and patient-level prediction analyses
- Develop reproducible analytic pipelines using R and SQL in relational database environments
- Apply and evaluate methods from causal inference (e.g., confounding control, bias assessment, sensitivity analyses)
- Apply machine learning approaches for predictive modeling using high-dimensional healthcare data
- Work with standardized data representations, including the OMOP Common Data Model and standardized clinical vocabularies for conditions, drugs, procedures, and measurements
- Collaborate with interdisciplinary teams including clinicians, statisticians, data engineers, and informaticians
- Contribute to scholarly outputs including manuscripts, presentations, and open-source analytic tools
- Support transparent, reproducible, and scalable research practices across distributed data networks

CAREERS AT COLUMBIA UNIVERSITY

Search All Jobs | Featured Career Areas | Temporary Opportunities | My Application | My Profile

Data Science Engineer

557250
Columbia University Medical Center
Biomedical Informatics
Full Time
Opening on: May 14 2026
Grade 106

Add to favorites | View favorites

- Job Type: Officer of Administration
- Bargaining Unit:
- Regular/Temporary: Regular
- End Date if Temporary:
- Hours Per Week: 35
- Standard Work Schedule: Monday - Friday
- Building: PH-20
- Salary Range: $160,000 - $180,000

The salary of the finalist selected for this role will be set based on a variety of factors, including but not limited to departmental budgets, qualifications, experience, education, licenses, specialty, and training. The above hiring range represents the University's good faith and reasonable estimate of the range of possible compensation at the time of posting.

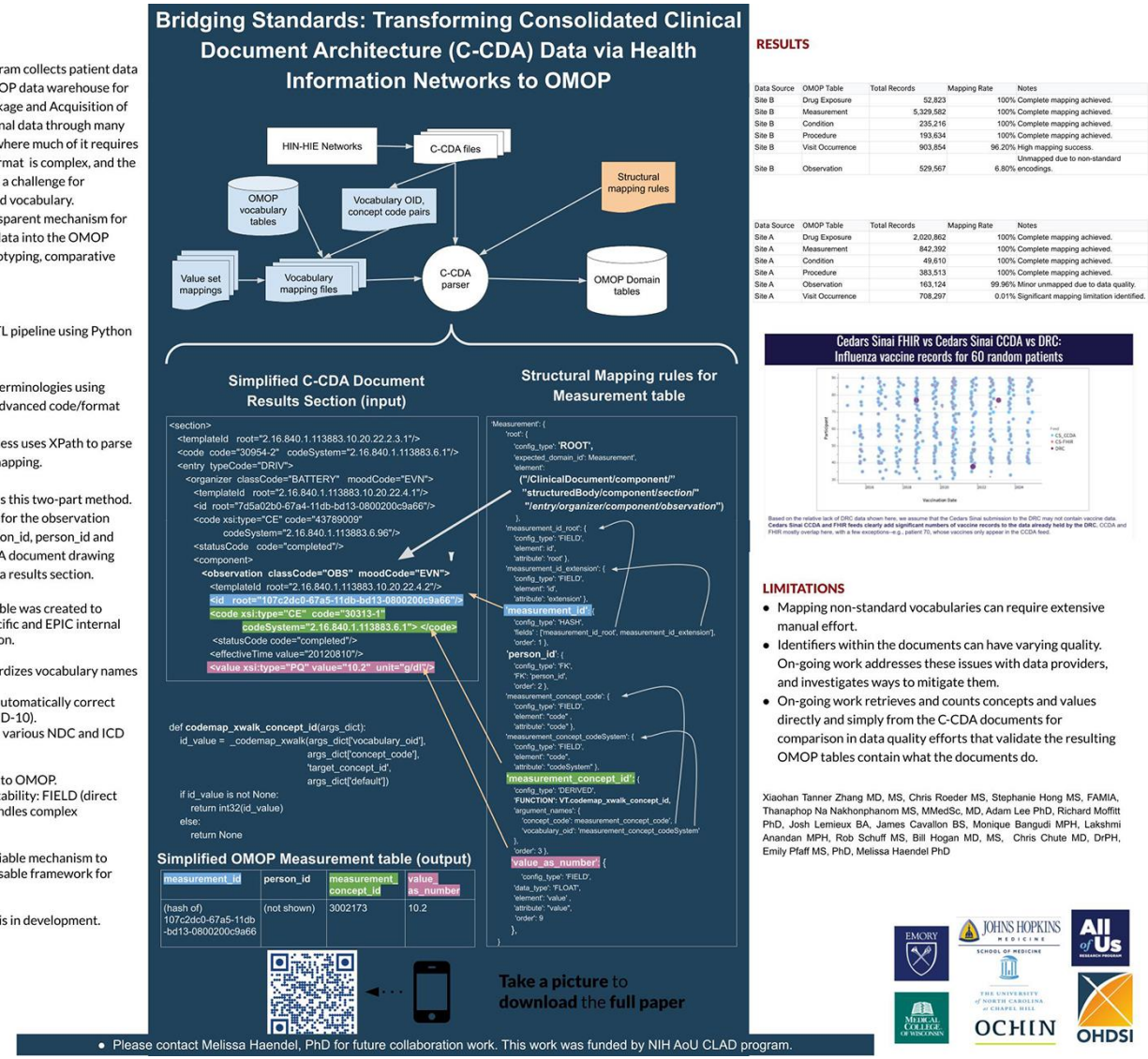
Maternal Fellowship Deadline: May 31

#OHDSISocialShowcase This Week

Monday

Bridging Standards: Transforming Consolidated Clinical Document Architecture (C-CDA) Data via Health Information Networks to OMOP

(Xiaohan Tanner Zhang, Chris Roeder, Stephanie Hong, Thanaphop Na Nakhonphanom, Adam Lee, Richard Moffitt, Josh Lemieux, James Cavallon, Monique Bangudi, Lakshmi Anandan, Rob Schuff, Bill Hogan, Chris Chute, Emily Pfaff, Melissa Haendel)

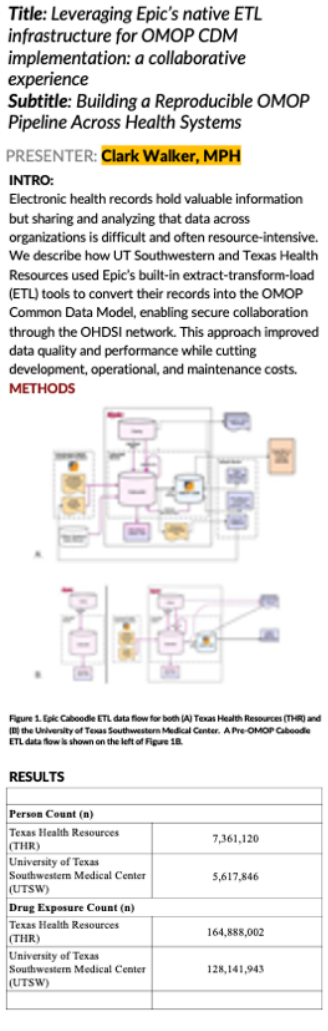


#OHDSISocialShowcase This Week

Tuesday

Leveraging Epic'S Native ETL Infrastructure for OMOP CDM Implementation: A Collaborative Experience

