Standardized Vocabulary 2026 Summer Refresh

OHDSI Community Call

Sept. 1, 2026 • 11 am ET

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Upcoming Community Calls

Date	Topic
Sept. 1	Vocabulary Refresh, Summer 2026 Version
Sept. 8	Medical Imaging Workgroup Spotlight, CDM 5.5 Review
Sept. 15	Global Symposium Preview
Sept. 22	TBA
Sept. 29	OHDSI/OMOP Research Spotlight
Oct. 6	DARWIN EU Progress
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

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Jennifer Hoppe, Jocelyn Lee, Tomi Akinyemiju, Danielle Bitterman, Michael Brudno, Michelle Green, Vojtech Huser, Jafi Lipson, Sanjay Mishra, Daniel Nussbaum, Jai Patel, Valentina Petkov, Kelli Rasmussen, Sahussapont Joseph Sirintrapun, Patricia Spears, Sebastiaan Van Sandijk, Anne-Marie Meyer, Umit Topaloglu, and Jeremy Warner** on the recent publication of **The GENIE Data Model: A Solid Tumor, Person-Centric Framework for Scalable, Harmonized Precision Oncology Data Collection** in *Cancer Research Communications*.

CANCER RESEARCH COMMUNICATIONS

ABOUT ⌄ JOURNAL SECTIONS ⌄ ONLINEFIRST JOURNAL ARCHIVE FOR AUTHORS ⌄ ALERTS

Article Contents

Abstract

Supplementary data

RESEARCH ARTICLE | AUGUST 27 2026

The GENIE Data Model: A Solid Tumor, Person-Centric Framework for Scalable, Harmonized Precision Oncology Data Collection

Jennifer N. Hoppe ; Jocelyn Lee ; Tomi F. Akinyemiju ; Danielle S. Bitterman ; Michael Brudno ; Michelle F. Green ; Vojtech Huser ; Jafi A. Lipson ; Sanjay Mishra ; Daniel P. Nussbaum ; Jai N. Patel ; Valentina I. Petkov ; Kelli M. Rasmussen ; Sahussapont Joseph. Sirintrapun ; Patricia A. Spears ; Sebastiaan Van Sandijk ; Anne-Marie Meyer ; Umit Topaloglu ; Jeremy L. Warner

Check for updates

+ Author & Article Information

Cancer Research Communications (2026)

https://doi.org/10.1158/2767-9764.CRC-26-0148 Article history

Split-Screen Views ⌄ PDF Share ⌄ Tools ⌄ Versions ⌄

Abstract

Adequately powered analyses in precision oncology often require combining cohorts across institutions. Yet integration is constrained by the least granular source and may become infeasible when data elements are too heterogeneous to harmonize and map to a common data model. This challenge is acute in multi-institutional precision oncology research, where real-world evidence requires harmonized clinico-omic data integration. Existing models often lack sufficient treatment patterns, outcomes, and genomic data, limiting interoperability and scalability. To address these gaps, AACR Project GENIE™ (Genomics Evidence Neoplasia Information Exchange) developed the GENIE Data Model (GDM), a comprehensive, open-source, oncology data model for scalable, consistent, and interoperable data collection across solid tumors designed to effectively capture the patient's journey with cancer. Through iterative consensus-building, four working groups comprising 13 subject matter experts defined data elements across multiple clinical domains: patient characteristics, imaging, diagnosis, surgery, histopathology, biomarkers, systemic therapy, radiation, clinical trial history, disease response and outcomes, and social determinants of health. Elements were defined using standardized terminologies and permissible values to support mapping to HL7 FHIR, OMOP, and other existing oncology standards. The model architecture distinguishes manually abstracted elements from computationally collected elements, enabling parallel workflows. The GDM provides an extensible framework that addresses critical gaps and enables scalable, harmonized data collection essential for precision oncology and real-world evidence generation.

Three Stages of The Journey

Where Have We Been?

