OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Workgroup Spotlight: Natural Language Processing

OHDSI Community Call
July 14, 2026 • 11 am ET

Upcoming Community Calls

Date	Topic
July 14	Workgroup Spotlight: Natural Language Processing
July 21	Workgroup Spotlight: Health Systems, Surgery and Perioperative Medicine
July 28	OHDSI Newcomer Introductions
Aug. 4	TBA
Aug. 11	Oncology Research in OHDSI
Aug. 18	Workgroup Spotlight: Eyecare and Vision Research, Rehabilitation

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Jisu Park, Ju Hwan Kim, Yoon Seob Kim, Gyeongmin Lim, Hong Ji Song, Sang Youl Rhee, Seonghoon Hwang, Daeun Cho, Minyoung Kim, Bo-Yeon Kim, Chang Won Jeong, Dae-Yeon Kim, Hang A Park, Min-Ho Kim, Seong Hwan Kim, Won-Woo Seo, Woo Jin Kim, Young Sung Suh, Rae Woong Park, and Ju-Young Shin** on the recent publication of **GLP-1 Receptor Agonists for Weight Loss and Risk of Major Safety Outcomes: A Multicentre Cohort Study** in *Diabetes, Obesity and Metabolism*.


Diabetes, Obesity and Metabolism

WILEY

ORIGINAL ARTICLE

GLP-1 Receptor Agonists for Weight Loss and Risk of Major Safety Outcomes: A Multicentre Cohort Study

Jisu Park[1] | Ju Hwan Kim[1] | Yoon Seob Kim[2,3,4] | Gyeongmin Lim[1] | Hong Ji Song[5,6] | Sang Youl Rhee[7,8] | Seonghoon Hwang[1] | Daeun Cho[1] | Minyoung Kim[9] | Bo-Yeon Kim[10] | Chang Won Jeong[11] | Dae-Yeon Kim[12] | Hang A Park[13] | Min-Ho Kim[14] | Seong Hwan Kim[15] | Won-Woo Seo[16] | Woo Jin Kim[17] | Young Sung Suh[18] | Rae Woong Park[2,3,4,19] | Ju-Young Shin[1,20]

[1]School of Pharmacy, Sungkyunkwan University, Suwon, South Korea | [2]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea | [3]Center for Biomedical Informatics Research, Ajou University Medical Center, Suwon, Republic of Korea | [4]Department of Medical Sciences, Graduate School of Ajou University, Suwon, Republic of Korea | [5]Department of Family Medicine, Hallym University Sacred Heart Hospital, Anyang, Republic of Korea | [6]Department of Data Science, Hallym University, Chuncheon, Republic of Korea | [7]Department of Endocrinology and Metabolism, Kyung Hee University College of Medicine, Seoul, Republic of Korea | [8]Center for Digital Health, Kyung Hee University, Seoul, Republic of Korea | [9]HUBASE Co., Ltd., Seoul, Republic of Korea | [10]Division of Endocrinology and Metabolism, Department of Internal Medicine, Soonchunhyang University Bucheon Hospital, Soonchunhyang University College of Medicine, Bucheon, Republic of Korea | [11]Medical AI Team, Wonkwang University Hospital, Iksan, Republic of Korea | [12]Department of Internal Medicine, Soonchunhyang University Cheonan Hospital, Cheonan, Republic of Korea | [13]Department of Emergency Medicine, Hallym University, Dongtan Sacred Heart Hospital, Hwaseong-si, Republic of Korea | [14]Informatization Department, Ewha Womans University Seoul Hospital, Seoul, Republic of Korea | [15]Department of Orthopedic Surgery, Chung-Ang University Hospital, Seoul, Republic of Korea | [16]Department of Internal Medicine, Kangdong Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Republic of Korea | [17]Department on Internal Medicine, Kangwon National University, Chuncheon, Republic of Korea | [18]Department of Family Medicine, Keimyung University Dongsan Hospital, Daegu, Republic of Korea | [19]BK21 R&E Initiative for Advanced Precision Medicine, Suwon, Republic of Korea | [20]Department of Clinical Research Design & Evaluation, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul, Republic of Korea

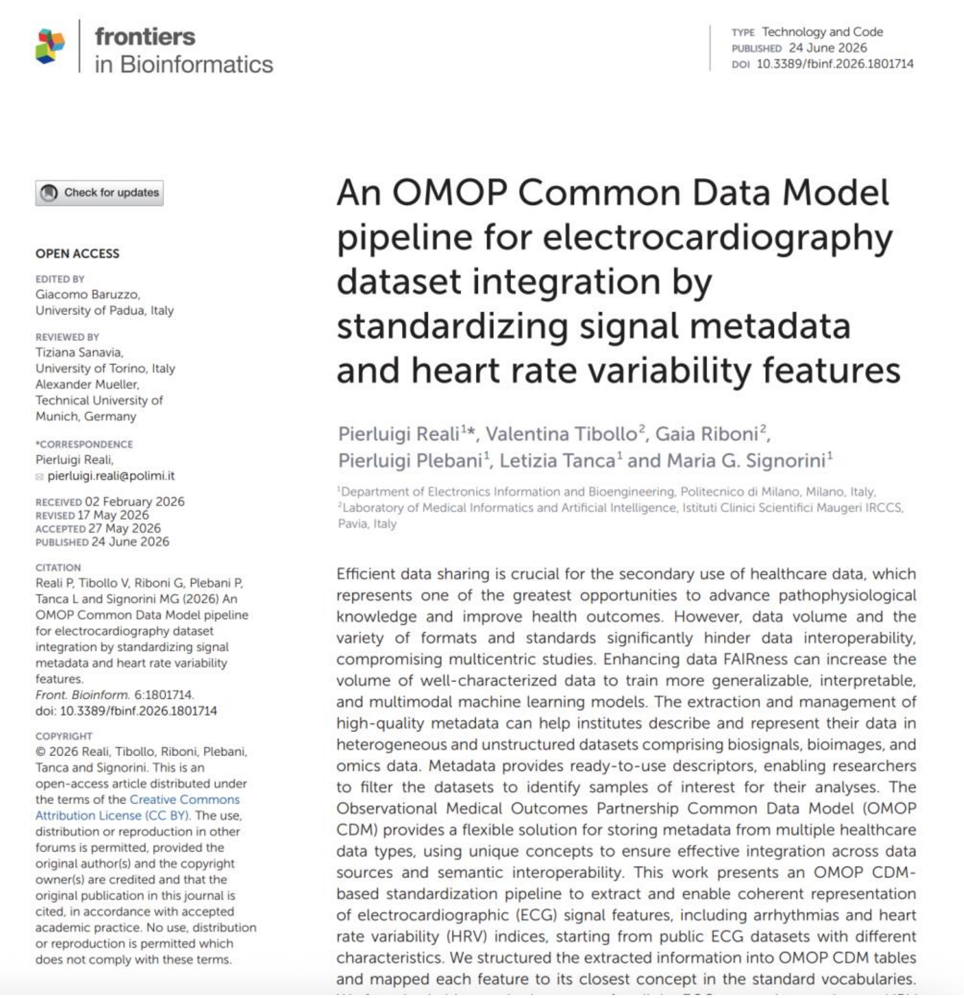
Correspondence: Rae Woong Park (veritas@ajou.ac.kr) | Ju-Young Shin (shin.jy@skku.edu)