(Lauren N. Cooper, Aamirah Vadsariya, Mereeja Varghese, Bhavini Nayee, Jessica Moon, Chaitanya Katterapalli, Clark Walker, Chris Gonzalez, Sonam Sohal, Christoph U. Lehmann, Ferdinand Velasco, Mujeeb Basit, DuWayne Willett)



Title: *Leveraging Epic's native ETL infrastructure for OMOP CDM implementation: a collaborative experience*
Subtitle: *Building a Reproducible OMOP Pipeline Across Health Systems*

PRESENTER: **Clark Walker, MPH**

INTRO:
Electronic health records hold valuable information but sharing and analyzing that data across organizations is difficult and often resource-intensive. We describe how UT Southwestern and Texas Health Resources used Epic's built-in extract-transform-load (ETL) tools to convert their records into the OMOP Common Data Model, enabling secure collaboration through the OHDSI network. This approach improved data quality and performance while cutting development, operational, and maintenance costs.

METHODS

Figure 1. Epic Caboodle ETL data flow for both (A) Texas Health Resources (THR) and (B) the University of Texas Southwestern Medical Center. A Pre-OMOP Caboodle ETL data flow is shown on the left of Figure 1B.

RESULTS

Person Count (n)	
Texas Health Resources (THR)	7,361,120
University of Texas Southwestern Medical Center (UTSW)	5,617,846
Drug Exposure Count (n)	
Texas Health Resources (THR)	164,888,002
University of Texas Southwestern Medical Center (UTSW)	128,141,943

Epic Caboodle OHDSI

How we used Epic's built-in data pipeline to create an OMOP database together

Take a picture to download the full paper

OmopConditionOccurrenceFactX	
Duration	
Texas Health Resources (THR)	3h 1m
University of Texas Southwestern Medical Center (UTSW)	3h 7m
Rows extraction (n)	
Texas Health Resources (THR)	246,266,249
University of Texas Southwestern Medical Center (UTSW)	269,116,345
Records created (n)	
Texas Health Resources (THR)	766,066
University of Texas Southwestern Medical Center (UTSW)	20,699
Records modified (n)	
Texas Health Resources (THR)	206,356
University of Texas Southwestern Medical Center (UTSW)	10,449

Table 2. Performance metrics for the ETL of OMOP CDM table, OmopConditionOccurrenceFactX for both Texas Health Resources executed on March 16, 2025, and the University of Texas Southwestern Medical Center, executed on March 24, 2025.

Aamirah Vadsariya, RN MSN, Mereeja Varghese, MS Bhavini Nayee, BS, Jessica Moon, BS, Chaitanya Katterapalli, MS, Clark Walker, MPH, Chris Gonzalez, LPN, Sonam Sohal, MHSM MBA, Christoph U. Lehmann, MD, Ferdinand Velasco, MD, Mujeeb Basit, MD, DuWayne Willett, MD

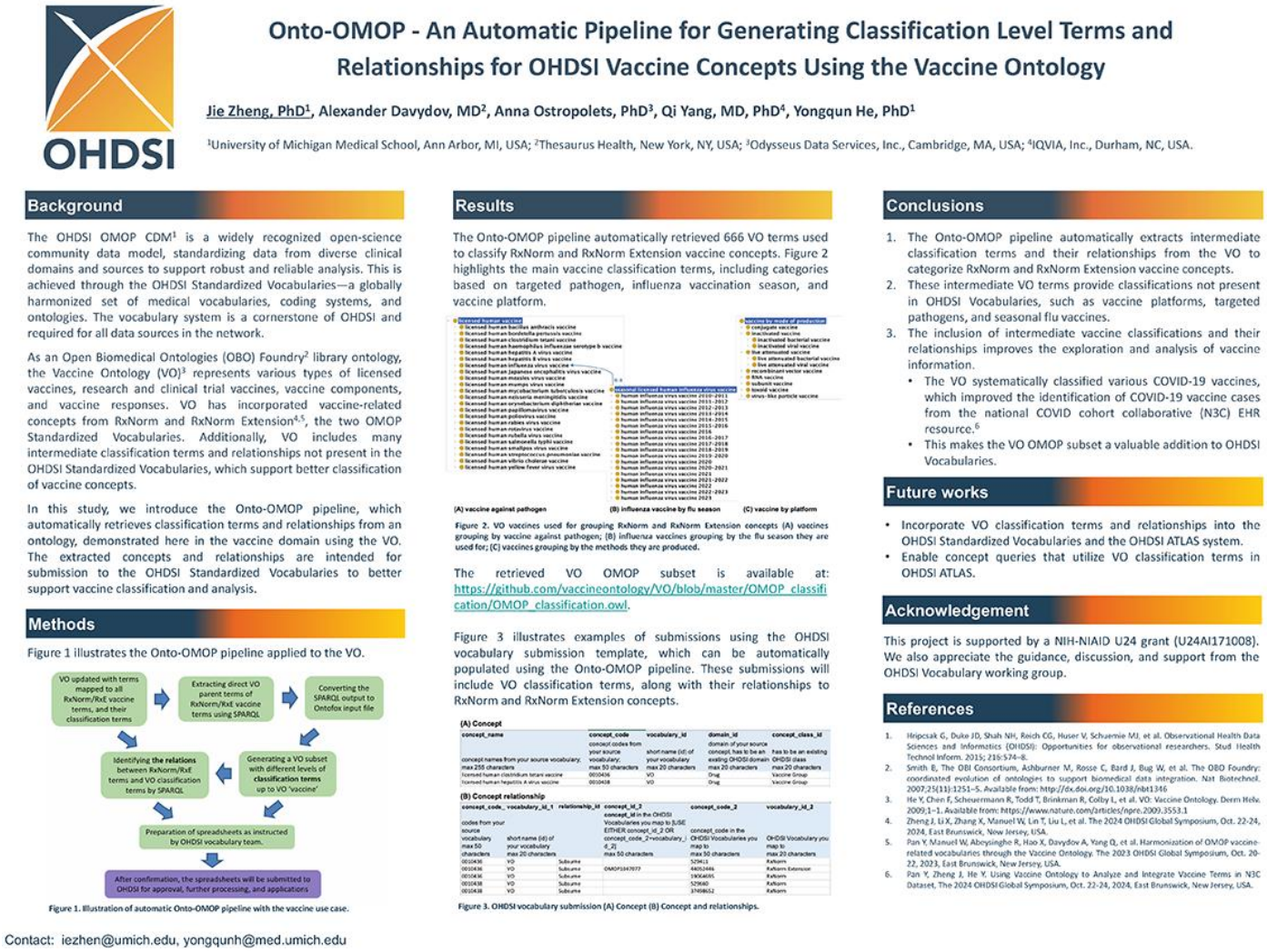
UT Southwestern Medical Center
Texas Health Resources