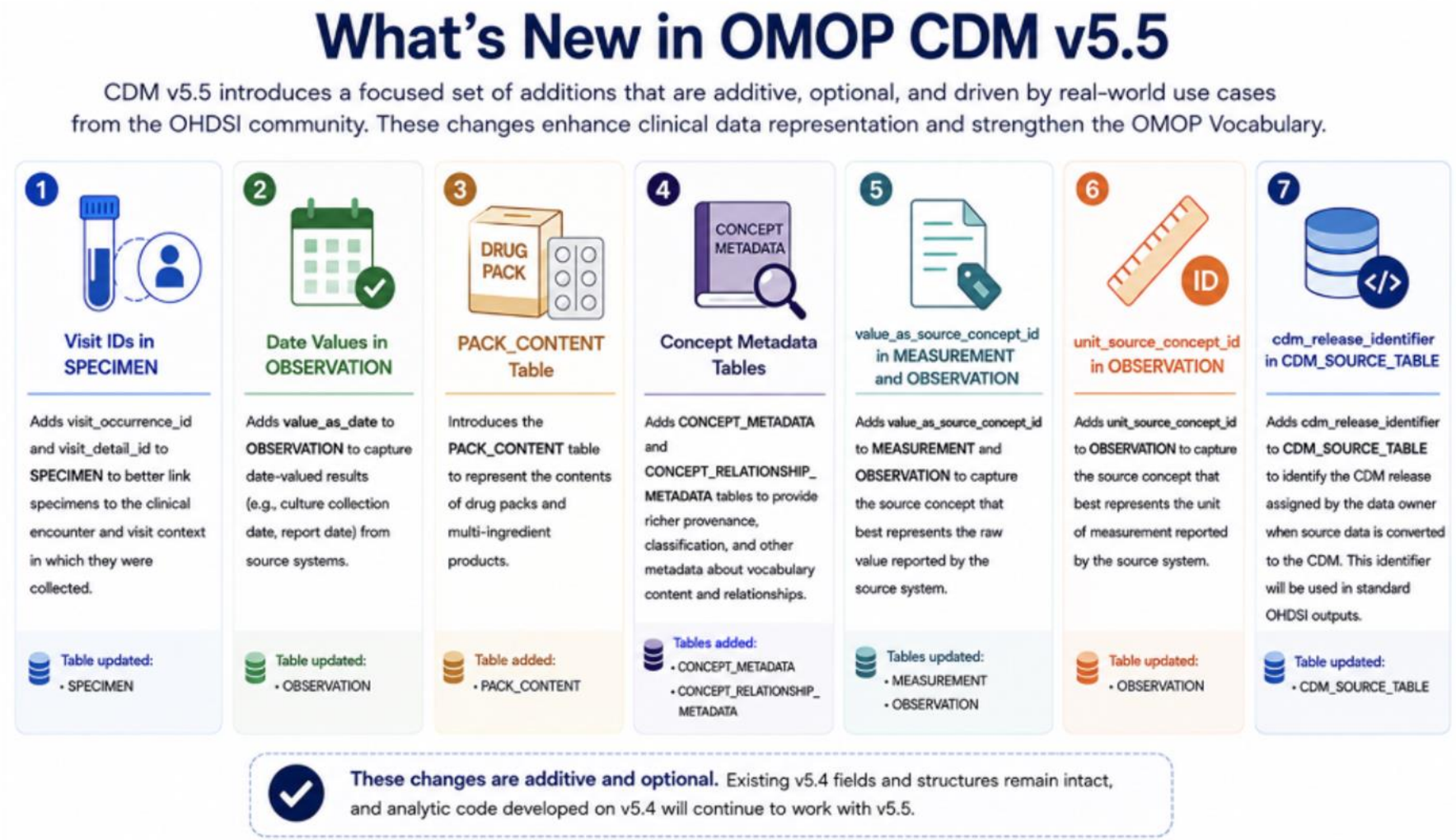
Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Wednesday	7 am	Medical Imaging
Wednesday	9 am	Tidy R Programming with OMOP
Wednesday	10 am	Common Data Model
Thursday	8 am	Medical Devices
Thursday	10 am	ATLAS/WebAPI
Thursday	10 am	GIS – Geographic Information System
Thursday	11 am	Themis
Thursday	12 pm	Methods Research
Thursday	1 pm	Oncology Vocabulary/Development Subgroup
Thursday	2 pm	Early-Stage Researchers
Thursday	2 pm	REVEAL
Thursday	7 pm	Dentistry
Friday	9 am	Waveform
Friday	10 am	Respiratory Research
Friday	10 am	Transplant
Friday	11:30 am	Steering
Tuesday	9 am	Oncology Genomic Subgroup
Tuesday	9 am	Data2Evidence

CDM v5.5 Released



September Newsletter is Available

OHDSI

The Journey Newsletter (September 2026)

The busiest months on the OHDSI calendar are officially underway, kicked off by our latest vocabulary release and an OMOP update. Between upcoming regional events, including the UK Symposium (Sept. 17–18), and next month's Global Symposium (Oct. 20-22), there is plenty happening across the community. Explore the latest news, updates, recent publications, and presentations in the September edition of the OHDSI newsletter.
#JoinTheJourney

Podcast: Regional Events, Global Plenaries

In the September 2026 On The Journey podcast, Patrick Ryan and Craig Sachson looked ahead to the final four months of the year, which will include multiple regional events, including symposia in four different continents. Then they look at the three community-developed plenary sessions that will be featured during Day 2 of the 2026 Global Symposium. *(If video does not appear, please click 'View this email in your browser')*

Community Updates

Where Have We Been?

- The Summer 2026 Vocabulary Refresh was recently shared by our Vocabulary team, and it will be presented during the Sept. 1 community call. CDM v5.5 was also shared, and more information can be found later in this newsletter.
- Oncology workgroup lead **Asieh Golozar** led an inspiring spotlight on the Oncology team's efforts to advance precision cancer research. The presentation highlighted key developments across standardized cancer vocabularies, enhanced ETL pipelines, and novel methods like the ARTEMIS algorithm for characterizing chemotherapy regimens.
- A critical aspect of the phenotype development and evaluation process is aligning on a clinical definition that provides a full specification for each of the indications, exposures, and outcomes of interest in our studies. But how good are we as a community at reaching that shared understanding? **Patrick Ryan** discussed during an interactive community call.

Where Are We Now?

- The OHDSI Titan Awards recognize those who have gone above and beyond to foster community engagement, lead research and development efforts, and make significant contributions towards OHDSI's mission. Individuals, teams or collaborating institutions can be nominated for an award, and you can make as many nominations as you would like. Nominations will be accepted until 8pm ET Tuesday, Sept. 22, 2026.
- The fourth annual UK Symposium will be held September 18 at the University of Nottingham, preceded by an OMOP Training Day on September 17. The main conference agenda is available.
- Take your OMOP CDM from implementation to impact with The OMOP Practitioner, an intensive three-day expert training hosted Sept. 7-9 by Erasmus MC in Rotterdam. This hands-on program moves beyond the basics, offering a maximum of 30 participants deep-dive experience with the Data Quality Dashboard, ATLAS, and real-world study execution.

OMOP v5.5: What's New, Should You Upgrade?

