



Poster/Demo Preview & OHDSI2026 Final Logistics

OHDSI Community Call
Oct. 13, 2026 • 11 am ET



Upcoming Community Calls

Date	Topic
Oct. 13	Global Symposium Poster Preview / Final Logistics
Oct. 20	NO CALL – Global Symposium
Oct. 27	Welcome to OHDSI: What We Do, How Can You Make an Impact
Nov. 3	Meet the Titans
Nov. 10	Collaborator Showcase Honoree Presentations
Nov. 17	Early-Stage Researcher Presentations
Nov. 24	Software Demos
Dec. 1	OHDSI/OMOP Research Spotlight
Dec. 8	How Did OHDSI Do This Year?
Dec. 15	Holiday-Themed Fairwell to 2026



Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?



OHDSI Shoutouts!



Congratulations to the team of **Jordan G Nestor, Sarath Babu Krishna Murthy, Casey Ta, Teena Zachariah, Mariela Navarro-Torres, Natasha S Freeman, Heedeok Han, Sindhuri Prakash, Woojin Ahn, Yilu Fang, Atlas Khan, Jai Radhakrishnan, Krzysztof Kiryluk, and Chunhua Weng** on the recent publication of **Lost in transformation: diagnostic aggregation and semantic drift in ICD-derived kidney phenotypes** in *JAMIA*.

Journal of the American Medical Informatics Association, 2026, 1-11
<https://doi.org/10.1093/jamia/ocag167>
Brief Communications

AMIA OXFORD
ADVANCING THE PROFESSIONAL, LEADING THE WAY

Lost in transformation: diagnostic aggregation and semantic drift in ICD-derived kidney phenotypes

Jordan G. Nestor, MD, MS^{1,2,*}, Sarath Babu Krishna Murthy, MS², Casey Ta, PhD³, Teena Zachariah, MD¹, Mariela Navarro-Torres, MD¹, Natasha S. Freeman, MD¹, Heedeok Han, MD¹, Sindhuri Prakash, MD, PhD⁴, Woojin Ahn, MD, PhD⁵, Yilu Fang, MA³, Atlas Khan, PhD^{1,2}, Jai Radhakrishnan, MD, MS¹, Krzysztof Kiryluk, MD, MS^{1,2}, Chunhua Weng, PhD³

¹Department of Medicine, Division of Nephrology, Columbia University, New York, NY 10032, United States

²Department of Medicine, Center for Precision Medicine and Genomics, Columbia University, New York, NY 10032, United States

³Department of Biomedical Informatics, Columbia University, New York, NY 10032, United States

⁴Department of Medicine, Division of Nephrology, New York University, New York, NY 10016, United States

⁵Department of Medicine, Division of Nephrology and Hypertension, Oregon Health & Science University, Portland, OR 97239, United States

*Corresponding author: Jordan G. Nestor, MD, MS, Department of Medicine, Center for Precision Medicine and Genomics, Columbia University, 622 West 168th Street, PH4-124, New York, NY 10032, United States (jgn2108@cumc.columbia.edu)

Abstract

Objectives: To evaluate how cross-terminology mapping of International Classification of Diseases (ICD)-coded diagnoses is associated with preservation of diagnostic distinctions and traceability in electronic health record (EHR)-derived phenotypes for genomic research.

Materials and Methods: We conducted a retrospective, cross-sectional study of real-world ICD code utilization and semantic fidelity among kidney-related diagnoses. Kidney-relevant ICD-9-CM and ICD-10-CM codes were identified using clinician-curated OHDSI ATLAS searches and evaluated in a large academic health system. ICD codes were mapped to Phecodes, SNOMED-CT, the Human Phenotype Ontology (HPO), and Online Mendelian Inheritance in Man (OMIM) using standardized crosswalks. Mapping coverage, richness, and multiplicity were quantified, with multiplicity operationalizing diagnostic aggregation. Three nephrologists independently rated label-level semantic fidelity (0-2 scale). Ordinal logistic regression evaluated associations between mapping multiplicity and semantic fidelity.

Results: Among 585 nephrology-relevant ICD codes (571 435 patients), mapping coverage and richness varied across terminologies. Phecodes showed high coverage and mapping multiplicity, whereas SNOMED-CT, HPO, and OMIM mappings were predominantly one-to-one. Among ICD-Phecode mappings, 39% were rated poor, whereas no poor matches were observed among mapped SNOMED-CT, HPO, or OMIM concepts. Greater mapping multiplicity was associated with lower semantic fidelity (OR, 0.33; 95% CI, 0.29-0.37; $P < .001$). Most ICD-9-CM codes (96%) were represented in at least 1 external cohort.

Discussion: Diagnostic aggregation was associated with reduced semantic fidelity and traceability, with potential implications for phenotype construction and interpretation in GWAS, PheWAS, and biobank-scale studies.

Conclusion: This study provides an empirical framework for evaluating semantic drift in EHR-derived phenotypes and highlights a central tradeoff: abstraction enables scalability, whereas traceability enables translation.

Key words: electronic health records; International Classification of Diseases; genome-wide association studies; kidney diseases; precision medicine.