Received: 27 March 2026 | **Revised:** 22 June 2026 | **Accepted:** 22 June 2026

Handling Editor: Hulin Tang

OHDSI Shoutouts!

Congratulations to the team of **Pierluigi Reali, Valentina Tibollo, Gaia Riboni, Pierluigi Plebani, Letizia Tanca, and Maria G Signorini** on the recent publication of **An OMOP Common Data Model pipeline for electrocardiography dataset integration by standardizing signal metadata and heart rate variability features** in *Frontiers in Bioinformatics*.



frontiers in Bioinformatics

TYPE Technology and Code
PUBLISHED 24 June 2026
DOI 10.3389/fbinf.2026.1801714

Check for updates

OPEN ACCESS

EDITED BY
Giacomo Baruzzo,
University of Padua, Italy

REVIEWED BY
Tiziana Sanavia,
University of Torino, Italy
Alexander Mueller,
Technical University of Munich, Germany

*CORRESPONDENCE
Pierluigi Reali,
pierluigi.reali@polimi.it

RECEIVED 02 February 2026
REVISED 17 May 2026
ACCEPTED 27 May 2026
PUBLISHED 24 June 2026

CITATION
Reali P, Tibollo V, Riboni G, Plebani P, Tanca L and Signorini MG (2026) An OMOP Common Data Model pipeline for electrocardiography dataset integration by standardizing signal metadata and heart rate variability features.
Front. Bioinform. 6:1801714.
doi: 10.3389/fbinf.2026.1801714

An OMOP Common Data Model pipeline for electrocardiography dataset integration by standardizing signal metadata and heart rate variability features

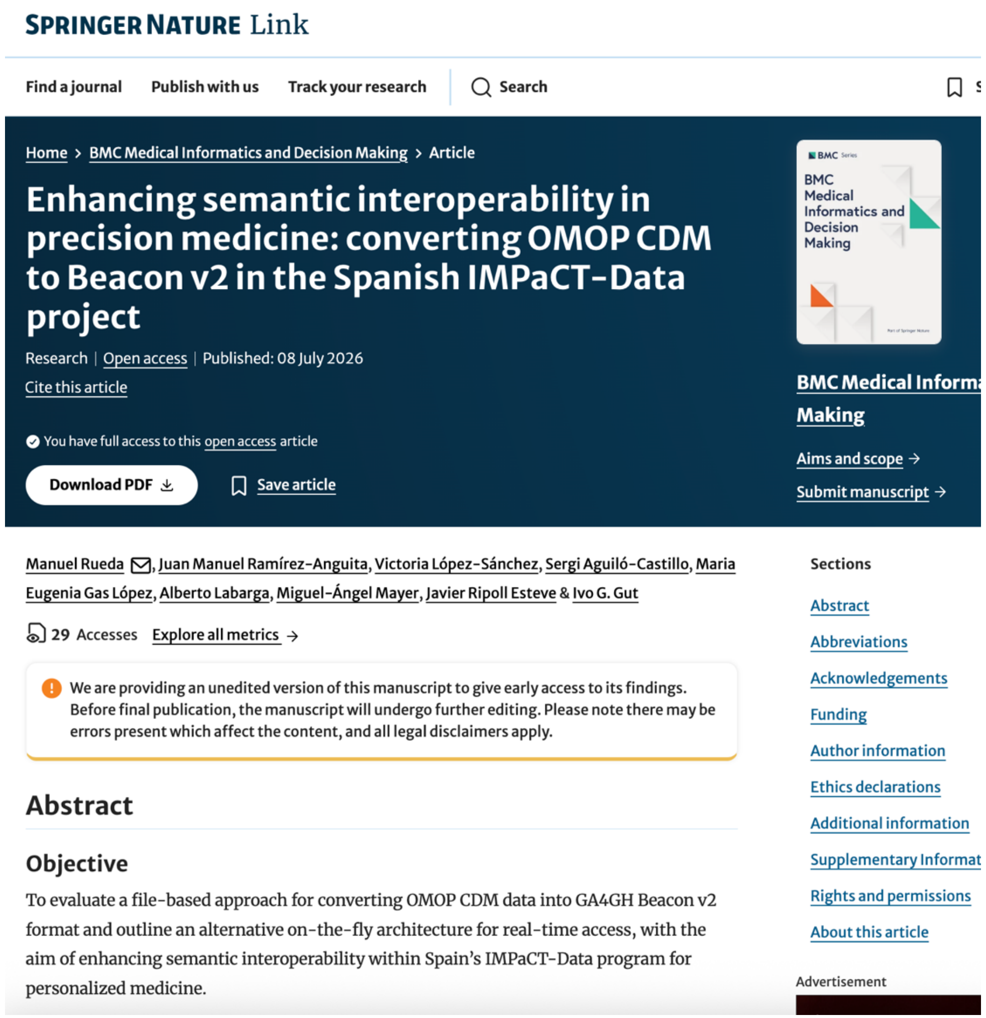
Pierluigi Reali[1]*, Valentina Tibollo[2], Gaia Riboni[2], Pierluigi Plebani[1], Letizia Tanca[1] and Maria G. Signorini[1]

