



DARWIN EU[®] Initiative Annual Report

OHDSI Community Call
Nov. 18, 2025 • 11 am ET



Upcoming Community Calls

Date	Topic
Nov. 18	DARWIN EU 2025 Update
Nov. 25	Early-Stage Researcher Presentations
Dec. 2	OHDSI/OMOP Research Spotlight
Dec. 9	How Did OHDSI Do This Year?
Dec. 16	Holiday Farewell To 2025



Nov. 25: Early-Stage Researcher Workgroup and Presentations



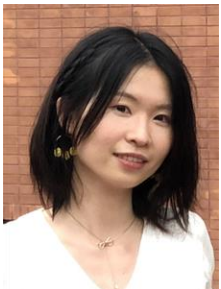
Sumin Lee and Kyulee Jeon, Yonsei College of Medicine

ARKE: An Ontology-Driven Framework for Standardizing Radiology Procedure Terminology Using LLMs and RAG



Markian Hromiak and Jacob Zelko, George Institute of Technology

Exploring Efficient and Scalable OMOP CDM Workflows by Leveraging dbt-synthea



Bingyu Zhang, Univ. of Pennsylvania

The Fine Art of Tolerance: Robustify p-value Calibration in Observational Studies with Partially Valid Negative Control Outcomes





Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?



OHDSI Shoutouts!



Congratulations to the team of **Berta Cuyàs, Edilmar Alvarado-Tapias, Eng Hooi Tan, Asieh Golozar, Talita Duarte-Salles, Antonella Delmestri, Josepmaria Argemi, Wai Yi Man, Edward Burn, Carlos Guarner-Argente, Daniel Prieto Alhambra, and Danielle Newby** on the publication of **Trends in incidence, prevalence, and survival of primary liver cancer in the United Kingdom (2000–2021)** in the *European Journal of Public Health*.

European Journal of Public Health, 2025, ckaf153

© The Author(s) 2025. Published by Oxford University Press on behalf of the European Public Health Association.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
<https://doi.org/10.1093/eurpub/ckaf153>

Trends in incidence, prevalence, and survival of primary liver cancer in the United Kingdom (2000–2021)

Berta Cuyàs^{1,2,3,†}, Edilmar Alvarado-Tapias^{1,2,3,†}, Eng Hooi Tan⁴, Asieh Golozar^{5,6}, Talita Duarte-Salles^{7,8}, Antonella Delmestri⁴, Josepmaria Argemi^{3,9,10}, Wai Yi Man⁴, Edward Burn⁴, Carlos Guarner-Argente¹, Daniel Prieto Alhambra^{4,8,*}, Danielle Newby⁴

¹Department of Gastroenterology, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain

²Medicine Department, Autonomous University of Barcelona (UAB), Barcelona, Spain

³Centre for Biomedical Research in Liver and Digestive Diseases Network (CIBERhd), Instituto de Salud Carlos III, Madrid, Spain

⁴Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

⁵Nemesis Health, New York, NY, United States

⁶OHDSI Center at the Roux Institute, Northeastern University, Boston, MA, United States

⁷Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

⁸Department of Medical Informatics, Erasmus University Medical Centre, Rotterdam, The Netherlands

⁹Liver Unit, Clínica Universidad de Navarra, DNA & RNA Medicine Program, CIMA University of Navarra, Pamplona, Spain

¹⁰Division of Gastroenterology Hepatology and Nutrition, University of Pittsburgh, Pittsburgh, PA, United States

*Corresponding author. Botnar Research Centre, Windmill Road, OX37LD, Oxford, United Kingdom. E-mail: daniel.prietoalhambra@ndorms.ox.ac.uk.

[†]joint first authors.