OHDSI Shoutouts!



Congratulations to the team of **Brandon Jernigan, Hannah L Walters, Diane Corey, Albert Garcia, Alan Premasiri, William T Roddy, Dillon Shearer, Terri G Thompson, Jason Tucker, Fernando Vieira, and Danielle Boyce** on the recent publication of **Harmonizing ALS research data across multimodal data sets using the OMOP Common Data Model** in *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*.

Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 2026; 0: 1–3



BRIEF REPORT

Harmonizing ALS research data across multimodal data sets using the OMOP Common Data Model

BRANDON JERNIGAN^{1,2}, HANNAH L. WALTERS³, DIANE COREY⁴, ALBERT GARCIA⁴, ALAN PREMASIRI³, WILLIAM T. RODDY⁴, DILLON SHEARER¹, TERRI G. THOMPSON^{1,2}, JASON TUCKER⁴, FERNANDO VIEIRA³ & DANIELLE BOYCE^{3,5}

¹Answer ALS, New Orleans, LA, USA, ²Onpoint Scientific, Inc, San Diego, CA, USA, ³ALS Therapy Development Institute, Watertown, MA, USA, ⁴Critical Path Institute, Tucson, AZ, USA, and ⁵Johns Hopkins University School of Medicine, Baltimore, MD, USA

Abstract

Three nonprofit organizations—ALS Therapy Development Institute (ALS TDI), Answer ALS, and the Critical Path Institute (C-Path)—collaborated to harmonize distinct ALS research datasets using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). This effort was designed to make otherwise separate ALS resources more comparable, enabling larger and more informative cross-cohort analyses. The teams coordinated vocabulary mapping, identified OMOP vocabulary gaps, and submitted new concepts to the Observational Health Data Science and Informatics (OHDSI) community for approval. Harmonized data from ALS TDI and Answer ALS were merged in the Neuromine data portal, while their clinical and digital health technology (DHT) datasets are also available through C-Path's Rare Disease Cures Accelerator-Data Analytics Platform (RDCA-DAP). This collaboration aligns with current regulatory priorities for data sharing and offers a model for harmonizing rare disease data. A robust, public GitHub repository was established to share real-time updates with the ALS research community.

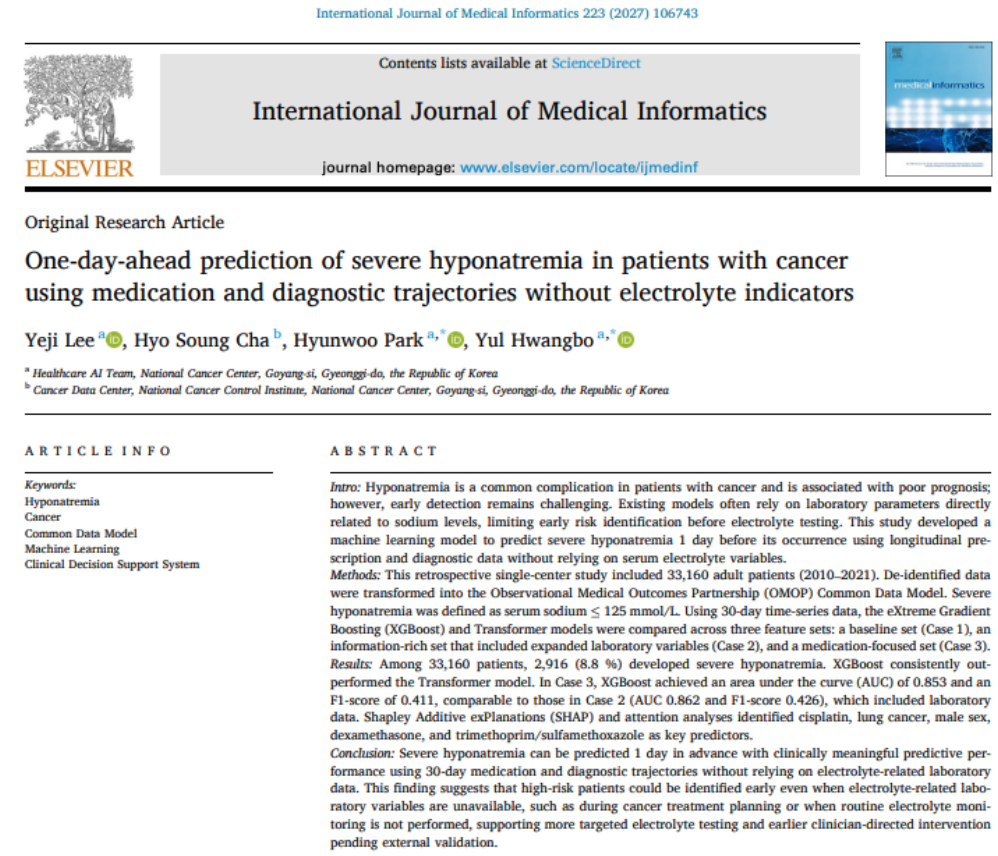
Keywords: ALS, OMOP, data harmonization, digital health technology, Act for ALS



OHDSI Shoutouts!