[1]Department of Electronics Information and Bioengineering, Politecnico di Milano, Milano, Italy,
[2]Laboratory of Medical Informatics and Artificial Intelligence, Istituti Clinici Scientifici Maugeri IRCCS, Pavia, Italy

Efficient data sharing is crucial for the secondary use of healthcare data, which represents one of the greatest opportunities to advance pathophysiological knowledge and improve health outcomes. However, data volume and the variety of formats and standards significantly hinder data interoperability, compromising multicentric studies. Enhancing data FAIRness can increase the volume of well-characterized data to train more generalizable, interpretable, and multimodal machine learning models. The extraction and management of high-quality metadata can help institutes describe and represent their data in heterogeneous and unstructured datasets comprising biosignals, bioimages, and omics data. Metadata provides ready-to-use descriptors, enabling researchers to filter the datasets to identify samples of interest for their analyses. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) provides a flexible solution for storing metadata from multiple healthcare data types, using unique concepts to ensure effective integration across data sources and semantic interoperability. This work presents an OMOP CDM-based standardization pipeline to extract and enable coherent representation of electrocardiographic (ECG) signal features, including arrhythmias and heart rate variability (HRV) indices, starting from public ECG datasets with different characteristics. We structured the extracted information into OMOP CDM tables and mapped each feature to its closest concept in the standard vocabularies.

OHDSI Shoutouts!

Congratulations to the team of **Manuel Rueda, Juan Manuel Ramírez-Anguita, Victoria López-Sánchez, Sergi Aguiló-Castillo, Maria Eugenia Gas López, Alberto Labarga, Miguel-Ángel Mayer, Javier Ripoll Esteve, and Ivo G. Gut** on the recent publication of **Enhancing semantic interoperability in precision medicine: converting OMOP CDM to Beacon v2 in the Spanish IMPaCT-Data project** in *BMC Medical Informatics and Decision Making*.


SPRINGER NATURE Link

Find a journal | Publish with us | Track your research | Search

Home > BMC Medical Informatics and Decision Making > Article

Enhancing semantic interoperability in precision medicine: converting OMOP CDM to Beacon v2 in the Spanish IMPaCT-Data project

Research | Open access | Published: 08 July 2026

Cite this article

You have full access to this open access article

Download PDF | Save article

BMC Medical Informatics and Decision Making

Aims and scope →

Submit manuscript →

Manuel Rueda, Juan Manuel Ramírez-Anguita, Victoria López-Sánchez, Sergi Aguiló-Castillo, Maria Eugenia Gas López, Alberto Labarga, Miguel-Ángel Mayer, Javier Ripoll Esteve & Ivo G. Gut

29 Accesses | Explore all metrics →

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

Sections: Abstract, Abbreviations, Acknowledgements, Funding, Author information, Ethics declarations, Additional information, Supplementary Informat, Rights and permissions, About this article

Abstract

Objective

To evaluate a file-based approach for converting OMOP CDM data into GA4GH Beacon v2 format and outline an alternative on-the-fly architecture for real-time access, with the aim of enhancing semantic interoperability within Spain's IMPaCT-Data program for personalized medicine.

Advertisement

OHDSI Shoutouts!

Congratulations to the team of **Hanjae Kim, Minseong Kim, Seonji Kim, and Seng Chan You** on the recent publication of **From study design to executable code: automating target trial emulation with large language models** in *JAMIA Open*.

JAMIA Open, 2026, 9(4), ooag131
https://doi.org/10.1093/jamiaopen/ooag131
Research and Applications

AMIA OXFORD

From study design to executable code: automating target trial emulation with large language models

Hanjae Kim, BS[1,2,3], Minseong Kim, BS[1], Seonji Kim, PhD[1,2], Seng Chan You, MD, PhD*[,1,2,3]

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea
[2]Yonsei Institute for Digital Health, Yonsei University, Seoul, Republic of Korea
[3]Institute for AI and Social Innovation, Yonsei University, Seoul, Republic of Korea
H. Kim and M. Kim contributed equally.
*Corresponding author: Seng Chan You, Department of Biomedical Systems Informatics, Yonsei University College of Medicine, 50-1, Yonsei-Ro, Seodaemun-gu, Seoul, Republic of Korea (chandryou@yuhs.ac)

Abstract

Objective: Implementing target trial emulation (TTE) studies as standardized, reproducible analytic workflows is technically demanding. We developed Text-guided Health-study Estimation and Specification Engine Using Strategus (THESEUS), which uses large language models (LLMs) to translate free-text study descriptions into structured analytic specifications and Strategus R scripts within the Observational Health Data Sciences and Informatics (OHDSI) ecosystem.