Abstract

Primary liver cancer (PLC) remains a global health challenge. Understanding trends in the disease burden and survival is crucial to inform decisions regarding screening, prevention, and treatment. Population-based cohort study using UK primary care data from the Clinical Practice Research Datalink (CPRD) GOLD (2000–2021), replicated in CPRD Aurum. Crude and age-standardized incidence rates (IRs), crude period prevalence (PP), and survival at 1, 5, and 10 years were calculated, and stratified by age, sex, and diagnosis year. The crude IR of PLC was 4.56 (95% CI 4.42–4.70) per 100 000 person-years between 2000 and 2021, with an increase over time across age and sex strata. Sex-specific IR for males was higher than females, 6.60 (95% CI 6.36–6.85) vs. 2.58 (95% CI 2.44–2.74) per 100 000 person-years. Age-standardized IR showed identical trends. Crude PP showed a seven-fold increase over the study period, with PP 0.02% (95% CI 0.019%–0.022%) in 2021, and a 2.8-fold higher PP in males. Survival at 1, 5, and 10 years after diagnosis was 41.7%, 13.2%, and 7.1%, respectively, for both sexes. One-year survival increased only in men, from 33.2% in 2005–2009 to 49.3% in 2015–2019. Over the past two decades, there has been a substantial increase in the number of patients diagnosed with PLC. Despite a slight improvement in median and one-year survival in men, prognosis remains poor. To improve the survival of PLC patients, it is necessary to understand the epidemiological changes and address preventable risk factors associated with liver disease and promote early detection and access to care.



OHDSI Shoutouts!



Congratulations to the team of
Melissa Finster, Markus Wenzel
and Elham Taghizadeh on the
publication of **Common data
models and data standards for
tabular health data: a systematic
review** in *BMC Medical
Informatics and Decision Making*.

Finster et al. *BMC Medical Informatics and Decision Making* (2025) 25:422
<https://doi.org/10.1186/s12911-025-03267-2>

BMC Medical Informatics
and Decision Making

SYSTEMATIC REVIEW

Open Access

Common data models and data standards for tabular health data: a systematic review



Melissa Finster^{1*}, Markus Wenzel^{1,2†} and Elham Taghizadeh^{1†}

Abstract

Background The use of health data supports knowledge-based decision-making in healthcare. Common Data Models (CDMs) and data standards facilitate the integration of diverse data sources and enable federated analysis by harmonizing data formats and terminologies.

Methods To determine the best approaches to harmonizing patient data, we undertook a comprehensive literature search, which allowed us to identify the most popular and established CDMs (i2b2, Sentinel CDM, PCORnet CDM, OMOP CDM) and data standards (CDA, HL7 version 2, FHIR, openEHR). We established a set of criteria across the categories of Suitability, Popularity, Adaptability, Interoperability, and Support.

Results The CDMs and data standards are evaluated based on the defined criteria. Overall criteria the OMOP CDM and FHIR scored best. We highlight the strongest CDM and data standard for each criteria category.

Conclusion Given the unique characteristics, strengths, and weaknesses of each CDM and data standard, no single global representation can be selected. To promote broad adoption of CDMs and data standards, it is essential to enable transformation between different representations and utilize various formats within a single tool to facilitate their interoperability. Only then seamless data exchange and research across borders can be achieved.

Clinical trial number Not applicable.

Keywords Common data model, FAIR-Principles, Data standard, Interoperability, Data harmonization



Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?



Upcoming Workgroup Calls



Date	Time (ET)	Meeting
Tuesday	12 pm	ATLAS/WebAPI
Wednesday	1 pm	Perinatal and Reproductive Health
Thursday	8 am	India Community Call
Thursday	9 am	Oncology Vocabulary/Development Subgroup
Thursday	11 am	Themis
Thursday	12 pm	HADES
Friday	10 am	GIS-Geographic Information System
Friday	10:30 am	Open-Source Community
Friday	11:30 am	Steering
Monday	9 am	Africa Chapter
Monday	10 am	Getting Started Subgroup
Tuesday	10 am	CDM Survey