Introduced in tandem with the Summer 2026 Vocabulary Refresh, OMOP CDM v5.5 is an incremental, non-breaking update to v5.4 designed to address specific needs identified by the OHDSI community while supporting the continued evolution of the OMOP Vocabulary. Rather than introducing major structural changes, v5.5 adds targeted capabilities for representing clinical data and introduces new vocabulary tables that provide richer information about drug products, concepts, and relationships.

The result is a relatively small CDM update with benefits for data partners who need these new capabilities, and a straightforward path forward for those who do not. You can see the specific upgrades in the graphic and links below.

The recommendation of the CDM WG is to upgrade when it makes sense for your organization—there is no need to rush. For existing v5.4 implementations, v5.5 is a relatively small upgrade that provides access to the latest CDM and Vocabulary capabilities and prepares your database for future OHDSI studies. If those capabilities are not currently relevant to you, remaining on v5.4 in the near term will not disrupt your existing workflows. **In short: v5.4 will continue to work; v5.5 gives you access to what comes next.**

My Journey: Nicole Pratt

In the latest episode of our "My Journey" series, Nicole Pratt (Professor of Biostatistics and Pharmacoepidemiology at Adelaide University) shares her passion for mentorship and global teamwork. From guiding early-career researchers to watching the community grow to massive global symposiums, Nicole highlights the generosity, friendship, and multidisciplinary collaboration driving the future of evidence generation. *(If video does not appear, please click 'View this email in your browser')*

Titan Award Nominations Are Open! Highlight A Community Individual, Team by Sept. 22

August Publications

Haynes K, Reps J, Hardin J, Didden EM, Gilbert J, Swerdel J, Bohn J, Conover MM, Sarayani A, Sena AG, Yost E, van Baalen V, Davis K, Ryan P, Schuemie M. Active Safety Surveillance Using Real-World Evidence (ASSURE) Implementation: Transparent and Reproducible Real-World Evidence Standardized Framework to Support Safety Signal Evaluation. Pharmacoepidemiol Drug Saf. 2026 Aug;35(8):e70435. doi: 10.1002/pds.70435. PMID: 42494324; PMCID: PMC13397054.

Mbouamba Yankam B, Luc Baudoin FT, Onana Akoa FA, Andeso P, Cygu SB, Iddi S, Ochola M, Bhattacharjee T, Charles Lebon MO, Barasa M, Kiragga A, Mbatchou Ngahane BH. Harmonizing patient health data into the OMOP common data model: a case study using tuberculosis data from Douala General Hospital, Cameroon. Front Digit Health. 2026 Aug 3;8:1828475. doi: 10.3389/fdgth.2026.1828475. PMID: 42609725; PMCID: PMC13479808.

Kern DM, Bohn J, Gilbert JP, Knoll C, Ryan PB. REWARD-an open-source framework for identifying the unknown benefits of existing medications to inform drug discovery, development, and repurposing. J Am Med Inform Assoc. 2026 Aug 6:ocag137. doi: 10.1093/jamia/ocag137. Epub ahead of print. PMID: 42562768.

Wang-Silvanto J, Jajoo M, Guo H, Ohri R, Nelissen J, Casañas I Comabella C. Evaluating AI Performance in Systematic Literature Reviews for HEOR: A Case Study. J Health Econ Outcomes Res. 2026 Aug 7;13(2):46-54. doi: 10.36469/001c.165173. PMID: 42572706; PMCID: PMC13453151.

Sheikhalishahi S, Kaspar M, Schwinn J, Morhart M, Simon P, Hinske LC. Cross-Silo Federated Learning for Predicting Successful Mechanical Ventilation Weaning Across 5 Intensive Care Unit Databases: Retrospective Database Analysis. JMIR Med Inform. 2026 Aug 14;14:e79713. doi: 10.2196/79713. PMID: 42600153.

Lohachab A, Jung F, Rommes S, Ertaylan G, Ankolekar A, Urovi V. SMART: structured, meaningful, auditable, responsible, and transparent documentation for clinical AI. J Am Med Inform Assoc. 2026 Aug 19:ocag117. doi: 10.1093/jamia/ocag117. Epub ahead of print. PMID: 42616679.

Alcalde-Herraiz M, Delmestri A, Omulo H, Bräuner E, Bruun S, Norris R, Vivirito A, Harms A, Vuković J, Pristaš I, Jurčević A, Čavlina M, Jezidžić A, Ivanko P, Hayati S, Trinh NTH, Nordeng H, Duarte-Salles T, Palomar-Cros A, Giuliodori A, Gómez-Outes A, Vrijlandt P, Restrepo-Méndez MC, Burn E, Prats-Uribe A, Saura-Lazaro A. Characterising Clinically Recognised Hypertrophic Cardiomyopathy in six European Countries Using Real-World Data. ESC Heart Fail. 2026 Aug 20:xvag218. doi: 10.1093/eschf/xvag218. Epub ahead of print. PMID: 42619650.

mailchi.mp/ohdsi/september2026

Canada Symposium Submissions Open

Members of the OHDSI Canada community are invited to showcase their work through poster presentations at the 2026 OHDSI Canada Symposium, which takes place November 17–18 in Toronto.

Posters do not need to present completed research. Work in progress, implementation experiences, methods and tools, lessons learned, and preliminary research results are all welcome.

Submissions are due Sept. 11!

Titan Nominations Are OPEN!

The OHDSI Titan Awards recognize those who have gone above and beyond to foster community engagement, lead research and development efforts and make significant contributions towards OHDSI's mission.

Individuals, teams or collaborating institutions can be nominated for an award, and you can make as many nominations as you would like.