Materials and Methods: THESEUS executes 2 steps: an LLM maps study descriptions to a JavaScript Object Notation (JSON) schema, and validated specifications are converted into Strategus R scripts through rule-based logic. For standardization evaluation, we compared specifications generated by 8 LLMs using 15 OHDSI-based TTE studies and 15 non-OHDSI studies under primary-analysis and full-analyses settings.

Results: Under the primary-analysis setting, overall standardization accuracy ranged from 0.93 to 0.97 across models in OHDSI studies and from 0.82 to 0.95 in non-OHDSI studies. Gemini-3.1-Pro achieved the highest overall accuracy in OHDSI studies, while Gemini-3.1-Pro and Gpt-5.5 jointly achieved the highest overall accuracy in non-OHDSI studies. Under the full-analyses setting, field-level sensitivity ranged from 0.83 to 0.97 in OHDSI studies, with 0.07–0.80 false positives (FPs) per study, and from 0.77 to 0.89 in non-OHDSI studies, with 0.53–1.20 FPs per study. Gpt-5.5 performed best at the field level. THESEUS was implemented as a web application and coding-agent tools.

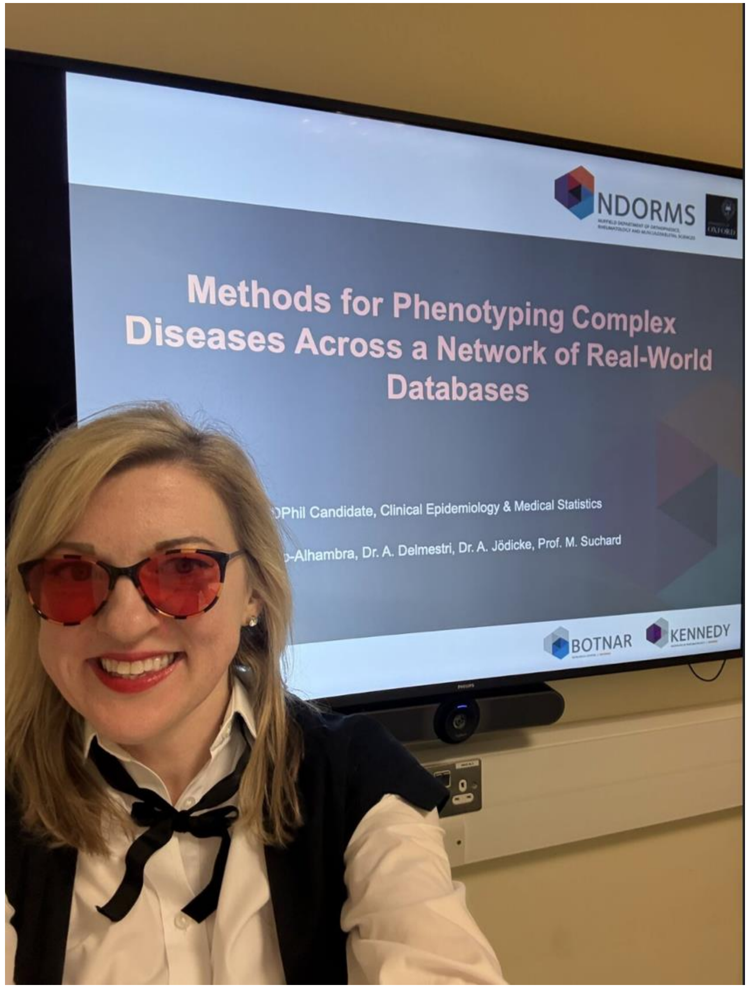
Discussion: Pairing a standardized data model with a structured analysis framework enables reliable LLM-assisted interpretation of study descriptions and deterministic workflow construction in observational research.

Conclusion: THESEUS supports translation of natural language study descriptions into executable, shareable code in standardized observational research settings.

Key words: target trial emulation; large language models; OHDSI; human-in-the-loop AI; reproducibility of results.

Congratulations Dr. Blacketer!

Congratulations Dr. Kostka!



Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Thursday	9 am	Oncology Vocabulary/Development Subgroup
Thursday	10 am	Africa Chapter
Thursday	10 am	GIS-Geographic Information System
Thursday	10:30 am	Evidence Network
Thursday	11 am	Themis
Friday	10:30 am	Education
Friday	11:30 am	Steering
Monday	9 am	Vaccine Vocabulary
Monday	2 pm	Electronic Animal Health Records
Tuesday	10 am	CDM Survey

Spotlight: Sulev Reisberg

In Estonia, the impact has been immense, especially among medical professionals. The dataset itself is not where the real value lies — the real value comes from the answers it can provide to doctors and researchers. One of the most rewarding aspects has been hearing feedback from highly experienced clinicians. Some top specialists with more than 20 years of experience in their fields have told us that this is the first time they have seen actual nationwide numbers for topics related to their specialty.

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Atlas 3 Demo Site

- Presented by Peter Hoffmann at 7/7 OHDSI Community Call
- We are asking for community feedback
- Published to https://atlas-preview.ohdsi.org/atlas/
  - Db Authenticaton enabled: use ohdsi/ohdsi
  - Will enable anonymous access in a future update
- Discussions at https://github.com/OHDSI/Atlas3/discussions

Wanted: HADES Developers

Help wanted - developers for HADES shiny viewer

Developers

jreps 5d

We have a set of R packages for developing modular shiny apps to explore the standardized results returned via HADES packages.

- OhdsiShinyAppBuilder - this builds the shiny app and defines what modules to include (e.g., cohort exporer, characterization, estimation, …)
- OhdsiShinyModules - the main shiny code for the key HADES packages are here
- OhdsiReportGenerator - this is a set of SQL that pulls out friendly data.framas with the results

However, there is a lot of potential for improvements in these packages including (but not limited to):

- adding bookmarking to the URL for easy sharing (Chapter 11 Bookmarking | Mastering Shiny)
- improving display time (improving data extraction speed)
- improving the UI

If you are a developer with knowledge or interest in R or shiny or web development/design and UI and want to help out with HADES development please post below or message me.