#OHDSISocialShowcase This Week

Wednesday

Onto-OMOP: An Automatic Pipeline for Generating Classification Level Terms and Relationships for OHDSI Vaccine Concepts Using the Vaccine Ontology

(Jie Zheng, Alexander Davydov, Anna Ostropolets, Qi Yang, Yongqun He)



Onto-OMOP - An Automatic Pipeline for Generating Classification Level Terms and Relationships for OHDSI Vaccine Concepts Using the Vaccine Ontology

Jie Zheng, PhD[1], Alexander Davydov, MD[2], Anna Ostropolets, PhD[3], Qi Yang, MD, PhD[4], Yongqun He, PhD[1]

[1]University of Michigan Medical School, Ann Arbor, MI, USA; [2]Thesaurus Health, New York, NY, USA; [3]Odysseus Data Services, Inc., Cambridge, MA, USA; [4]IQVIA, Inc., Durham, NC, USA.

Background

The OHDSI OMOP CDM[1] is a widely recognized open-science community data model, standardizing data from diverse clinical domains and sources to support robust and reliable analysis. This is achieved through the OHDSI Standardized Vocabularies—a globally harmonized set of medical vocabularies, coding systems, and ontologies. The vocabulary system is a cornerstone of OHDSI and required for all data sources in the network.

As an Open Biomedical Ontologies (OBO) Foundry[2] library ontology, the Vaccine Ontology (VO)[3] represents various types of licensed vaccines, research and clinical trial vaccines, vaccine components, and vaccine responses. VO has incorporated vaccine-related concepts from RxNorm and RxNorm Extension[4,5], the two OMOP Standardized Vocabularies. Additionally, VO includes many intermediate classification terms and relationships not present in the OHDSI Standardized Vocabularies, which support better classification of vaccine concepts.

In this study, we introduce the Onto-OMOP pipeline, which automatically retrieves classification terms and relationships from an ontology, demonstrated here in the vaccine domain using the VO. The extracted concepts and relationships are intended for submission to the OHDSI Standardized Vocabularies to better support vaccine classification and analysis.

Methods

Figure 1 illustrates the Onto-OMOP pipeline applied to the VO.

Figure 1. Illustration of automatic Onto-OMOP pipeline with the vaccine use case.

Contact: iezhen@umich.edu, yongqunh@med.umich.edu

Results

The Onto-OMOP pipeline automatically retrieved 666 VO terms used to classify RxNorm and RxNorm Extension vaccine concepts. Figure 2 highlights the main vaccine classification terms, including categories based on targeted pathogen, influenza vaccination season, and vaccine platform.

(A) vaccine against pathogen (B) influenza vaccine by flu season (C) vaccine by platform

Figure 2. VO vaccines used for grouping RxNorm and RxNorm Extension concepts (A) vaccines grouping by vaccine against pathogen; (B) influenza vaccines grouping by the flu season they are used for; (C) vaccines grouping by the methods they are produced.

The retrieved VO OMOP subset is available at: https://github.com/vaccineontology/VO/blob/master/OMOP_classification/OMOP_classification.owl.

Figure 3 illustrates examples of submissions using the OHDSI vocabulary submission template, which can be automatically populated using the Onto-OMOP pipeline. These submissions will include VO classification terms, along with their relationships to RxNorm and RxNorm Extension concepts.

(A) Concept

concept_name	concept_code	vocabulary_id	domain_id	concept_class_id
concept names from your source vocabulary; max 255 characters	concept codes from your source vocabulary; max 50 characters	short name (id) of your vocabulary max 20 characters	domain of your source concept; has to be an existing OHDSI domain	has to be an existing OHDSI class max 20 characters
licensed human clostridium tetani vaccine	0010436	VO	Drug	Vaccine Group
licensed human hepatitis A virus vaccine	0010438	VO	Drug	Vaccine Group

(B) Concept relationship

concept_code_	vocabulary_id_1	relationship_id	concept_id_2	concept_code_2	vocabulary_id_2
codes from your source vocabulary max 50 characters	short name (id) of your vocabulary; max 20 characters		concept_id in the OHDSI Vocabularies you map to (USE EITHER concept_id_2 OR concept_code_2+vocabulary_id_2) max 50 characters	concept_code in the OHDSI Vocabularies you map to max 50 characters	OHDSI Vocabulary you map to max 20 characters
0010436	VO	Subsume	529613		Reform
0010436	VO	Subsume	OMOP1047077	44052446	Reform Extension
0010436	VO	Subsume		19004695	Reform
0010438	VO	Subsume		529640	Reform
0010438	VO	Subsume		37498652	Reform

Figure 3. OHDSI vocabulary submission (A) Concept (B) Concept and relationships.

Conclusions

1. The Onto-OMOP pipeline automatically extracts intermediate classification terms and their relationships from the VO to categorize RxNorm and RxNorm Extension vaccine concepts.
2. These intermediate VO terms provide classifications not present in OHDSI Vocabularies, such as vaccine platforms, targeted pathogens, and seasonal flu vaccines.
3. The inclusion of intermediate vaccine classifications and their relationships improves the exploration and analysis of vaccine information.
   - The VO systematically classified various COVID-19 vaccines, which improved the identification of COVID-19 vaccine cases from the national COVID cohort collaborative (N3C) EHR resource.[6]
   - This makes the VO OMOP subset a valuable addition to OHDSI Vocabularies.

Future works

- Incorporate VO classification terms and relationships into the OHDSI Standardized Vocabularies and the OHDSI ATLAS system.
- Enable concept queries that utilize VO classification terms in OHDSI ATLAS.

Acknowledgement

This project is supported by a NIH-NIAID U24 grant (U24AI171008). We also appreciate the guidance, discussion, and support from the OHDSI Vocabulary working group.