Nominations will be accepted until 8pm ET Tuesday, Sept. 22, 2026.

ohdsi.org/titan-awards

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
- OHDSI Standardized Vocabularies on FHIR: A Deep Dive Using the Echidna Terminology Server
- 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	11:30 pm	OHDSI Jam Session	Martijn Schuemie

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

EUROPE

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

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

Job Openings

Associate Director, Data Science (Real World Data)

Apply

Boston, MA; New York, NY

About the Position

As Associate Director of RWD Intelligence at Formation Bio, you will lead the strategy and execution of our real-world data (RWD) capabilities, building the data foundations that power drug acquisition, clinical development, and portfolio decision-making. You will own the end-to-end lifecycle of RWD: sourcing, procurement, ingestion, harmonization, quality assurance, and delivery of analysis-ready datasets to downstream consumers across Product, Data Science, Clinical Development, and Business Development.

This role sits at the intersection of data engineering, data science, and drug development. You will build and maintain scalable data infrastructure (pipelines, data models, lakes/marts) while ensuring semantic interoperability across heterogeneous data sources through ontology-driven harmonization frameworks such as OMOP. You will also manage vendor relationships and data procurement, evaluating and integrating new data assets as the portfolio evolves. The ideal candidate combines deep RWD domain expertise with strong data fluency, enabling Formation Bio to treat real-world evidence as a first-class strategic asset.

Responsibilities

- Lead the RWD Intelligence function within Data Science, owning data strategy, sourcing, and delivery of analysis-ready datasets
- Architect and maintain the supporting infrastructure (pipelines, ingestion workflows, data models, lakes/marts) across EHR/EMR, claims, registries, and genomics-linked cohorts
- Drive adoption and extension of harmonization frameworks (e.g., OMOP CDM) across heterogeneous data sources, leveraging AI/ML tools for entity resolution, ontology mapping, data quality monitoring, and automated harmonization
- Manage RWD vendor relationships end-to-end: evaluate providers, negotiate data use agreements, broker new partnerships, and integrate acquired datasets into the platform
- Partner with Data Science, Clinical Development, Business Development, and Engineering teams to define RWD use cases (trial feasibility, synthetic control arms, epidemiology, label expansion) and productize ad hoc pipelines into scalable, production-grade systems
- Foster a culture of data quality rigor, documentation, and reproducibility across all RWD assets

Senior Data Scientist - Real World Data

Apply

New York, NY; Boston, MA; San Francisco, CA; Raleigh-Durham, NC

About the Position

We're looking for a Senior Data Scientist to join the Scientific Data Intelligence (SDI) team at Formation Bio to help transform Real World Data (RWD)—spanning electronic health records, claims, and other longitudinal patient data sources—into structured, analytics-ready assets. In this role, you'll be partnering closely with our Data Science team not only to model and transform data, but also to actively analyze it: answering research questions, generating evidence, and supporting scientific decision-making across our drug portfolio.

This position sits at the intersection of healthcare data engineering, real-world evidence analysis, and generative AI. While a strong foundation in building reliable, scalable pipelines is essential, you'll be equally expected to roll up your sleeves and work directly with the data—constructing cohorts, running analyses, and translating findings into actionable insights for scientific and business stakeholders.

The ideal candidate is a hybrid of data engineer and applied scientist: someone who can build the infrastructure and then use it, with familiarity in RWD study design, GenAI fluency (e.g., LLM-based entity extraction, summarization, classification), and strong technical expertise with modern data tooling. You'll play a key role in shaping how real-world patient data becomes discoverable, structured, and impactful across the organization.

Responsibilities

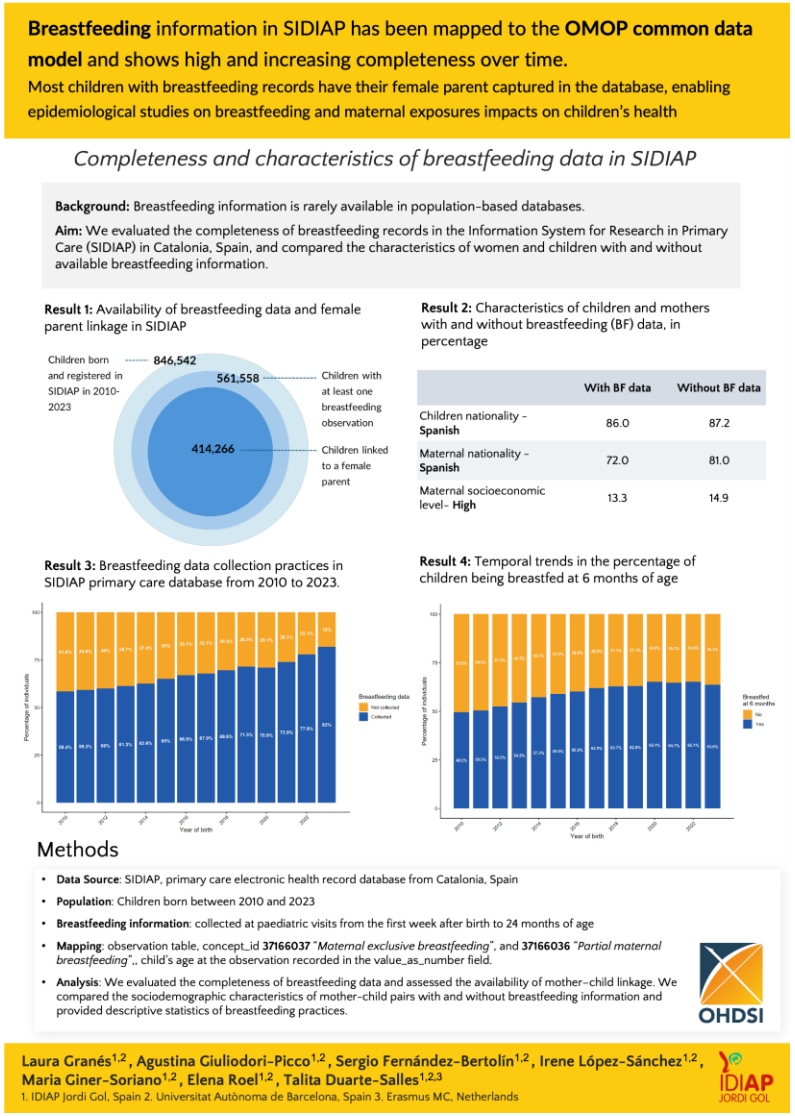
- Model and transform raw EHR and claims data into clean, canonical, and analytics-ready datasets using SQL, Python, and clinical standards like OMOP.
- Build and manage scalable data pipelines using Dagster for orchestration, dbt for transformation, and Snowflake as the primary compute and storage engine.
- Conduct hands-on RWD analyses to answer scientific and strategic research questions—including disease epidemiology, treatment patterns, patient journey characterization, and comparative effectiveness.
- Partner with Data Scientists and clinical leads to design and execute observational studies, translating scientific questions into well-structured, reproducible analyses.
- Implement data validation, completeness, and observability frameworks to ensure real-world datasets are accurate, comprehensive, and trustworthy for downstream research and product use.
- Apply Generative AI techniques within transformation and analysis layers to accelerate data structuring and insight generation.
- Communicate findings clearly to both technical and non-technical stakeholders, including summaries for portfolio teams and leadership.