OHDSI Africa Symposium 2025

Nov 10-12 Kampala, Uganda





Day 1 Tutorial @ JCRC

Session Name	Time	Instructor(s)
OHDSI/OMOP Intro	9:00 – 9:20 AM	Michael Ochola Mui Van Zandt Cynthia Sung
OMOP CDM and Vocabulary	9:20 – 10:00 AM	Sebastian van Sandijk Cynthia Sung Mui Van Zandt
OMOP Conversion Process	10:00 – 10:30 AM	Rachel Odhiambo Freija Descamps Ousmane Diop
<i>Break</i>	10:30 – 10:40 AM	
ETL Exercises	10:40 – 12:00 noon	Pauline Andeso Freija Descamps Reinpeter Momanyi
<i>Lunch</i>	12:00 – 1:30 PM	
Data Quality Dashboard	1:30 – 2:00 PM	Reinpeter Momanyi David Amadi Sebastian van Sandijk
Evidence Generation with OHDSI Tools	2:00 – 3:00 PM 3:15 – 4:15 PM	Edward Burns Anna Saura-Lazaro Mui van Zandt



Instructors



Mui van Zandt
IQVIA



Michael Ochola
APHRC



Cynthia Sung
Duke-NUS



Sebastian van Sandijk
EPAM



Rachel Odhiambo
APHRC



David Amadi
LSHTM & APHRC



Freija Descamps
EdenceHealth



Ousmane Diop
IRESSEF



Pauline Andeso
APHRC



Ed Burn
Oxford



Reinpeter Momanyi
APHRC



Anna Saura Lazaro
Oxford



Tutorial Group Photo @ JCRC





Day 2-3 Symposium @ Mestil Hotel





Day 2-3 Symposium @ Mestil Hotel



Dr. Cissy Kityo
Executive Director
Joint Clinical Research Center

Paul Mbaka, Asst Commissioner
Dept Health Information, MOH



Brenda Nakazibwe
Sci, Tech & Innovation Secretariat
Office of the President, Uganda



2025 OHDSI Africa Symposium Group Photo





Next OHDSI Africa Symposium



Uganda

© Vemaps.com



Ethiopia

© Vemaps.com



India Symposium



The graphic features a central blue hexagon with a white border, containing the OHDSI logo and the text "OHDSI INDIA SYMPOSIUM 2025". Below this, a white rounded rectangle contains the date "TUESDAY 02 DECEMBER 2025" and the location "SVM HOSPITALS, BANGALORE". A blue banner at the bottom contains the text "We're gearing up for OHDSI Symposium 2025 — where data meets discovery! Be part of a global network driving open science in healthcare !" and a "Register Now" button. A small globe icon and the text "For More Information : contact@ohdsi-india.org" are at the very bottom.

OHDSI INDIA SYMPOSIUM 2025

TUESDAY 02 DECEMBER 2025 | SVM HOSPITALS, BANGALORE

We're gearing up for OHDSI Symposium 2025 — where data meets discovery! Be part of a global network driving open science in healthcare !

[Register Now](#)

For More Information : contact@ohdsi-india.org





APAC Symposium: Dec. 6-7

Day 1 (December 6) – Tutorial at Room 102, Dongxia Yuan Building (Zheng-Cai Cuiju Teaching Building)

Morning Session

- 09:00-09:20 Introduction of OHDSI/OMOP
- 09:20-10:00 OMOP CDM and Vocabulary
- 10:00-10:30 OMOP Conversion Process
- 10:40-12:00 ETL Exercises

Afternoon Session

- 13:30-14:50 OHDSI Analyses: Building Cohorts & Hands-on
- 14:50-15:30 CohortDiagnostics and Population-Level Estimation
- 15:50-16:30 Interpreting Results

Day 2 (December 7) – Main conference at Room A100, 1F, Student Center

Session 1 – From Global to Regional Impact: OHDSI across APAC & Africa

- 09:00 – 09:15 Opening Speech
- 09:15 – 09:45 Keynote Speech from OHDSI Global
- 09:45 – 10:45 APAC Regional Chapter Updates
- 10:45 – 11:00 OHDSI Africa