Congratulations to the team of **Yeji Lee, Hyo Soung Cha, Hyunwoo Park, and Yul Hwangbo** on the recent publication of **One-day-ahead prediction of severe hyponatremia in patients with cancer using medication and diagnostic trajectories without electrolyte indicators** in the *International Journal of Medical Informatics*.





Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?



Upcoming Workgroup Calls



Date	Time (ET)	Meeting
Tuesday	12 pm	Vocabulary
Tuesday	12 pm	Generative AI and Analytics
Wednesday	9 am	Patient-Level Prediction
Wednesday	10 am	Common Data Model
Wednesday	2 pm	Canada Chapter
Wednesday	7 pm	Medical Imaging
Wednesday	7 pm	Eyecare and Vision Research
Thursday	8 am	Medical Devices
Thursday	9 am	Oncology Vocabulary/Development Subgroup
Thursday	10 am	GIS - Geographic Information System
Thursday	10:30 am	Evidence Network
Thursday	12 pm	HADES
Thursday	7 pm	Dentistry
Friday	9 am	Waveform
Friday	10 am	Transplant
Friday	10:30 am	Education
Friday	11:30 am	Steering



Congratulations to our 2026 Titan Nominees!

ASSURE team, David Bard, Robert Barrett, Clair Blacketer, Rich Boyce, Danielle Boyce, Evanette Burrows, Cindy Cai, Carrot and Bunny Team, Cindy Chen, Pascal Coorevits, Data4Life, Scott Duvall, Noémie Elhadad, Lee Evans, Eye Care and Vision Research WG, FastOMOP, Mornin Feng, Davera Gabriel, Josh Gagne, Sarah Gasman, Lars Halvorsen, Peter Hoffmann, Christian Högberg, Michelle Hribar, Jason Hsu, Ben Illigens, Andy Kanter, Elisse Katzman, Masha Khitrin, Mack Kigada, Chungsoo Kim, Chris Knoll, Haeun (Hannah) Lee, Gaeun Lee, Renske Los, Benjamin Martin, Valentina Martufi, Mind Meets Machine, Evan Minty, Karthik Natarajan, Bill O'Brien, Sean O'Reilly, OHDSI UK organizers, Julio Oliveria, Anna Ostropolets, Woo Yeon (Jen) Park, Rowan Parry, Paul Petraro, Melanie Philofsky, Hanieh Razzaghi, Jenna Reys, Harry Reyes Nieva, Marti Sabate, Craig Sachson, Alexander Saelmans, Martijn Schuemie, Sarah Seager, Anthony Sena, Cynthia Sung, Jim Todd, Mui Van Zandt, Annelies Verbiest, Fleur Vereijken, Katia Verhamme, Ilse Vermeulen, David Vizcaya, Tom White, Qi Yang, Seng Chan You, Bingyu Zhang



Spotlight: Rupa Makadia

Rupa Makadia is a Director in Observational Health Data Sciences at Johnson & Johnson, based in Raritan, New Jersey. Her work focuses on generating real-world evidence to evaluate the safety and effectiveness of medical products, with particular emphasis on oncology and pregnancy research.

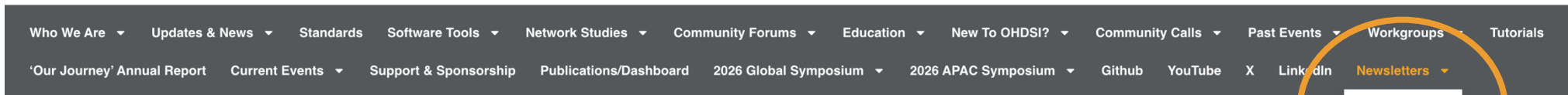
She leads the Perinatal and Reproductive Health Workgroup and contributes to methods development and network studies. She discusses her career journey and research interests, upcoming initiatives for the workgroup and more in the latest edition of the Collaborator Spotlight.

mailchi.mp/ohdsi/october2026





October Newsletter is Available



Welcome to OHDSI!

The Observational Health Data Sciences and Informatics (or OHDSI, pronounced "Odyssey") program is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics. All our solutions are open-source.

OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University.

Learn more [about our community](#), about [our mission, vision and values](#), and how you can [help support the OHDSI community](#).

[Join the Journey](#)

Join us at the 2026 OHDSI Global Symposium

Registration is OPEN for the 2026 OHDSI Global Symposium, which will be held Oct. 20-22 in New Brunswick, NJ. This event unites hundreds of collaborators to showcase scientific innovations and build bridges for future research. Together, we are advancing our mission to generate real-world evidence that informs better health decisions and improves patient care.

[Global Symposium Homepage](#)

[Register Me for OHDSI 2026](#)

What OHDSI is all about

OHDSI's mission

To improve health by empowering a

Latest News

• [Weekly OHDSI Digest – October 5, 2026](#)

[mailchi.mp/ohdsi/october2026](https://us4.campaign-archive.com/home/?u=9402e3dd40021ebda4b15a4b7&id=af9e5362f5)



APAC Symposium: Nov. 13-15

The 2026 APAC Symposium will be held Nov. 13-15 at Yonsei University in Seoul, South Korea.