I'm planning to set up regular meetings for this work group.

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

2026 OHDSI Global Symposium

There are opportunities to be both a **sponsor** and an **exhibitor** at the Global Symposium.

Please reach out to symposium@OHDSI.org for more information.

ohdsi.org/OHDSI2026

First Latin America Symposium – July 30-31

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

Day 1

Strategic panels with government, academia and industry

Thursday, July 30, 2026

Opening and keynote

Common Data Model for Health Equity: the Role of Latin America.

Panel 1 — Health data interoperability and standards

Panelists from the Ministry of Health, Bahia State Health Department, PAHO and Latin American Governments.

Panel 2 — The power of administrative data for health research

Panelists from the Ministry of Health, CONASS, Fiocruz, Latin American Governments, Industry and OHDSI Global.

Panel 3 — The future of interoperability in healthcare in Latin America

A public–private debate.

Panelists from the Ministry of Health, CONASS, Fiocruz, private hospitals and Latin American Governments.

Day 2

Hands-on workshops and scientific collaboration

Friday, July 31, 2026

Introductory OMOP CDM workshops

- Introduction to OMOP
- Building cohorts with OHDSI tools

Parallel tracks of specialized workshops

- ETL to OMOP
- Scientific collaboration

Closing

Future perspectives and next steps for the OHDSI Latin America community.

ohdsilatam.org

UK Symposium: Registration Closes Aug. 20

HDR UK Event

OHDSI UK 2026

We're delighted to announce that OHDSI UK 2026 will be held on the 18th of September at the University of Nottingham. For the first time, there will also be an OMOP training day on the 17th of September.

Share this page

OHDSI (Observational Health Data Sciences and Informatics, pronounced "Odyssey") is an international community of stakeholders dedicated to unlocking the value of health data through large-scale analytics. OHDSI promotes open science and collaboration in health data research with a key focus on adoption of the OMOP Common Data Model, a global standard for harmonising data and facilitating federated analytics across institutions. **Find out more about OHDSI.**

Registration is now open!

Registration is now open for both the OHDSI UK 2026 Symposium and training day. Registration is scheduled to close on **20th August.** Please be aware that this event often sells out early.

FALCON Symposium & Bladder Cancer Studyathon

Sep 28 – Oct 1, 2026 Dunden, Antwerp, Belgium

Sep 28: Symposium — Public afternoon with oncologists, data scientists, regulators, and hospital leaders. Short talks and a network drink.

Sep 29 – Oct 1: Studyathon — Three hands-on days: cohort building, analytics, and abstract drafting.

Study related materials:

- All study related materials are here: GDE2025 - Bladder cancer treatment | OHDSI | Microsoft Teams
- GitHub repo: FALCON Bladder Git Repo

Register now

Details, agenda updates, venue information, and registration are available on the FALCON website.

www.falcon-network.org

You've built your OMOP CDM. Now let's make it shine!

Take your OMOP implementation to the next level

- 2.5-day hands-on expert training (max. 30 participants)
- Work with Data Quality Dashboard, ATLAS, CohortDiagnostics
- Learn study design, validation, execution
- Optional: bring your own ETL + OMOP CDM

What to expect

- Practical OMOP expertise
- Concrete CDM improvements
- Skills for evidence generation

Practical

📅 7–9 September 2026
📍 Rotterdam

#OHDSISocialShowcase This Week

Monday

ICD 10 to SNOMED-CT and OMOP Mapping: Performance, Limitations, and Clinical Implications of Automated Athena Mappings

(**Kirsten Colpaert**, Hadewig Claeys, Athiana Ryland, Maartje Deschacht, Mieke Deschepper)

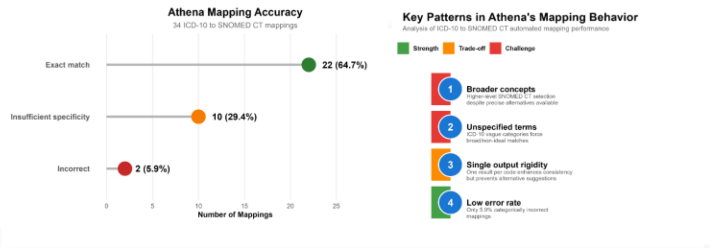
Athena automated mapping for ICD-10 → SNOMED CT loses specificity: **human expert review remains essential**

ICD 10 to SNOMED-CT and OMOP Mapping: Performance, Limitations, and Clinical Implications of Automated Athena Mappings

Background: Interoperability between ICD-10 and SNOMED CT is crucial for high-quality health data use, but differences in their structures make accurate mapping challenging. While Athena offers automated ICD-10-to-SNOMED CT conversions, this study evaluates how well these mappings preserve clinical specificity for muscular-dystrophy-related codes compared to manual expert mapping.

Methos: ICD-10 G71 muscular dystrophy codes were entered individually into the Athena "Non-standard → Standard (OMOP → SNOMED)" tool, which generated a single SNOMED CT concept for each code. In parallel, two researchers performed an independent manual SNOMED CT mapping using the international browser, reviewing synonyms, hierarchical relations, and clinical context, resolving uncertainties by consensus. Athena's outputs were then blindly compared to the manual mappings and classified as exact, insufficiently specific, or incorrect, after which results were summarized in absolute numbers and percentages. Preliminary testing of LLM-based mapping was excluded due to inconsistent and unreliable performance.