Day 2 (December 7) – Main conference at Room A100, 1F, Student Center

Session 1 – From Global to Regional Impact: OHDSI across APAC & Africa

- 09:00 – 09:15 Opening Speech
- 09:15 – 09:45 Keynote Speech from OHDSI Global
- 09:45 – 10:45 APAC Regional Chapter Updates
- 10:45 – 11:00 OHDSI Africa

Session 2 – From Research to Reflection: 2025 APAC Studies and Lessons Learned

- 11:15 – 11:30 2025 APAC Study 1 by Fudan University
- 11:30 – 11:45 2025 APAC Study 2 by Peking University
- 11:45 – 12:00 2025 APAC Study 3 by University of Science and Technology of China (USTC)
- 12:00 – 12:10 Journal's Perspectives
- 12:10 – 12:30 Panel Discussion

Session 3 – From Regional Insights to Local Challenges: Real-World Evidence and OHDSI/OMOP in China

- 13:30 – 14:30 Collaborator Showcase: Lightning Talks
- 14:30 – 14:45 Real-World Evidence Talk 1
- 14:45 – 15:00 Real-World Evidence Talk 2
- 15:00 – 15:15 Real-World Evidence Talk 3
- 15:30 – 15:50 Real-World Evidence Using OHDSI/OMOP
- 15:50 – 16:10 Panel Discussion: Opportunities and Challenges Using OHDSI/OMOP for Real-World Evidence in China
- 16:10 – 16:50 Closing & Networking



ohdsi.org/apac2025



2026 Global Symposium

The 2026 OHDSI Global Symposium will return to the Hyatt Regency Hotel in New Brunswick, N.J., on **Oct. 20-22.**

More details to come.





Tutorials Homepage

OHDSI Tutorials

Education is at the heart of OHDSI's mission, and these tutorials showcase the community's commitment to sharing knowledge. Developed and taught by OHDSI faculty, they highlight tools, standards, and best practices that empower collaborators at every level to engage in open science and generate reliable evidence.

2025 Global Symposium

An Introduction to the Journey from Data to Evidence Using OHDSI

Introduction to OHDSI

OHDSI2025 Tutorial: An Introduction to the Journey from Data to Evidence Using OHDSI

Observational Health Data Sciences and Informatics (OHDSI, pronounced "Odyssey") was founded in 2014.

- Central coordinating center housed at Columbia University.
- A multi-stakeholder, interdisciplinary evidence collaborative to bring out the value of health data for large-scale analytics.

Faculty: Erica Voss, Yong Chen, Katy Sadowski, Nicole Pratt, Roger Carlson, Chongliang (Jason) Luo

Developing and Evaluating Your Extract, Transform, Load (ETL) Process to the OMOP Common Data Model

OMOP Common Vocabulary Model

reference data for the OMOP CDM

What it is

- Compiled standards from disparate public and private sources and some OMOP-grown concepts
- Standardized structure to house existing vocabularies used in the public domain

What it's not

- Static dataset - the vocabulary updates regularly to keep up with the continual evolution of the sources
- Product - vocabulary and improvement is ongoing activity that requires community participation and support

Faculty: Clair Blacketer, Karthik Natarajan, Evanette Burrows, Max Adulyanuksol, Maxim Molnat

Using the OHDSI Standardized Vocabularies for Research

OHDSI Standardized Vocabularies

OHDSI2025 Tutorial: Using the OHDSI Standardized Vocabularies for Research

Data in US after 2005 ICD-10

Data in US after 2005 ICD-9

Data in UK after 2005 ICD-9

Data in US after 2005 ICD-10

Common reference standard: SNOMED

Faculty: Anna Ostropelets, Vlad Korsik, Polina Talapova, Masha Khitrin

Population-Level Effect Estimation Applications to Generate Reliable Real-World Evidence