#OHDSISocialShowcase This Week

Monday

Completeness and characteristics of breastfeeding data in SIDIAP

(Laura Granés, Agustina Giuliodori-Picco, Sergio Fernández-Bertolín, Irene López-Sánchez, Maria Giner-Soriano, Elena Roel Talita Duarte-Salles)



Breastfeeding information in SIDIAP has been mapped to the **OMOP common data model** and shows high and increasing completeness over time.

Most children with breastfeeding records have their female parent captured in the database, enabling epidemiological studies on breastfeeding and maternal exposures impacts on children's health

Completeness and characteristics of breastfeeding data in SIDIAP

Background: Breastfeeding information is rarely available in population-based databases.

Aim: We evaluated the completeness of breastfeeding records in the Information System for Research in Primary Care (SIDIAP) in Catalonia, Spain, and compared the characteristics of women and children with and without available breastfeeding information.

Result 1: Availability of breastfeeding data and female parent linkage in SIDIAP

Result 2: Characteristics of children and mothers with and without breastfeeding (BF) data, in percentage

	With BF data	Without BF data
Children nationality - **Spanish**	86.0	87.2
Maternal nationality - **Spanish**	72.0	81.0
Maternal socioeconomic level- **High**	13.3	14.9

Result 3: Breastfeeding data collection practices in SIDIAP primary care database from 2010 to 2023.

Result 4: Temporal trends in the percentage of children being breastfed at 6 months of age

Methods

- **Data Source**: SIDIAP, primary care electronic health record database from Catalonia, Spain
- **Population**: Children born between 2010 and 2023
- **Breastfeeding information**: collected at paediatric visits from the first week after birth to 24 months of age
- **Mapping**: observation table, concept_id **37166037** "*Maternal exclusive breastfeeding*", and **37166036** "*Partial maternal breastfeeding*"., child's age at the observation recorded in the value_as_number field.
- **Analysis**: We evaluated the completeness of breastfeeding data and assessed the availability of mother-child linkage. We compared the sociodemographic characteristics of mother-child pairs with and without breastfeeding information and provided descriptive statistics of breastfeeding practices.

OHDSI

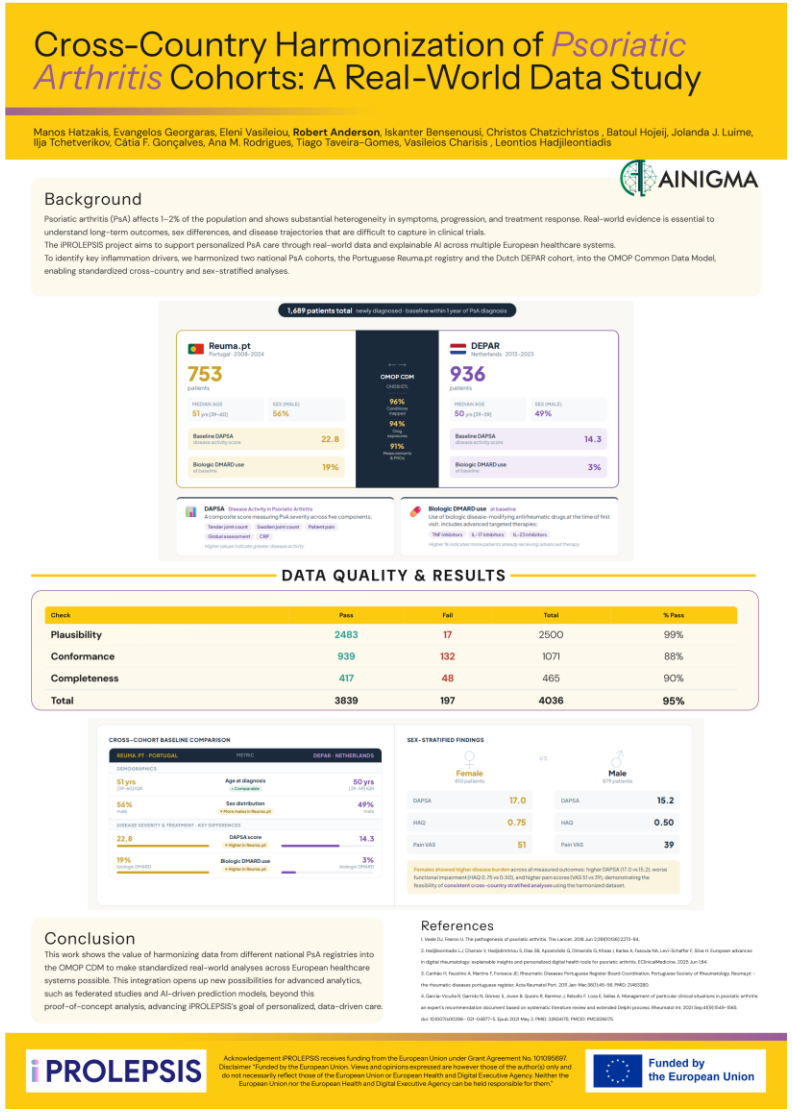
Laura Granés[1,2], Agustina Giuliodori-Picco[1,2], Sergio Fernández-Bertolín[1,2], Irene López-Sánchez[1,2], Maria Giner-Soriano[1,2], Elena Roel[1,2], Talita Duarte-Salles[1,2,3]
1. IDIAP Jordi Gol, Spain 2. Universitat Autònoma de Barcelona, Spain 3. Erasmus MC, Netherlands
IDIAP JORDI GOL