References

1. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. Stud Health Technol Inform. 2015; 216:574–8.
2. Smith B, The OBI Consortium, Ashburner M, Rosse C, Bard J, Bug W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25(11):1251–5. Available from: http://dx.doi.org/10.1038/nbt1346
3. He Y, Chen E, Scheuermann R, Todd T, Brinkman R, Colby L, et al. VO: Vaccine Ontology. Derm Helv. 2009;1–1. Available from: https://www.nature.com/articles/npre.2009.3553.1
4. Zheng J, Li X, Zhang X, Manuel W, Lin T, Liu L, et al. The 2024 OHDSI Global Symposium, Oct. 22-24, 2024, East Brunswick, New Jersey, USA.
5. Pan Y, Manuel W, Abeysinghe R, Hao X, Davydov A, Yang Q, et al. Harmonization of OMOP vaccine-related vocabularies through the Vaccine Ontology. The 2023 OHDSI Global Symposium, Oct. 20-22, 2023, East Brunswick, New Jersey, USA.
6. Pan Y, Zheng J, He Y. Using Vaccine Ontology to Analyze and Integrate Vaccine Terms in N3C Dataset, The 2024 OHDSI Global Symposium, Oct. 22-24, 2024, East Brunswick, New Jersey, USA.

#OHDSISocialShowcase This Week

Thursday

Assessing Temporal Data Quality of Rheumatoid and Psoriatic Arthritis Patients in the All of Us Research Program

(Matthew Spotnitz, John Giannini, Emily Clark, Yechiam Ostchega, Tamara R Litwin, Lewis Berman)

OHDSI

Assessing Temporal Data Quality of Rheumatoid and Psoriatic Arthritis Patients in the *All of Us* Research Program

Matthew Spotnitz, MD, MPH[1], John Giannini, PhD[1], Emily Clark, MPH[2], Yechiam Ostchega, PhD, RN (retired)[1], Tamara R Litwin, PhD, MPH[1], Lewis Berman, PhD, MS (retired)[1]

[1]*All of Us* Research Program, Office of the Director, National Institutes of Health, [2]GAP Solutions, Inc.

All of Us — THE FUTURE OF HEALTH BEGINS WITH YOU

Key Points

- Data quality assessments are essential for rigorous and reproducible observational healthcare research.
- However, many of those assessments lack an evaluation of temporality, which is a measure of whether the data elements follow an expected temporal order.
- In this analysis, we present a novel temporality assessment for rheumatoid and psoriatic arthritis cohorts.

Background

- Rheumatoid and Psoriatic Arthritis (RA and PsA) are autoimmune diseases that have overlapping clinical symptoms and can cause debilitating joint pain [1,2].
- Previously, we used a data quality framework, to evaluate cohorts of *All of Us* participants who had a ductal carcinoma in situ (DCIS) diagnosis, a mastectomy procedure, and multiple surgical oncology procedures [3 - 5].
- In this analysis we assessed the data quality of participant diagnoses for RA and PsA with the portion of the framework that focused on temporality.

Methods

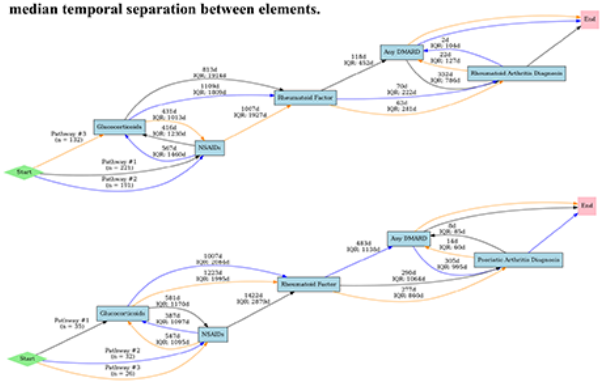
- Phenotype definition: The first occurrence of an ICD-9 or SNOMED CT diagnosis code for RA or PsA in adult *All of Us* Research Program participants.
- Temporality was measured by calculating the median intervals in days between the following five clinical concept sets: 1) Nonsteroidal anti-inflammatory drugs (NSAIDs), 2) glucocorticoids, 3) rheumatoid factor, 4) any disease-modifying antirheumatic drug (DMARD), and 5) diagnosis. The interval and count calculations are shown in the table.
- We plotted the temporal order of those concept sets for the subset of participants who had data on all of them. The plots were generated using the Python graphviz library. No automated tools were used. The three most common pathways are shown in the figure.

Results

Table: Counts and median intervals in days for Rheumatoid Arthritis and Psoriatic Arthritis Cohorts in different temporal analyses.

Analysis	Rheumatoid Arthritis No. (%)	Median (IQR) (In Days)	Psoriatic Arthritis No. (%)	Median (IQR) (In Days)
RF to Diagnosis	4638 (43.1)	19.4 (779.8)	585 (31.9)	62.3 (929.6)
Diagnosis to csDMARD	4856 (45.2)	0 (535.5)	705 (38.4)	0 (557.9)
Diagnosis to NSAID	8714 (81.1)	-279.3 (2094.4)	1394 (76.0)	-588 (-2417.1)
Diagnosis to Glucocorticoids	9123 (84.9)	-117.6 (1852.2)	1491 (81.3)	-560 (-2198)
RF to csDMARD	2780 (25.9)	46.9 (593.6)	314 (17.1)	126 (827.4)
RF to Biologic DMARD	1415 (13.2)	461.3 (1682.1)	296 (16.1)	304.5 (1260.7)
RF to NSAID	4042 (37.6)	-492.4 (2112.6)	497 (27.1)	-685.3 (2517.9)
RF to Glucocorticoids	4126 (38.4)	-212.5 (1871.8)	509 (27.8)	-851.9 (2377.9)

Figure: Timeline analysis in days of subsets of Rheumatoid (top) and Psoriatic (bottom) Arthritis cohorts with representative concept sets. For subset cohorts where participants have all relevant concept sets, these timelines show the most common orderings and median temporal separation between elements.



Abbreviations: Rheumatoid factor = RF; conventional synthetic disease-modifying antirheumatic drug = csDMARD; disease-modifying antirheumatic drug = DMARD; nonsteroidal anti-inflammatory drug = NSAID; interquartile range = IQR; count = n; days = d. Data source = The *All of Us* Research Program.

Discussion

- We have shown successful implementation of our data quality dimensions analysis to evaluate temporality of RA and PsA cohorts.
- For the subgroup of participants in the cohorts with all five clinical concept sets, the general chronological progression was from symptomatic management to laboratory tests to DMARD therapy or diagnosis.
- The lag between symptomatic treatment and diagnosis ranged from 4 to 5 years in the RA subgroups and was 6 years in the PsA subgroups.
- Our methods can be used to compare the temporality of cohorts from different data sources, including gold standard data sources.
- Our analysis was limited by the amount of data available and manual concept selection. Also, it is unclear if alternative phenotype definitions would have yielded different results.