Results



Conclusions: Athena provides an efficient baseline for large-scale ICD-10 to SNOMED CT conversion (OMOP), but the 64.7% exact-match rate reveals that over one-third of mappings lack clinical precision. Loss of specificity, not outright errors, is the dominant issue. A **hybrid approach,** automated mapping followed by targeted expert validation, is the most practical

Deschepper Mieke, Claeys Hadewig, Ryland Athiana, Deschacht Maartje, Colpaert Kirsten – Ghent University Hospital, Belgium
Contact: mieke.deschepper@uzgent.be

#OHDSISocialShowcase This Week

Tuesday

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

(Dmytro Dymshyts, Anna Ostropolets, Martijn Schuemie)

LLMs can support automated mapping of ICD to SNOMED concepts

Title: Large Language Model–Based Classification of ICD-10-CM to SNOMED Mappings for Improved Semantic Fidelity in OHDSI

Background: Concept mapping in OHDSI can lose/corrupt clinical meaning when source and target terminologies don't align perfectly. LLMs are proposed as a solution to define type of semantic equivalence of mappings.

Methods

1. Data retrieval. Candidate SNOMED CT targets and synonyms retrieved for each ICD-10-CM source concept with 1:1 mapping (tested on a subset yet).

2. ICD10CM descriptions normalization. Non-specific ICD-10-CM modifiers ("unspecified", "NOS", "NEC", etc.) removed from source names using GPT-4o, preserving clinically meaningful attributes.
"AND" in selected ICD10CM concepts is replaced with "OR"

3. Mapping classification steps
3.1. text based match: Pairs with identical bags of words (after lowercasing, hyphen normalization, and stop-word removal) labelled "equal" ;
3.2. remaining pairs classified by GPT-o3 reasoning model into five labels: **equal, uphill, downhill, up&downhill, or incorrect**, with confidence levels.
3.3. Low/moderate-confidence labels validated through adversarial re-check* and majority-vote tiebreaker; all-differ cases flagged for manual review.

4. Output preparation. Final mappings split into a consistent set (targets in agreement or ≥1 "equal") and an inconsistent set (conflicting labels)

5. Quality check:
5.1. rerunning the model several times on a same subset and expecting the same results
5.2. review by a medical professional

*built-in validation first call prompt part:
"A previous classification assigned the label '<label>' with <confidence> confidence. Carefully reconsider. Apply all rules strictly and return your final label and confidence."
2nd validation call is on contrary independent from the previous calls.

Results: see output examples below

concept code	orig_source_name	source_norm_final	target_concept_name	outcome step1	confidence step1	outcome step2	confidence step2	outcome step3	confidence step3	Final outcome	Final confidence	decision method
G02	Meningitis in other infectious and parasitic diseases classified elsewhere	Meningitis in infectious or parasitic diseases	Infective meningitis	equal	moderate	equal	high	NA	NA	**equal**	high	LLM (validated)
G04.00	Acute disseminated encephalitis and encephalomyelitis, unspecified	Acute disseminated encephalitis or encephalomyelitis	Acute disseminated encephalomyelitis	NA	NA	NA	NA	NA	NA	**equal**	high	LLM
G04.01	Postinfectious acute disseminated encephalitis and encephalomyelitis (postinfectious ADEM)	Postinfectious acute disseminated encephalitis or encephalomyelitis (postinfectious ADEM)	Acute disseminated encephalomyelitis following infectious disease	downhill	moderate	downhill	moderate	NA	NA	**downhill**	moderate	LLM (validated)
G05.3	Encephalitis and encephalomyelitis in diseases classified elsewhere	Encephalitis or encephalomyelitis	Encephalitis, myelitis and encephalomyelitis	NA	NA	NA	NA	NA	NA	**uphill**	high	LLM
G12.1	Other inherited spinal muscular atrophy	Inherited spinal muscular atrophy	Spinal muscular atrophy	uphill	moderate	equal	moderate	uphill	high	**equal**	high	(tiebreak)
G31.09	Other frontotemporal neurocognitive disorder	Frontotemporal neurocognitive disorder	Frontotemporal dementia	NA	NA	NA	NA	NA	NA	**equal**	high	LLM
G31.83	Neurocognitive disorder with Lewy bodies	Neurocognitive disorder with Lewy bodies	Senile dementia of the Lewy body type	downhill	moderate	equal	moderate	equal	high	**equal**	high	(tiebreak)
G31.84	Mild cognitive impairment of uncertain or unknown etiology	Mild cognitive impairment of uncertain or unknown etiology	Minimal cognitive impairment	equal	moderate	equal	moderate	NA	NA	**equal**	moderate	LLM (validated)
G40.219	Localization-related (focal) (partial) symptomatic epilepsy and epileptic syndromes with complex partial seizures, intractable, without status epilepticus	Localization-related epilepsy or epileptic syndromes with complex partial seizures, intractable	Focal epilepsy	NA	NA	NA	NA	NA	NA	**uphill**	high	LLM
G40.5	Epileptic seizures related to external causes	Epileptic seizures related to external causes	Epileptic seizure	NA	NA	NA	NA	NA	NA	**uphill**	high	LLM
G40.8	Other epilepsy and recurrent seizures	Epilepsy or recurrent seizures	Epilepsy	downhill	moderate	equal	moderate	equal	moderate	**equal**	moderate	(tiebreak)
G40.824	Epileptic spasms, intractable, without status epilepticus	Epileptic spasms, intractable	Refractory infantile spasms	NA	NA	NA	NA	NA	NA	**downhill**	high	LLM
G04.3	Acute necrotizing hemorrhagic encephalopathy	Acute necrotizing hemorrhagic encephalopathy	Acute hemorrhagic leukoencephalitis	NA	NA	NA	NA	NA	NA	**incorrect**	high	LLM
G40.11	Localization-related (focal) (partial) symptomatic epilepsy and epileptic syndromes with simple partial seizures, intractable	Localization-related (focal) (partial) symptomatic epilepsy or epileptic syndromes with simple partial seizures, intractable	Focal onset aware epileptic seizure	NA	NA	NA	NA	NA	NA	**incorrect**	high	LLM
G40.A0	Absence epileptic syndrome, not intractable	Absence epileptic syndrome, not intractable	Absence seizure	uphill	moderate	incorrect	high	incorrect	moderate	**incorrect**	high	LLM (tiebreak)
G43.B0	Ophthalmoplegic migraine, not intractable	Ophthalmoplegic migraine, not intractable	Recurrent painful ophthalmoplegic neuropathy	uphill	moderate	uphill	high	NA	NA	**uphill**	high	(validated)