#OHDSISocialShowcase This Week

Tuesday

Cross-Country Harmonization of Psoriatic Arthritis Cohorts: A real-world data study

(Manos Hatzakis, Evangelos Georgaras, Eleni Vasileiou, Robert Anderson, Iskanter Bensenousi, Christos Chatzichristos, Batoul Hojeij, Jolanda J. Luime, Ilja Tchetverikov, Cátia F. Gonçalves, Ana M. Rodrigues, Tiago Taveira-Gomes, Vasileios Charisis, Leontios Hadjileontiadis)

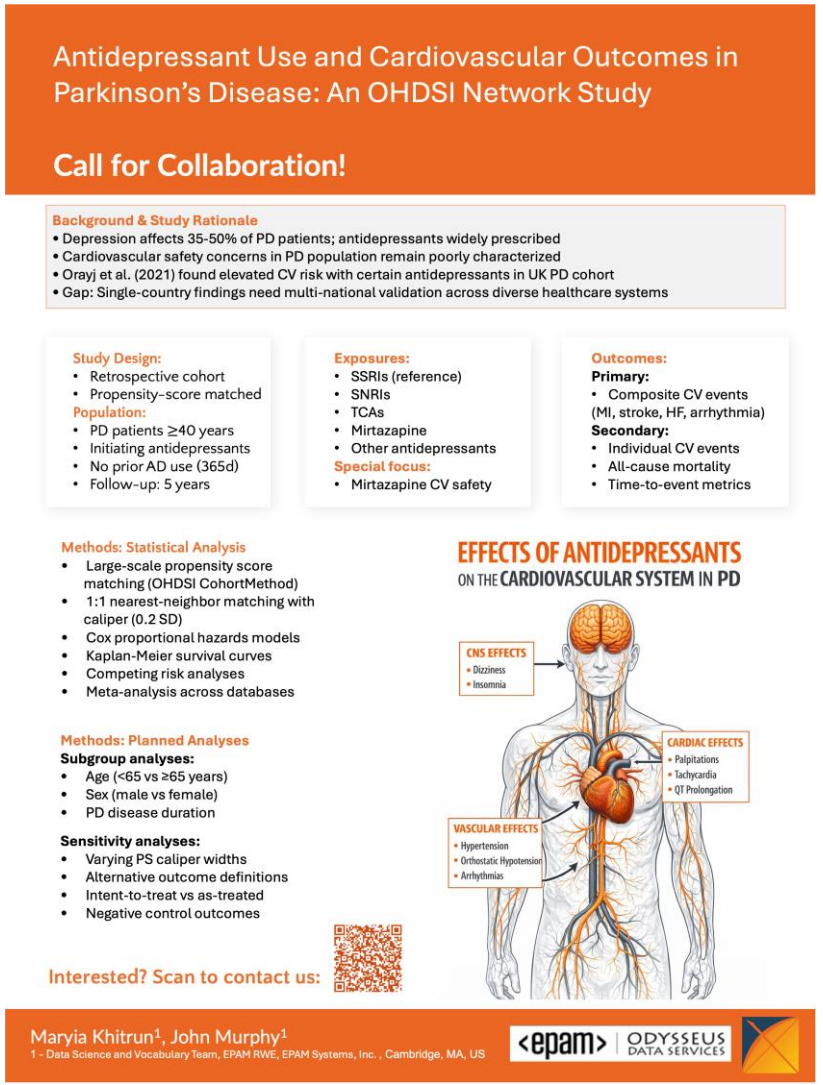


#OHDSISocialShowcase This Week

Wednesday

Antidepressant Use and Cardiovascular Outcomes in Parkinson's Disease: An OHDSI Network Study

(Maria Khitrun, John Murphy)