Nov. 13: Main Conference

Nov. 14: Tutorials, Datathon

Nov. 15: Datathon Team Activities

Registration is open! All details can be found on the event homepage.



ohdsi.org/APAC2026



Canada Symposium: Nov. 17-18

Registration is now open for the 2026 OHDSI Canada Symposium, held Nov. 17-18 in Toronto, Ont.

Nov. 17: Workshops

Track A: From Source Data to OMOP

Track B: From OMOP to Evidence

Nov. 18: Symposium

The main symposium day will bring the Canadian OHDSI community together to share experiences, showcase emerging work, and explore opportunities for collaboration.



JOIN US IN TORONTO THIS NOVEMBER

2026 OHDSI Canada Symposium

DATE	LOCATION
Nov 17–18, 2026	Toronto, ON

Join the OHDSI Canada community in Toronto for two days of hands-on workshops, engaging talks, strategic discussions, and networking.

Workshop Day

Practical sessions on vocabulary mapping, ETL, and using OMOP and OHDSI to answer real-world research questions.

Symposium Day

Presentations, strategic discussions, and opportunities to connect with colleagues from across the OHDSI Canada community.

India Symposium: Dec. 11



OHDSI INDIA SYMPOSIUM 2026

11 DECEMBER 2026

CHENNAI

PARTNERING FOR IMPACT, GROWTH, AND VISIBILITY

COLLABORATE. INNOVATE. TRANSFORM.

A STRONG COMMUNITY | **DATA-DRIVEN INSIGHTS** | **REAL-WORLD IMPACT** | **GLOBAL VISION, LOCAL ACTION**

Be a Part of the Change



WHY TO ATTEND?

- **UNDERSTAND INDIA'S HEALTHCARE DATA LANDSCAPE**
Learn about the current state of healthcare data in India, emerging national priorities, and how global standards are shaping the future of healthcare and research.
- **EXPLORE OMOP AND HEALTHCARE DATA INITIATIVES**
Discover the latest developments in OMOP, Real-World Evidence (RWE), FHIR, AI, and other data standardization initiatives being implemented across India.
- **LEARN HOW YOUR ORGANIZATION CAN PARTICIPATE**
Understand how hospitals, medical colleges, and healthcare organizations can contribute to national data initiatives, multicenter research, training programs, and collaborative projects through OHDSI India.
- **LEARN FROM EXPERTS**
Gain insights from national and international experts on healthcare data standards, OMOP, FHIR, Real-World Evidence, AI, and observational research.
- **NETWORK WITH THE HEALTHCARE COMMUNITY**
Connect with clinicians, researchers, hospitals, academia, industry partners, and global OHDSI leaders to build meaningful collaborations
- **SHOWCASE YOURWORK**
SHOWCASE YOURWORK Present your research through posters, lightning talks, or case studies and receive valuable feedback from the community
- **BUILD FUTURE COLLABORATIONS**
Identify opportunities to participate in multicenter research, training programs, industry collaborations, and Real-World Evidence projects.



#OHDSISocialShowcase This Week

Monday

Scalable Extraction of APGAR Scores from Electronic Health Records Clinical Notes and Standardization for OMOP CDM

(**Carla Arxé**, Angela Leis, Juan Manuel Ramírez-Anguita, Antonio Arcas, Jan Carreras, Ricard García-Isern, Miguel-Angel Mayer)

- Enabling Large-Scale Observational Research on Neonatal Outcomes by Standardising Unstructured APGAR Data into the OMOP CDM
- Building High-Quality, Research-Ready Pregnancy Episode Cohorts by Overcoming EHR Text Redundancy and Format Heterogeneity

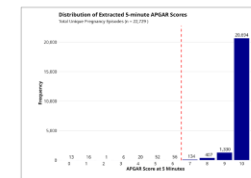
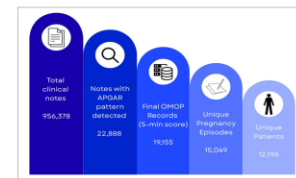
Scalable Extraction of APGAR Scores from Electronic Health Records Clinical Annotations and Standardization for OMOP CDM

Background: The APGAR is a quick, standardised assessment used right after birth to sum up how a newborn is doing and whether immediate support is needed. The score is based on five observable signs: appearance (skin colour), pulse (heart rate), grimace (reflex irritability in response to stimulation), activity (muscle tone), and respiration (breathing effort). Each component is rated 0, 1, or 2, and the total ranges from 0 to 10. In practice, a total of 7–10 is generally reassuring. To overcome the barrier of unstructured clinical text, we developed a scalable pipeline to extract and standardise it into the OMOP CDM.

Results

We processed 950,000 clinical notes, successfully filtering redundant documentation. Our temporal clustering logic (150-day window) identified 15,049 biologically distinct pregnancy episodes, covering the period 2008 to 2024.

The resulting distribution of 5-minute APGAR scores aligns with established neonatal population norms. The left-skew (majority of scores ≥ 7) serves as an internal validation of the extraction accuracy and clinical plausibility of our rule-based engine.