Conclusions

We have shown successful implementation of our data quality dimensions analysis to evaluate temporality of RA and PsA cohorts. Our findings may have implications for whether RA and PsA diagnosis codes are fit for use as temporal anchors.

References

[1] Ritchlin CT, Colbert RA, and Gladman DD. Psoriatic Arthritis. N Engl J Med 2017;376:957-70. doi:10.1056/NEJMra1505557.

[2] Matteo AD, Barbon JM, and Emery P. Rheumatoid Arthritis. Lancet 2025; 405: 2019–33. https://doi.org/10.1016/

[3] Bertram L, Ostchega Y, Giannini J, et. al. Application of a Data Quality Framework to Ductal Carcinoma in Situ Using Electronic Health Record Data from the All of Us Research Program. JCO Clinical Cancer Informatics. 2024;8:e2400052. doi: 10.1200/CCI.24.00052.

[4] Spotnitz M, Giannini J, Ostchega Y, et. al. Assessing the Data Quality Dimensions of Partial and Complete Mastectomy Cohorts in the All of Us Research Program: Cross-Sectional Study. JMIR Cancer 2025;11:e76204. doi: 10.2196/76204. PMID: 40068169.

[5] Spotnitz M, Giannini J, Clark E, et al. Assessing the Data Quality Dimensions of Surgical Oncology Cohorts in the All of Us Research Program. JCO Clinical Cancer Informatics. (2025):9:e2500078. https://doi.org/10.1200/CCI-25-00.

[6] Kahn MG, Callahan TJ, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC) 2016;4(1):1244. Published 2016 September 11. PMID: 27713905.

[7] Albers DJ, Elhadad N, Tabak E, Perotte A, Hripcsak G (2014) Dynamical Phenotyping: Using Temporal Analysis of Clinically Collected Physiologic Data to Stratify Populations. PLoS ONE 9(6): e96443. doi:10.1371/journal.pone.0096443.

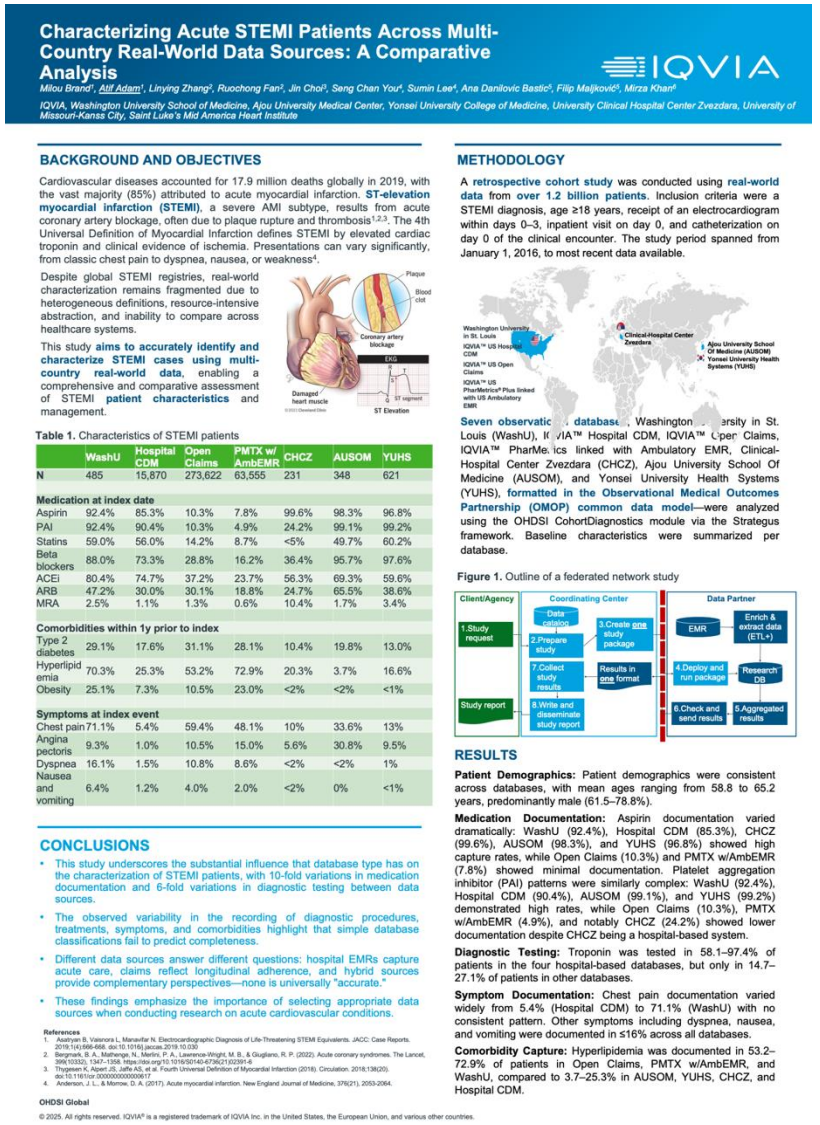
Contact: matthew.spotnitz@nih.gov

#OHDSISocialShowcase This Week

Friday

Characterizing Acute STEMI Patients Across Multi-Country Real-World Data Sources: A Comparative Analysis

(Milou Brand, Atif Adam, Linying Zhang, Ruochong Fan, Jin Choi, Seng Chan You, Sumin Lee, Ana Danilovic Bastic, Filip Maljkovi, Mirza Khan)



Characterizing Acute STEMI Patients Across Multi-Country Real-World Data Sources: A Comparative Analysis

Milou Brand[1], Atif Adam[1], Linying Zhang[2], Ruochong Fan[2], Jin Choi[3], Seng Chan You[4], Sumin Lee[4], Ana Danilovic Bastic[5], Filip Maljković[5], Mirza Khan[6]

IQVIA, Washington University School of Medicine, Ajou University Medical Center, Yonsei University College of Medicine, University Clinical Hospital Center Zvezdara, University of Missouri-Kansas City, Saint Luke's Mid America Heart Institute

BACKGROUND AND OBJECTIVES

Cardiovascular diseases accounted for 17.9 million deaths globally in 2019, with the vast majority (85%) attributed to acute myocardial infarction. **ST-elevation myocardial infarction (STEMI)**, a severe AMI subtype, results from acute coronary artery blockage, often due to plaque rupture and thrombosis[1,2,3]. The 4th Universal Definition of Myocardial Infarction defines STEMI by elevated cardiac troponin and clinical evidence of ischemia. Presentations can vary significantly, from classic chest pain to dyspnea, nausea, or weakness[4].