#OHDSISocialShowcase This Week

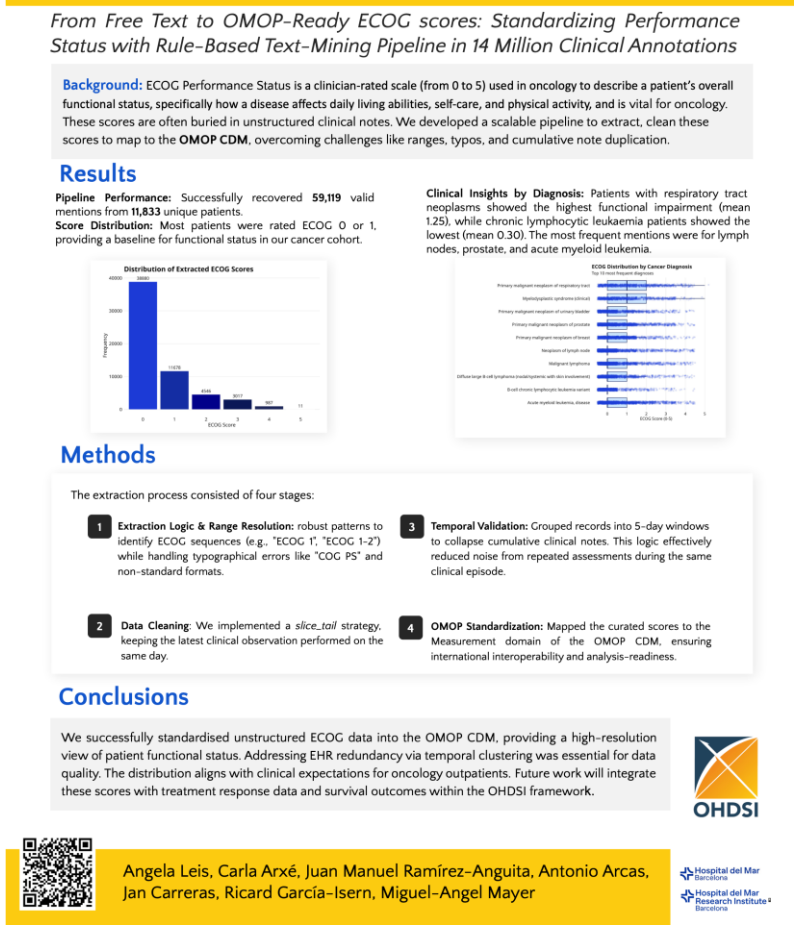
Thursday

From Free Text to OMOP-Ready ECOG scores: Standardizing Performance Status with Rule-Based Text-Mining Pipeline in 14 Million Clinical Annotations

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

From Unstructured Clinical Notes to Actionable Insights: A Workflow for Standardising Critical **ECOG Performance Status for OMOP-Based** Real-World Oncology Research.

Unlocking the Patient's Functional Trajectory: **Precision Rule-Based Text-Mining** and Analysis-Ready Data on Vital Status for 11,833 Patients.



From Free Text to OMOP-Ready ECOG scores: Standardizing Performance Status with Rule-Based Text-Mining Pipeline in 14 Million Clinical Annotations

Background: ECOG Performance Status is a clinician-rated scale (from 0 to 5) used in oncology to describe a patient's overall functional status, specifically how a disease affects daily living abilities, self-care, and physical activity, and is vital for oncology. These scores are often buried in unstructured clinical notes. We developed a scalable pipeline to extract, clean these scores to map to the **OMOP CDM**, overcoming challenges like ranges, typos, and cumulative note duplication.

Results

Pipeline Performance: Successfully recovered **59,119** valid mentions from **11,833** unique patients.
Score Distribution: Most patients were rated ECOG 0 or 1, providing a baseline for functional status in our cancer cohort.

Clinical Insights by Diagnosis: Patients with respiratory tract neoplasms showed the highest functional impairment (mean 1.25), while chronic lymphocytic leukaemia patients showed the lowest (mean 0.30). The most frequent mentions were for lymph nodes, prostate, and acute myeloid leukemia.

Methods

The extraction process consisted of four stages:

1. **Extraction Logic & Range Resolution:** robust patterns to identify ECOG sequences (e.g., "ECOG 1", "ECOG 1-2") while handling typographical errors like "COG PS" and non-standard formats.
2. **Data Cleaning:** We implemented a *slice_tail* strategy, keeping the latest clinical observation performed on the same day.
3. **Temporal Validation:** Grouped records into 5-day windows to collapse cumulative clinical notes. This logic effectively reduced noise from repeated assessments during the same clinical episode.
4. **OMOP Standardization:** Mapped the curated scores to the Measurement domain of the OMOP CDM, ensuring international interoperability and analysis-readiness.

Conclusions

We successfully standardised unstructured ECOG data into the OMOP CDM, providing a high-resolution view of patient functional status. Addressing EHR redundancy via temporal clustering was essential for data quality. The distribution aligns with clinical expectations for oncology outpatients. Future work will integrate these scores with treatment response data and survival outcomes within the OHDSI framework.