Methods

The extraction process consisted of four stages:

- 1 Regex Extraction & Parsing:** Developed robust Regular Expression patterns to identify APGAR sequences (e.g., "9/10/10") while handling heterogeneous formats and common character/OCR errors.
- 2 Handling Multiple Births:** Detected linguistic markers (e.g., "twins", "tetus 2") to extract multiple sets of scores and preserve sibling relationships within the same clinical visit.
- 3 Episode Definition (Temporal Clustering):** Grouped records into 150-day windows to collapse redundant notes into unique and valid pregnancy episodes.
- 4 OMOP Standardization:** Mapped scores to the measurement table and the Perinatal Extension (PET) of the OMOP CDM to ensure interoperability and mother-child traceability.

Conclusions

We successfully standardised unstructured APGAR data into the OMOP CDM with 99.6% coverage. A temporal clustering logic was essential to eliminate EHR redundancy and accurately define unique pregnancy episodes. Future work will compare this rule-based pipeline against Large Language Models (LLMs) to improve edge-case extraction and provide an automated validation layer.



Carla Arxé, Angela Leis, Juan Manuel Ramírez-Anguita, Antonio Arcas, Jan Carreras, Ricard García-Isern, Miguel-Angel Mayer



EUROPEAN OHDSI Symposium 16-20 April 2026
Rotterdam





#OHDSISocialShowcase This Week

Tuesday

Personalized Federated Framework for OMOP-CDM Multi-Hospital Data: Application to prediction of 30-day hospital readmission

(**Dani Manjah**, Pierre Remacle, Maryna Borshchivska, Thibaut Helleputte)

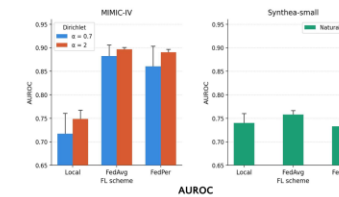
Personalized Federated Framework for OMOP-CDM Multi-Hospital Data

Application to prediction of 30-day hospital readmission

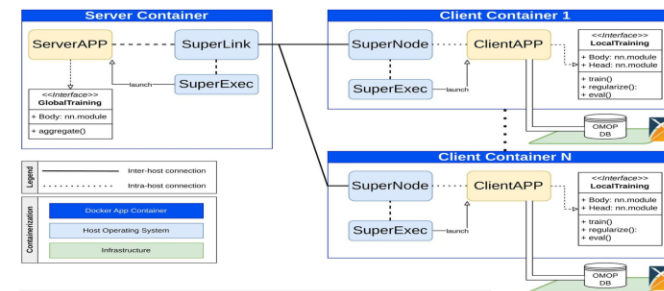
Background: OMOP-CDM standardizes observational health data across institutions, but privacy laws (GDPR) prevent centralized pooling. Existing FL frameworks require consortium-wide agreement on ML pipelines, overlooking client-specific needs. Our key innovation is dynamic runtime configuration: each hospital maintains its own model architecture and feature pipeline while participating in standardized, containerized, privacy-preserving distributed training.

Results

Methods	Metric	Datasets	
		MIMIC-IV $\alpha = 0.7$	Synthea $\alpha = 2$
Local	mean	0.762 ± 0.044	0.731 ± 0.020
	min	0.707	0.706
FedAvg ¹	mean	0.806 ± 0.036	0.842 ± 0.002
	min	0.757	0.839
FedPer ²	mean	0.831 ± 0.007	0.841 ± 0.003
	min	0.822	0.836
Pooled	-	0.86	0.75



Methods



Limitation: Validated on MIMIC-IV (partitioned) and Synthea (synthetic).
Real multi-site deployment ongoing. github.com/manjahdani/fedomop



Dani Manjah, Pierre Remacle, Maryna Borshchivska, Thibaut Helleputte



UCLouvain



#OHDSISocialShowcase This Week

Wednesday

Validating LLM-Based extraction of clinical events: Integrating unstructured EHRs into the OMOP Common Data Model: Suicide use case

(**Hermenegildo Martinez-Alcala**, Alicia Peñaranda-Navazo, Cristina Justo-Astorgano, Virginia Arroyo, Miguel Ángel Macía, Mar Martín, Elisa Martín-Merino, Miguel Gil, Jesús Santamaría)

Title: Validating LLM-Based extraction of clinical events: Integrating unstructured EHRs into the OMOP Common Data Model: Suicide use case

A. PRESENTER:
Hermenegildo Martinez-Alcala

INTRO

- **The Problem:** Structured diagnostic codes suffer from low sensitivity and high underreporting in complex behaviors like suicidality.
- **The Opportunity:** Large Language Models (LLMs) offer transformative potential for extracting and normalizing unstructured clinical narratives into the OMOP Common Data Model (CDM), but require validation to mitigate model errors and ensure evidentiary reliability.
- **The Scale:** This study utilizes the BIFAP database, containing over 2 billion unstructured health records from 30 million patients in Spain.

METHODS

The study implemented a six-step process to transform raw narratives into standardized evidence.