Despite global STEMI registries, real-world characterization remains fragmented due to heterogeneous definitions, resource-intensive abstraction, and inability to compare across healthcare systems.

This study **aims to accurately identify and characterize STEMI cases using multi-country real-world data**, enabling a comprehensive and comparative assessment of STEMI **patient characteristics** and management.

Table 1. Characteristics of STEMI patients

	WashU	Hospital CDM	Open Claims	PMTX w/ AmbEMR	CHCZ	AUSOM	YUHS
N	485	15,870	273,622	63,555	231	348	621
Medication at index date							
Aspirin	92.4%	85.3%	10.3%	7.8%	99.6%	98.3%	96.8%
PAI	92.4%	90.4%	10.3%	4.9%	24.2%	99.1%	99.2%
Statins	59.0%	56.0%	14.2%	8.7%	<5%	49.7%	60.2%
Beta blockers	88.0%	73.3%	28.8%	16.2%	36.4%	95.7%	97.6%
ACEi	80.4%	74.7%	37.2%	23.7%	56.3%	69.3%	59.6%
ARB	47.2%	30.0%	30.1%	18.8%	24.7%	65.5%	38.6%
MRA	2.5%	1.1%	1.3%	0.6%	10.4%	1.7%	3.4%
Comorbidities within 1y prior to index							
Type 2 diabetes	29.1%	17.6%	31.1%	28.1%	10.4%	19.8%	13.0%
Hyperlipidemia	70.3%	25.3%	53.2%	72.9%	20.3%	3.7%	16.6%
Obesity	25.1%	7.3%	10.5%	23.0%	<2%	<2%	<1%
Symptoms at index event							
Chest pain	71.1%	5.4%	59.4%	48.1%	10%	33.6%	13%
Angina pectoris	9.3%	1.0%	10.5%	15.0%	5.6%	30.8%	9.5%
Dyspnea	16.1%	1.5%	10.8%	8.6%	<2%	<2%	1%
Nausea and vomiting	6.4%	1.2%	4.0%	2.0%	<2%	0%	<1%

CONCLUSIONS

- This study underscores the substantial influence that database type has on the characterization of STEMI patients, with 10-fold variations in medication documentation and 6-fold variations in diagnostic testing between data sources.
- The observed variability in the recording of diagnostic procedures, treatments, symptoms, and comorbidities highlight that simple database classifications fail to predict completeness.
- Different data sources answer different questions: hospital EMRs capture acute care, claims reflect longitudinal adherence, and hybrid sources provide complementary perspectives—none is universally "accurate."
- These findings emphasize the importance of selecting appropriate data sources when conducting research on acute cardiovascular conditions.

References

1. Asaliyan B, Vaisnora L, Manavifar N. Electrocardiographic Diagnosis of Life-Threatening STEMI Equivalents. JACC: Case Reports. 2019;1(4):666-668. doi:10.1016/j.jaccas.2019.10.030
2. Bergmark, B. A., Mathenge, N., Merlini, P. A., Lawrence-Wright, M. B., & Giugliano, R. P. (2022). Acute coronary syndromes. The Lancet, 399(10332), 1347–1358. https://doi.org/10.1016/S0140-6736(21)02391-6
3. Thygesen K, Alpert JS, Jaffe AS, et al. Fourth Universal Definition of Myocardial Infarction (2018). Circulation. 2018;138(20). doi:10.1161/cir.0000000000000617
4. Anderson, J. L., & Morrow, D. A. (2017). Acute myocardial infarction. New England Journal of Medicine, 376(21), 2053-2064.

OHDSI Global

METHODOLOGY

A **retrospective cohort study** was conducted using **real-world data** from **over 1.2 billion patients**. Inclusion criteria were a STEMI diagnosis, age ≥18 years, receipt of an electrocardiogram within days 0–3, inpatient visit on day 0, and catheterization on day 0 of the clinical encounter. The study period spanned from January 1, 2016, to most recent data available.

Seven observatio[nal] databas[es], Washington [U]niversity in St. Louis (WashU), I[Q]VIA™ Hospital CDM, IQVIA™ [O]pen Claims, IQVIA™ PharM[etr]ics linked with Ambulatory EMR, Clinical-Hospital Center Zvezdara (CHCZ), Ajou University School Of Medicine (AUSOM), and Yonsei University Health Systems (YUHS), **formatted in the Observational Medical Outcomes Partnership (OMOP) common data model**—were analyzed using the OHDSI CohortDiagnostics module via the Strategus framework. Baseline characteristics were summarized per database.

Figure 1. Outline of a federated network study

RESULTS

Patient Demographics: Patient demographics were consistent across databases, with mean ages ranging from 58.8 to 65.2 years, predominantly male (61.5–78.8%).

Medication Documentation: Aspirin documentation varied dramatically: WashU (92.4%), Hospital CDM (85.3%), CHCZ (99.6%), AUSOM (98.3%), and YUHS (96.8%) showed high capture rates, while Open Claims (10.3%) and PMTX w/AmbEMR (7.8%) showed minimal documentation. Platelet aggregation inhibitor (PAI) patterns were similarly complex: WashU (92.4%), Hospital CDM (90.4%), AUSOM (99.1%), and YUHS (99.2%) demonstrated high rates, while Open Claims (10.3%), PMTX w/AmbEMR (4.9%), and notably CHCZ (24.2%) showed lower documentation despite CHCZ being a hospital-based system.

Diagnostic Testing: Troponin was tested in 58.1–97.4% of patients in the four hospital-based databases, but only in 14.7–27.1% of patients in other databases.

Symptom Documentation: Chest pain documentation varied widely from 5.4% (Hospital CDM) to 71.1% (WashU) with no consistent pattern. Other symptoms including dyspnea, nausea, and vomiting were documented in ≤16% across all databases.

Comorbidity Capture: Hyperlipidemia was documented in 53.2–72.9% of patients in Open Claims, PMTX w/AmbEMR, and WashU, compared to 3.7–25.3% in AUSOM, YUHS, CHCZ, and Hospital CDM.

Where Are We Going?