OHDSI

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

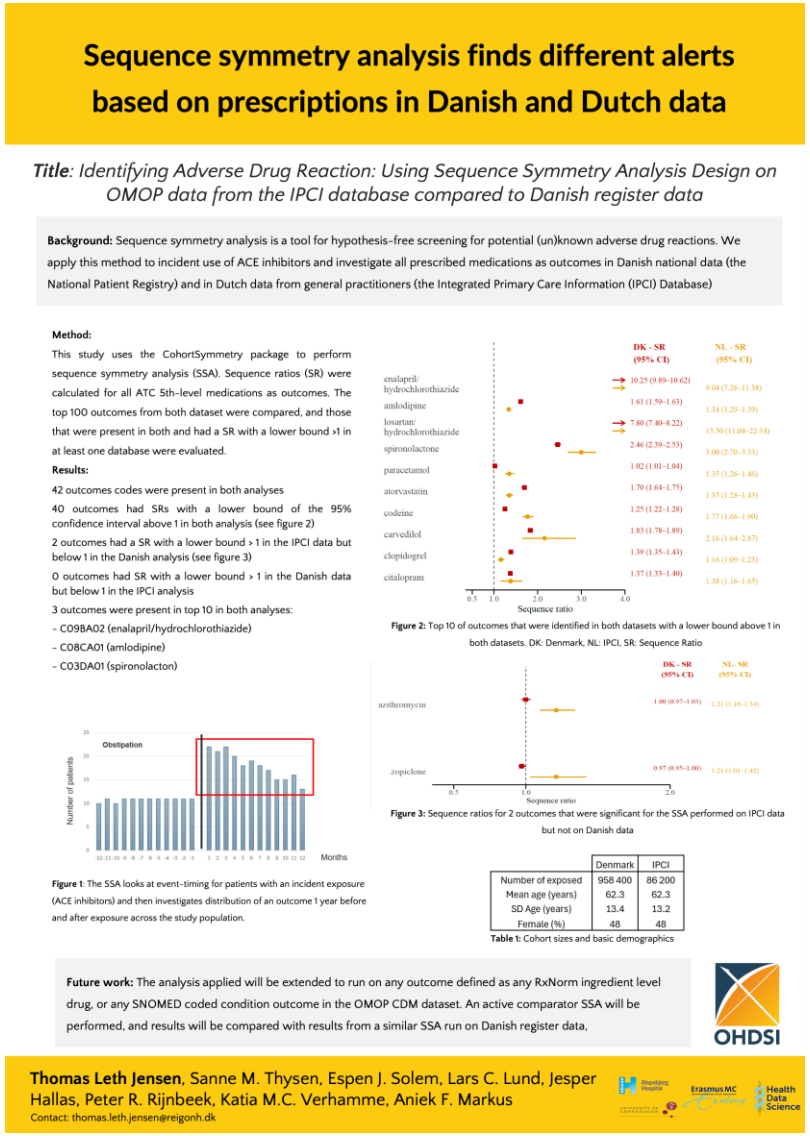
Hospital del Mar Barcelona
Hospital del Mar Research Institute Barcelona

#OHDSISocialShowcase This Week

Friday

Identifying Adverse Drug Reaction Using Sequence Symmetry Analysis Design on OMOP data

(Thomas Leth Jensen, Sanne Møller Thysen, Espen Jimenez Solem, Lars Christian Lund, Jesper Hallas, Peter R. Rijnbeek, Katia M.C. Verhamme, Aniek F. Markus)



Sequence symmetry analysis finds different alerts based on prescriptions in Danish and Dutch data

***Title**: Identifying Adverse Drug Reaction: Using Sequence Symmetry Analysis Design on OMOP data from the IPCI database compared to Danish register data*

Background: Sequence symmetry analysis is a tool for hypothesis-free screening for potential (un)known adverse drug reactions. We apply this method to incident use of ACE inhibitors and investigate all prescribed medications as outcomes in Danish national data (the National Patient Registry) and in Dutch data from general practitioners (the Integrated Primary Care Information (IPCI) Database)

Method:

This study uses the CohortSymmetry package to perform sequence symmetry analysis (SSA). Sequence ratios (SR) were calculated for all ATC 5th-level medications as outcomes. The top 100 outcomes from both dataset were compared, and those that were present in both and had a SR with a lower bound >1 in at least one database were evaluated.

Results:

42 outcomes codes were present in both analyses

40 outcomes had SRs with a lower bound of the 95% confidence interval above 1 in both analysis (see figure 2)

2 outcomes had a SR with a lower bound > 1 in the IPCI data but below 1 in the Danish analysis (see figure 3)

0 outcomes had SR with a lower bound > 1 in the Danish data but below 1 in the IPCI analysis

3 outcomes were present in top 10 in both analyses:

- C09BA02 (enalapril/hydrochlorothiazide)
- C08CA01 (amlodipine)
- C03DA01 (spironolacton)

Figure 1: The SSA looks at event-timing for patients with an incident exposure (ACE inhibitors) and then investigates distribution of an outcome 1 year before and after exposure across the study population.

Figure 2: Top 10 of outcomes that were identified in both datasets with a lower bound above 1 in both datasets. DK: Denmark, NL: IPCI, SR: Sequence Ratio

Figure 3: Sequence ratios for 2 outcomes that were significant for the SSA performed on IPCI data but not on Danish data

	Denmark	IPCI
Number of exposed	958 400	86 200
Mean age (years)	62.3	62.3
SD Age (years)	13.4	13.2
Female (%)	48	48

Table 1: Cohort sizes and basic demographics

Future work: The analysis applied will be extended to run on any outcome defined as any RxNorm ingredient level drug, or any SNOMED coded condition outcome in the OMOP CDM dataset. An active comparator SSA will be performed, and results will be compared with results from a similar SSA run on Danish register data.

OHDSI

Thomas Leth Jensen, Sanne M. Thysen, Espen J. Solem, Lars C. Lund, Jesper Hallas, Peter R. Rijnbeek, Katia M.C. Verhamme, Aniek F. Markus
Contact: thomas.leth.jensen@reigonh.dk

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?

Sept. 1: Summer Vocabulary Refresh

Anna Ostropolets

Associate Director, Johnson & Johnson;
Adjunct Associate Professor, Columbia University

Vlad Korsik

Head of Research & Development, Omophoria

Christian Reich

Chief Scientific Officer, Nemesis Health

Masha Khitrun

Lead Scientific Curation Specialist, EPAM Systems

Highlights:

- contributions from the Oncology Working Group
- CDM 5.5 support
- enhancements to rare disease vocabularies
- RxNorm Extension maintenance

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