- I. **Use case definition:** Outcomes were categorized into completed suicide, suicide attempt, suicidal ideation, and not related.
- II. **Corpus generation:** Large-scale filtering of 2 billion records using 70+ regular expressions to identify event-related text.
 - **LLM pre-annotation:** Auto-labeling: GPT-4 (14B) for initial classification.
 - **Human validation:** Expert review of 5,040 texts (>90% inter-annotator agreement).
- III. **Model training:** Fine-tuning was performed on both encoder (XLM-RoBERTa) and decoder (GPT family) models.
- IV. **The Cascaded System:** To balance speed and accuracy, an encoder model acts as the first filter; any record with a confidence score below specific thresholds is delegated to a GPT-4 (14B) decoder or human review.
- V. **Large-Scale Execution:** The best model was applied across the entire 2-billion-record dataset.
- VI. **OMOP integration:** Validated events were mapped to SNOMED CT concepts and stored in the Note_rpt table for global interoperability.

RESULTS

- **High-Fidelity performance:** Preliminary results show robust classification, achieving an F1-score of 0.988 for "completed suicide" and 0.931 for "suicidal ideation".
- **Efficiency:** The system maintained high-throughput, with a delegation rate of only 9% for uncertain cases.
- **Uncertainty:** The "not related" category accounted for the highest level of model uncertainty.

Clinician-Driven AI

How 5,000 expert annotations drive LLM precision across BIFAP database
(>2 Billion health records)



Fig 1. Workflow of the process.

CONCLUSIONS

This hybrid architecture successfully balances computational scalability for multi-billion-record datasets with high-performance accuracy. By integrating these outputs into the OMOP CDM, we enable more precise cohort definitions and robust epidemiological analyses for public health challenges.

LIMITATIONS & FUTURE WORK

> **Population-Level validation**
While preliminary results are robust, future efforts will prioritize cross-validation against external, heterogeneous data sources to establish the generalizability of this volume is currently unfeasible, necessitating the development of more advanced automated uncertainty resolution or secondary LLM-based verification layers.

> **Scalability of human review**
Although the cascaded architecture successfully filters the vast majority of records, the 9% delegation rate for uncertain cases presents a significant operational challenge. At a scale of 2 billion records, manual clinical review of this volume is currently unfeasible, necessitating the development of more advanced automated uncertainty resolution or secondary LLM-based verification layers.

© H. Martinez-Alcala, A. Peñaranda-Navazo, C. Justo-Astorgano, V. Arroyo, M. Macía, M. Martín, E. Martín-Merino, M. Gil, J. Santamaría

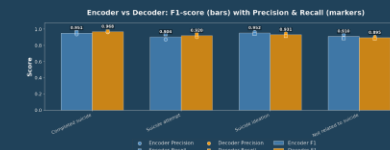


Fig 2. Benchmark evaluation of fine-tuned transformer models: precision, recall, and F1-score across suicidal behavior categories.



Take a picture to visit BIFAP





#OHDSISocialShowcase This Week

Thursday

Bilingual Clinical De-identification in Free-text Clinical Notes: A Hybrid Approach Combining Biomedical Transformers and Heuristic Rules

(Ricard Garcia-Isern, Antonio Arcas, Angela Leis, Juan Manuel Ramírez-Anguita, Carla Arxé, Jan Carreras, Miguel-Angel Mayer)

A hybrid AI + rules system detects more patient data than any single approach alone, ensuring strict GDPR compliance in bilingual hospitals

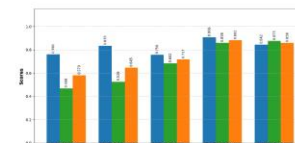
Bilingual Clinical De-Identification in Free-Text Annotations: A Hybrid Approach Combining Biomedical Transformers and heuristic Rules

Background: Clinical notes at Hospital del Mar mix Spanish and Catalan languages, and patient identifiers are sparse. This study presents ClinicaAño, a custom de-identification pipeline designed for Hospital del Mar. We propose a methodology that moves beyond simple fine-tuning, employing data-centric AI techniques to adapt a state-of-the-art Spanish model to the specific dialect and sparsity constraints of our hospital. **ClinicaAño** combines deterministic rules with a biomedical transformer to safely de-identify free-text.

Result 1 (The Privacy Win): The hybrid system achieves the highest recall (0.875), leaking the least patient data. The Transformer alone scores higher F1 (0.881), but in privacy, missing a name is worse than over-redacting.

Result 2 (Complementarity):

- Transformer detects context (Hospitals, Professions: F1=1.00)
- Rules Engine catches formats the AI misses (Phone numbers, Exact times: F1=1.00).
- Neither works alone.



Methods

- 1 Stratified Active Sampling:** Scanned 5,000 notes; extracted the top 100 Spanish & 100 Catalan notes with the highest PHI density.
- 2 Synthetic Header Injection:** Added fake patient headers to training data so the AI wouldn't ignore sparse names.
- 3 Silver Standard Annotation:** AI pre-labeled the 200 notes; human experts corrected errors in Label Studio. Faster with gold-standard quality.
- 4 Hybrid Fusion Engine:** Rules and AI run in parallel. A conflict-resolution engine prioritizes maximum PHI detection.