Out of 715 1:1 mappings of entire ICD10CM F chapter, 6 were labeled for manual review, 7 as incorrect, 45 as downhill, 4 as up&downhill, 154 as uphill, 510 as equal. After review by medical professional 17 labels were considered wrong - model was mostly confused by *harmful use* vs *abuse* vs *disorder related to use of psychoactive substance*

Conclusion: The automated mapping classification results are promising and mostly consistent. Non-equal mappings detected will be replaced with better target concept equivalents (decision to be made: what to do if no equivalent target exists)

OHDSI

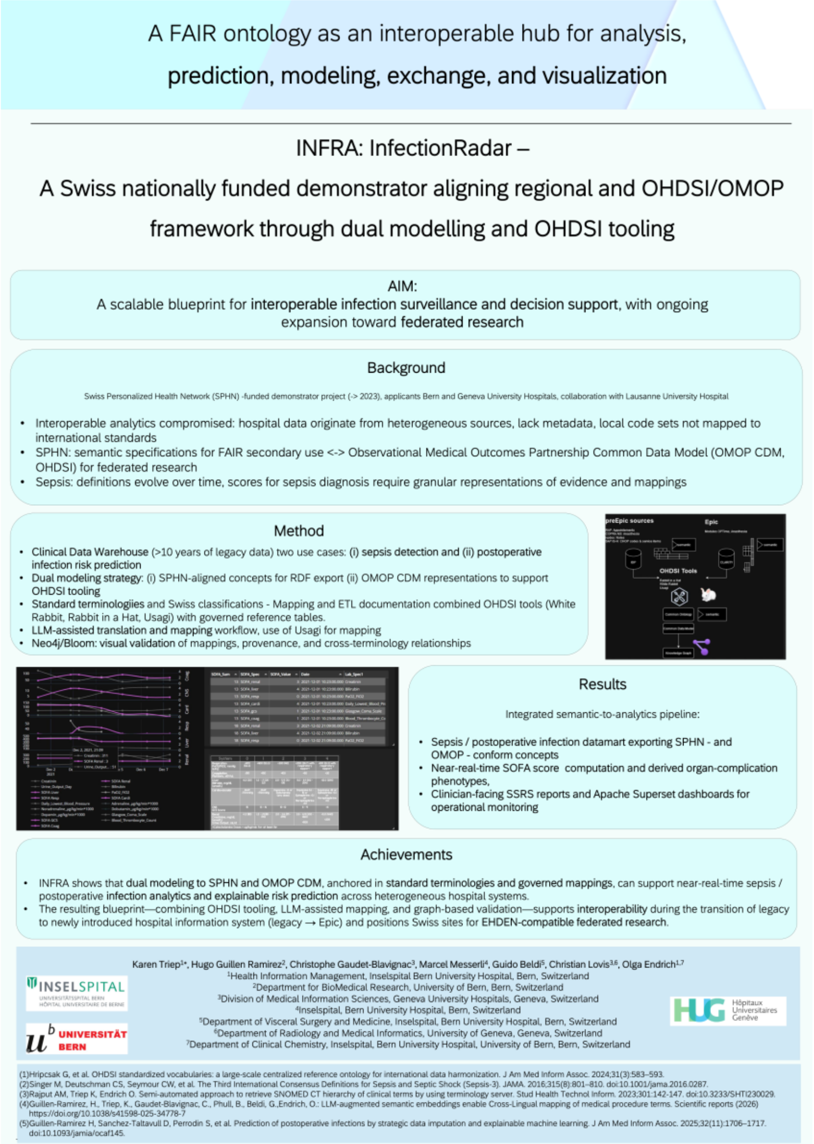
Dmytro Dymshyts, Anna Ostropolets, Martijn Schuemie
Johnson &Johnson

#OHDSISocialShowcase This Week

Wednesday

INFRA: InfectionRadar – A Swiss nationally funded demonstrator aligning regional and OHDSI/OMOP framework through dual modelling and OHDSI tooling

(Karen Triep, Hugo Guillen Ramirez, Christophe Gaudet-Blavignac, Marcel Messerli, Guido Beldi, Christian Lovis, Olga Endrich)