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

2026 OHDSI Vaccine Vocabulary WG

WG Co-Leads: Yongqun “Oliver” He & Alexander Davydov

- History of OHDSI Vaccine Vocabulary WG:
  - The ultimate challenge: classification and standardization of OHDSI vaccines.
  - 2021: Issues identified by Merck Lixia Yao group
  - 2022: Semi-automated method proposed by Licong Cui et al. PMID: 36029954
  - 2022: 2022: Identification of missing hierarchical links in the Vaccine Ontology. PMID: 35964149.
  - Jan 2023 - May 2026: Oliver He and Asiyah Yu Lin co-chaired the WG:
    - Members: Alexander Davydov, **Qi Yang (IQVIA)**, **Jie Zheng (U-M)**, Penny Pan, Lixia Yang, Anna Ostropolets, etc.
    - Key idea: Improve and use the Vaccine Ontology (VO) to support OHDSI vaccine classification.
    - We presented posters at the annual OHDSI Symposium every year.
    - Organized a Vaccine Vocabulary WG meeting in 2024.
    - Attended the OHDSI Vocabulathon in 2025.
    - Vaccine Ontology (VO) published in *Nature Scientific Data*: https://www.nature.com/articles/s41597-026-07298-w . PMID: 42062309.
- Since May 2026, Oliver He and Alexander Davydov have co-chaired the Vaccine Vocabulary WG
  - Overall Goal: Implement the VO vaccine classification system for practical OHDSI use.

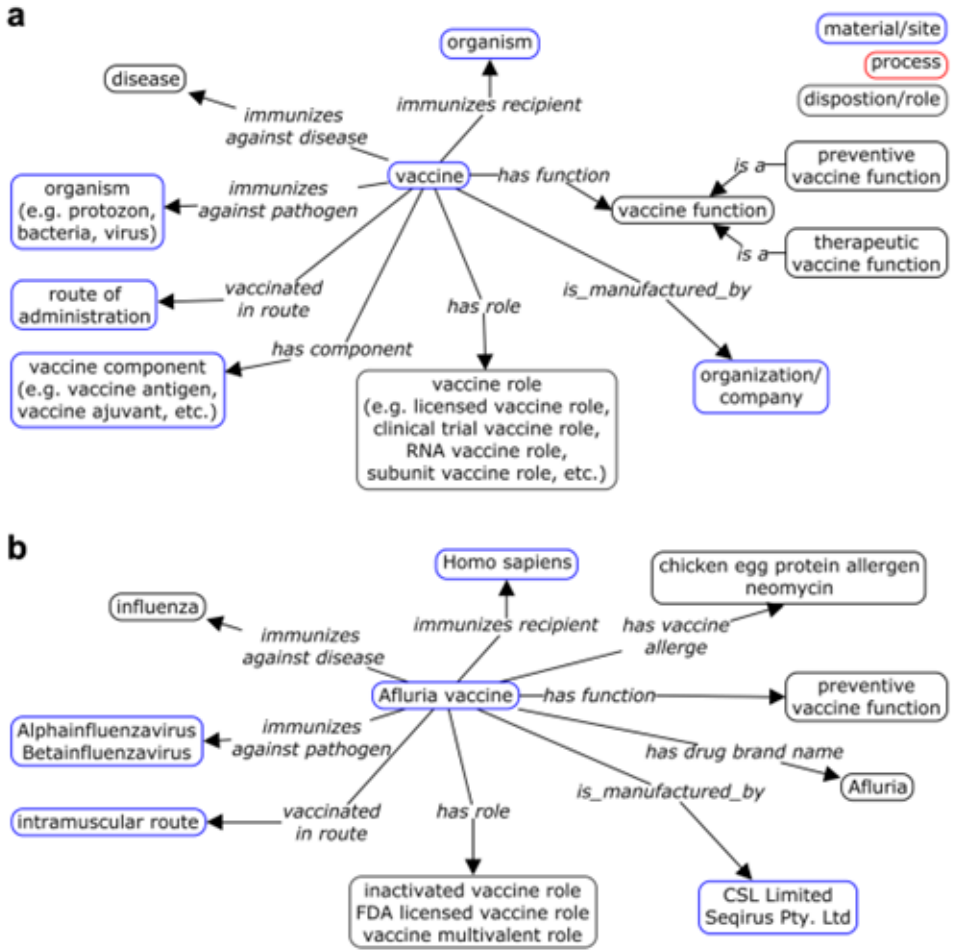
Summary of Vaccines in Vaccine Ontology (VO)

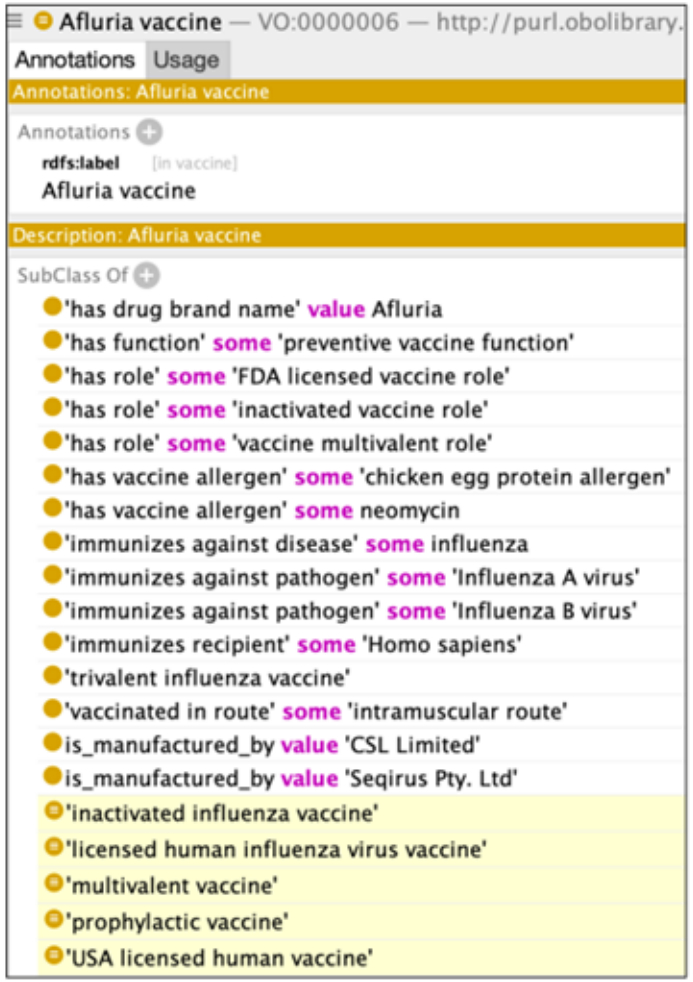
Development phase	Target host	Term resource	Vaccine Count	Total
Licensed/Authorized	Human	FDA	122	6,848
		CVX code	272	
		RxNorm	2,029	
		RxNorm Extension	4,074	
		Other	350	
	Veterinary animals	USDA	788	875
		Other	87	
In Clinical trial	Human	clinicaltrials.gov & NCI Thesaurus	427	805
	Human	Other	378	
In Research	Human + Veterinary animals	Literature and VO users	1,273	1,273