ClinicaAño successfully de-identifies bilingual clinical notes by fusing rule-based precision with AI recall to meet strict GDPR standards. Future work includes expanding the training corpus and deploying the system as a REST API for OMOP CDM integration.



Ricard Garcia-Isern, Antonio Arcas, Angela Leis, Juan Manuel Ramírez-Anguita, Carla Arxé Jan Carreras, Miguel-Angel Mayer



EUROPEAN OHDSI Symposium 18-20 April 2023 Rotterdam





#OHDSISocialShowcase This Week

Friday

Challenges in Generating Clinically Coherent Synthetic Electronic Health Records in OMOP: The Semantic Gap

(Jan Carreras, Juan Manuel Ramírez-Anguita, Ricard Garcia, Antonio Arcas, Carla Arxé, Angela Leis, Miguel Angel Mayer)

Developing a functional pipeline for generating synthetic OMOP CDM data

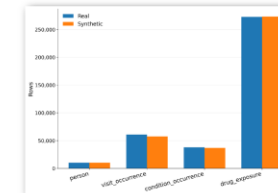
While structural validity and referential integrity were readily achieved, preserving high-level clinical semantics logic remains a key challenge for these models

Generating Clinically Coherent Synthetic Electronic Health Records in OMOP: Bridging the Semantic Gap

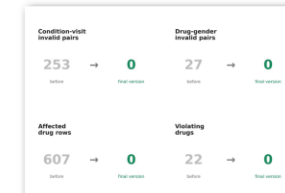
Background: Biomedical datasets are often constrained by stringent privacy requirements and limited accessibility. These factors hinder the widespread use of Electronic Health Records (EHRs) for observational research and methodological development. While Synthetic Data Generation (SDG) has emerged as a promising solution to share granular patient data while preserving privacy, producing synthetic clinical data of sufficient quality to serve as a reliable proxy for real patient records remains challenging. We aimed to build a usable synthetic OMOP dataset that preserves core clinical structure and coherence across four linked tables.

Results

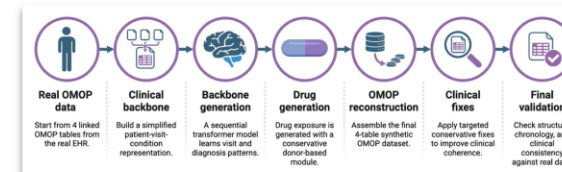
A full 4-table OMOP dataset at realistic scale
Synthetic row counts closely matched the original data across all 4 OMOP tables, supporting a complete and usable relational release rather than isolated synthetic records.



Targeted post-processing removed clinically invalid relationships
Conservative post-processing removed impossible condition-visit and drug-gender combinations, improving clinical coherence without changing the overall dataset structure.



Methods



We used a hybrid pipeline combining transformer-based backbone generation, conservative drug generation, OMOP reconstruction, and targeted clinical fixes.

Conclusions

Although the final 4-table synthetic OMOP dataset is usable and clinically safer than earlier candidates, concept-level calibration remains imperfect, and rare concepts are still under-represented. Future work will extend the pipeline to additional OMOP tables and improve long-tail clinical fidelity.



Jan Carreras, Juan Manuel Ramírez-Anguita, Ricard Garcia, Antonio Arcas, Carla Arxé, Angela Leis, Miguel-Angel Mayer



EUROPEAN OHDSI Symposium 16-20 April 2026
Barcelona

Hospital del Mar
Barcelona
Hospital del Mar
Research Institute
Barcelona



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?



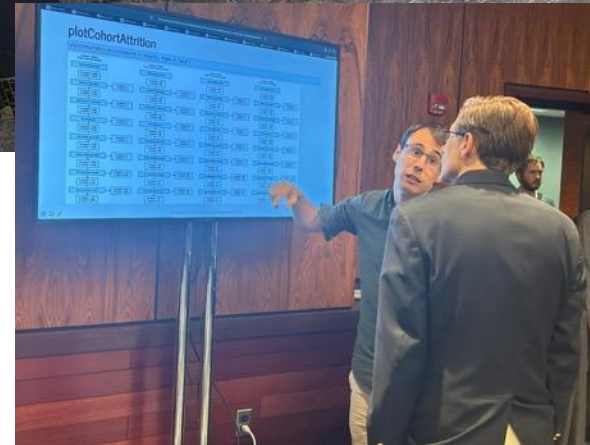
2026 OHDSI Global Symposium

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 Symposium Tutorials – Session 1

- **An Introduction to the Journey from Data to Evidence Using OHDSI**
- **An Introduction to ATLAS**
- **Bringing FAIR to Imaging Research with the Medical Imaging OMOP Extension**
- **Complex Phenotyping at Scale with and without LLMs Using PhenotypeR**
- **OHDSI Leadership Storytelling Workshop**
- **Mastering OMOP: Transforming EHR Data with Practical Strategies, Best Practices, and OHDSI Integration**



2026 Symposium Tutorials – Session 2

- **Building and Using the OHDSI Evidence Network: From Data Partner to Federated Study Execution**
- **From Multi-Modal Data to Real-World Evidence: Hands-on with the Data2Evidence Platform for OMOP Data Curation and Analytics**
- **Integrating Geospatial Data Into OMOP CDM**
- **Introduction to OHDSI Phenotype Development & Evaluation**
- **Using OMOP Model in Registry Context & Clinical Trials Standardization Context: Conventions, Past Use Cases, SDTM & Regulatory Consideration, Challenges**