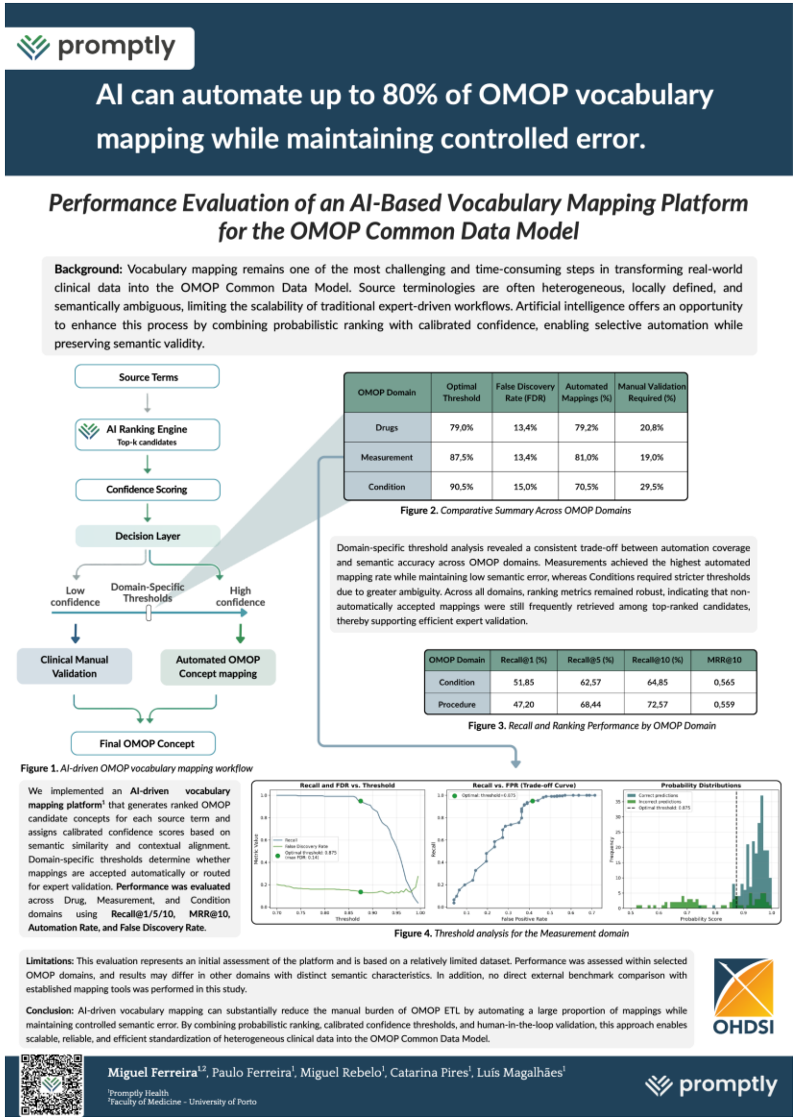


#OHDSISocialShowcase This Week

Thursday

Performance Evaluation of an AI-Based Vocabulary Mapping Platform for the OMOP Common Data Model

(Miguel Ferreira, Paulo Ferreira, Miguel Rebelo, Catarina Pires, Luís Magalhães)


promptly

AI can automate up to 80% of OMOP vocabulary mapping while maintaining controlled error.

Performance Evaluation of an AI-Based Vocabulary Mapping Platform for the OMOP Common Data Model

Background: Vocabulary mapping remains one of the most challenging and time-consuming steps in transforming real-world clinical data into the OMOP Common Data Model. Source terminologies are often heterogeneous, locally defined, and semantically ambiguous, limiting the scalability of traditional expert-driven workflows. Artificial intelligence offers an opportunity to enhance this process by combining probabilistic ranking with calibrated confidence, enabling selective automation while preserving semantic validity.

Figure 1. *AI-driven OMOP vocabulary mapping workflow*

OMOP Domain	Optimal Threshold	False Discovery Rate (FDR)	Automated Mappings (%)	Manual Validation Required (%)
Drugs	79,0%	13,4%	79,2%	20,8%
Measurement	87,5%	13,4%	81,0%	19,0%
Condition	90,5%	15,0%	70,5%	29,5%

Figure 2. *Comparative Summary Across OMOP Domains*

Domain-specific threshold analysis revealed a consistent trade-off between automation coverage and semantic accuracy across OMOP domains. Measurements achieved the highest automated mapping rate while maintaining low semantic error, whereas Conditions required stricter thresholds due to greater ambiguity. Across all domains, ranking metrics remained robust, indicating that non-automatically accepted mappings were still frequently retrieved among top-ranked candidates, thereby supporting efficient expert validation.

OMOP Domain	Recall@1 (%)	Recall@5 (%)	Recall@10 (%)	MRR@10
Condition	51,85	62,57	64,85	0,565
Procedure	47,20	68,44	72,57	0,559

Figure 3. *Recall and Ranking Performance by OMOP Domain*

We implemented an **AI-driven vocabulary mapping platform**[1] that generates ranked OMOP candidate concepts for each source term and assigns calibrated confidence scores based on semantic similarity and contextual alignment. Domain-specific thresholds determine whether mappings are accepted automatically or routed for expert validation. **Performance was evaluated** across Drug, Measurement, and Condition domains using **Recall@1/5/10, MRR@10, Automation Rate, and False Discovery Rate.**

Figure 4. *Threshold analysis for the Measurement domain*

Limitations: This evaluation represents an initial assessment of the platform and is based on a relatively limited dataset. Performance was assessed within selected OMOP domains, and results may differ in other domains with distinct semantic characteristics. In addition, no direct external benchmark comparison with established mapping tools was performed in this study.

Conclusion: AI-driven vocabulary mapping can substantially reduce the manual burden of OMOP ETL by automating a large proportion of mappings while maintaining controlled semantic error. By combining probabilistic ranking, calibrated confidence thresholds, and human-in-the-loop validation, this approach enables scalable, reliable, and efficient standardization of heterogeneous clinical data into the OMOP Common Data Model.

OHDSI

Miguel Ferreira[1,2], Paulo Ferreira[1], Miguel Rebelo[1], Catarina Pires[1], Luís Magalhães[1]

[1]Promptly Health
[2]Faculty of Medicine - University of Porto

promptly

#OHDSISocialShowcase This Week

Friday

Semantic Drift in Real-World Data: A Framework for Detecting and Managing Concept Evolution in OMOP-CDM

(Yogesh Kumar Gupta)

Where Are We Going?

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

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

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

Everybody is invited!

Links are sent out weekly and available at:

ohdsi.org/community-calls-2026

Find your workgroup.

Fuel our mission.

ohdsi.org/workgroups