VO includes vaccines from different development phases and term resources

https://www.nature.com/articles/s41597-026-07298-w

Vaccine Ontology (VO) Design Pattern



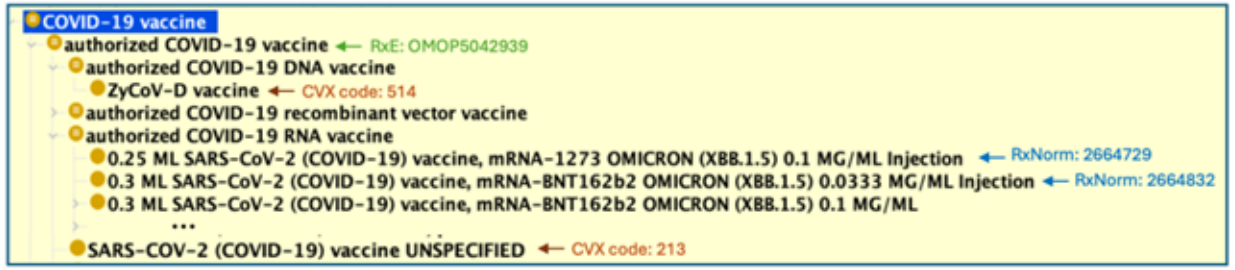


https://www.nature.com/articles/s41597-026-07298-w

Integration and Semantic Querying of OHDSI Vaccine Terminologies within VO

https://www.nature.com/articles/s41597-026-07298-w

a OHDSI COVID-19 vaccines under the VO hierarchical structure



b Find influenza vaccines in OHDSI vocabularies

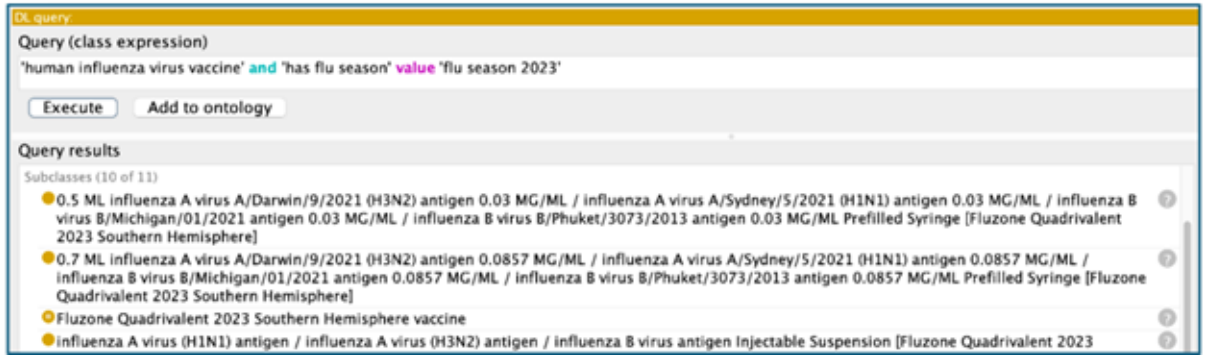
Query Text

```
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
SELECT DISTINCT ?VO_ID ?OMOP_Concept_ID ?CVX_code ?RxNorm_CUI ?RxNorm_Extension_ID ?name
FROM <http://purl.obolibrary.org/obo/merged/VO>
WHERE {
 ?entity rdfs:subClassOf* obo:VO_0003495 ;  # human influenza virus vaccine
       rdfs:label ?name ;
       obo:VO_0010151 ?OMOP_Concept_ID .
 BIND(COALESCE(IF(CONTAINS(STR(?entity), "/obo/"), STRAFTER(STR(?entity), "/obo/"), ?entity)) AS ?VO_ID)
 OPTIONAL { ?entity obo:VO_0005438 ?CVX_code }
 OPTIONAL { ?entity obo:VO_0003198 ?RxNorm_CUI }
 OPTIONAL { ?entity obo:VO_0010152 ?RxNorm_Extension_ID }}
ORDER BY ASC(?OMOP_Concept_ID)
```

Query Results

VO_ID	OMOP_Concept_ID	CVX_code	RxNorm_CUI	RxNorm_Extension_ID	name
VO_0018874	"1146353"		"2379630"		"influenza A virus A/Guangdong-Maonan/SWL1536/2019 (H1N1) antigen 0.03 MG/ML"
VO_0018875	"1146354"		"2379631"		"influenza A virus A/Hong Kong/2671/2019 (H3N2) antigen 0.03 MG/ML"
VO_0019759	"1146356"		"2379635"		"influenza A virus A/Guangdong-Maonan/SWL1536/2019 (H1N1) antigen 0.03 MG/ML /

c Find human influenza virus vaccines used in flu season 2022-2023 using DL query



2026 OHDSI Vaccine Vocabulary WG OKR

Objective 1: Re-examine and refine the **Vaccine Ontology (VO)** semantic and hierarchical model of licensed human vaccine terms **mapped to OMOP** vaccine standards (CVX, RxNorm, RxNorm Extension).

Key results:

1. Scheduled to present at the June OHDSI Vocabulary Workgroup meeting. Timeline: June 2026.
2. Submit an abstract(s) to the OHDSI 2026 Symposium. Deadline: June 5, 2026.
3. Reach agreement with the OHDSI Vocabulary WG. Timeline: Summer 2026.

2026 OHDSI Vaccine Vocabulary WG

Objective 2: Prepare a **submission** for the inclusion of VO to the OHDSI Vocabularies as a **non-standard/classification vocabulary**.

Key results:

1. Agree on pilot 2-3 use cases (e.g., using IQVIA and N3C OMOP data) with Workgroups and research leads.
2. Prepare a pilot submission, socialize and test it via a pre-release. Timeline: Summer 2026.
3. Execute a community contribution for VO inclusion to OMOP. Timeline: Q2-3 2026.
   - *We will refine and use our automated computational approach for a community contribution submission via a scripting pipeline or simple forms*

2026 OHDSI Vaccine Vocabulary WG

Objective 3: Use Case **Studies** and Dissemination of **Results**.

Key results:

1. Develop and evaluate the use of the VO vaccine classification for EHR data analysis using IQVIA and N3C data in OMOP format. Timeline: Q2-4 2026.
2. Attend and present at the **OHDSI 2026 symposium**. Timeline: Oct 20-22, 2026, New Brunswick, N.J. https://www.ohdsi.org/ohdsi2026/
3. Prepare a journal article. Timeline: Q3-4 2026.

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

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

Everybody is invited!

Links are sent out weekly and available at:

ohdsi.org/community-calls-2026

Find your workgroup.

Fuel our mission.

ohdsi.org/workgroups