2026 Global Symposium Agenda

Start	End	Topic	Presenter/Lead
7:15 am	8:00 am	Newcomer Welcome	Paul Nagy/Marc Suchard
8:00 am	8:30 am	State of the Community	George Hripcsak
8:30 am	9:15 am	OHDSI Year In Review	Early-Stage Researcher WG
9:15 am	10:00 am	Collaborator Showcase: Posters and Demos (Session 1)	
10:00 am	11:00 am	Plenary 1: Federated Learning Meets Negative Control Calibration: Toward Reliable Multi-Site Evidence Generation	Yong Chen
11:00 am	12:00 pm	Plenary 2: Beyond the Defaults: How the OHDSI Community is Adapting, Extending, and Reimagining Its Tools	Scott Duvall
12:00 pm	1:00 pm	Network & Lunch	
1:00 pm	2:00 pm	Plenary 3: The role of national initiatives in supporting sustainability, collaboration, and growth of OHDSI	Ed Burn
2:00 pm	2:45 pm	Collaborator Showcase: Lightning Talks (Session 1)	5 presenters
2:45 pm	3:30 pm	Collaborator Showcase: Posters and Demos (Session 2)	
3:30 pm	4:15 pm	Collaborator Showcase: Posters and Demos (Session 3)	
4:15 pm	5:00 pm	Collaborator Showcase: Lightning Talks (Session 2)	5 presenters
5:00 pm	6:00 pm	Titan Awards, Closing	Patrick Ryan
6:00 pm	8:30 pm	Dinner on your own	
8:30 pm	11:00 pm	OHDSI Jam Session	Martijn Schuemie



OHDSI 2026 Lightning Talks – Session 1

Evaluating Semantic Loss in EHR Terminology Mapping in Common and Rare Diseases (**Melanie Philofsky**, EPAM Systems)

Using OWL Reasoning to Surface Hidden Tensions in the OHDSI Standardized Vocabularies (**William Baumgartner**, University of Colorado Anschutz Medical Campus)

Ontology-Governed LLM Search for Safe Drug Concept Mapping in Air-Gapped Healthcare Environments (**Seung Ju Yun**, Chonnam National University Hwasun Hospital)

EPISODE/EPISODE_EVENT-Based Representation of Microbiology Results and Antimicrobial Susceptibility Testing in OMOP CDM v5.4 (**Polina Talapova**, SciForce/Tufts CTSI)

Making the most of What We Have: Regional Characterization of OHDSI Network Database (**Evelyn Goh**, National University of Singapore)



OHDSI 2026 Lightning Talks – Session 2

ORBIT: One-shot Reconstruction-Based Inference via Tensorization (**Qiaoshi Yang**, University of Pennsylvania)

Comparative performance of the concurrent comparator design with existing vaccine safety surveillance approaches on real-world observational health data (**Shounak Chattopadhyay**, University of Virginia)

Implementing Population Health Risk Analytics on a Standardized EHR Data Layer: Evaluation and Framework Development (**Haeun Lee**, Johns Hopkins University)

UNO: Neural Pruning Enhances AI Reliability of Observational Data-based Findings (**Yue Wu**, University of Pennsylvania)

A Multi-Database, Retrospective, Long-term, Observational Safety Study of Pediatric Patients with Inflammatory Bowel Disease (**Kevin Haynes**, Johnson & Johnson)



2026 Symposium Workgroup Activities

Session 1 (8 am – 10 am): Asia-Pacific Chapter; Early-Stage Researchers; Eyecare & Vision Research; Gen AI, Phenotyping, Vocabularies: AI-assisted Phenotyping Connectathon (1 of 2); Geographic Information System; HADES Hackathon (1 of 3); Health Equity (1 of 2); Industry (1 of 2); Oncology (1 of 2); Tidy R programming with OMOP (1 of 2)

Session 2 (10:30 am – 12:30 pm): Gen AI, Phenotyping, Vocabularies: AI-assisted Phenotyping Connectathon (2 of 2); HADES Hackathon (2 of 3); Health Equity (2 of 2); Industry (2 of 2); Medical Imaging; Oncology (2 of 2); Perinatal and Reproductive Health; Rare Disease; Tidy R programming with OMOP (2 of 2); Waveform

Session 3 (1:30 pm – 3:30 pm): ATLAS/WebAPI; CDM Survey; Dentistry; Evidence Network Data Partners; GIS/Waveform Cross Pollination Meeting-Extension Table Status, Integration, and Approaches; HADES Hackathon (3 of 3); Health Economics & Value Assessment; Natural Language Processing (NLP); Surgery and Perioperative Medicine; Women of OHDSI

Session 4 (3:30 pm – 5 pm): Workgroup Summary Session (with prizes!)



**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](https://www.ohdsi.org/community-calls-2026)



Find your workgroup.

Fuel our mission.

ohdsi.org/workgroups