OHDSI2025 Tutorial: Population-Level Effect Estimation Applications to Generate Reliable Real-World Evidence

Population-Level Effect Estimation Applications to Generate Reliable Real-World Evidence

George Hripsak
Martijn Schuemie
Linying Zhang
Tara Anand

Faculty: George Hripsak, Martijn Schuemie, Linying Zhang, Tara Anand

Clinical Characterization Applications to Generate Reliable Real-World Evidence

OHDSI2025 Tutorial: Clinical Characterization Applications to Generate Reliable Real-World Evidence

Complementary evidence to inform the patient

Clinical characterization: What happened to me?

Patient profile: What is it like to be me?

Population-level effect estimation: What is the causal effects?

Faculty: Patrick Ryan, Aniek Markus, Hsin Yi "Cindy" Chen, Azza Shoaibi

Patient-Level Prediction Applications to Generate Reliable Real-World Evidence

OHDSI2025 Tutorial: Patient-Level Prediction Applications to Generate Reliable Real-World Evidence

Prediction Problem Definition

Observation Window

Time-at-risk

outcome

Faculty: Jenna Reys, Egill Fridgerisson, Ross Williams

ohdsi.org/tutorials

www.ohdsi.org

#JoinTheJourney





#OHDSISocialShowcase This Week

Monday

PREPARE-Rehab: Personalized rehabilitation via novel AI patient stratification strategies using the OMOP-CDM standard

(**Carlotte Kiekens**, Esther Janssen, Lisa Hoogendam, Ruud Selles, Philip van der Wees, Alberto Negrini, Stefano Negrini and the PREPARE group)

PREPARE-Rehab helps establish best practices for applying OHDSI tools in **rehabilitation**, and the **extension of standardized vocabularies and CDM domains** to increase its usefulness for applications in **biopsychosocial domains**

Personalized rehabilitation via novel AI patient stratification strategies using the OMOP-CDM standard

- **Rehabilitation** is a complex, multimodal, collaborative person-centered process.
- Clinical decision support systems (**CDSSs**), like prediction models, aid shared decision-making (clinicians and patients).
- Many health conditions lack validated **prediction models**; existing models often use simple statistics and small datasets.
- Current **models lack Application Programming Interface** for integration of new data and continuous improvement.

Results

- Many clinical rehabilitation variables lack appropriate concepts in existing vocabularies, indicating the **need to extend current libraries**.
- **Overlapping data across clinical cases** offers a chance to establish uniform mapping for rehabilitation data.
- The WHO's International Classification of Functioning, Disability and Health (**ICF**) framework supports the extension of existing vocabularies and class hierarchy tailored to rehabilitation.
- **Extending CDM domains may improve mapping of rehabilitation data**, including (PROs) and long-term treatment.
- These extensions can significantly enhance data mapping accuracy in rehabilitation and for many chronic conditions.
- Comparisons between different cases will guide the newly developed **Rehabilitation OHDSI workgroup**.

Lessons learned from the OMOP CDM implementation in pilot case "Scoliosis during growth (SICO)"

- Extending the ETL (Extract, Transform, Load) to transform data from MySQL to PostgreSQL was crucial for OHDSI tool compatibility and long-term maintainability.
- Many rehab-specific concepts need to be added to **OHDSI Athena**, after discussion and consensus in our community.
- **Recording therapy sessions and other rehabilitation interventions**, including compliance or device changes was challenging, as well as results of questionnaires.
- A dedicated OHDSI CDM set of concepts for **non-pharmacological prescriptions** is lacking.
- The **rehabilitation processes and the databases** used to construct the CDSS were compared for commonalities and distinctions, using Excel files and piloting **GUIDE-Rehab**.

Methods

OHDSI Promote better rehabilitation care by leveraging the OHDSI collaborative to produce large scale observational rehabilitation research.

Mission statement

PREPARE is a €7 million HaDEA-Horizon European project (4 years; 20 partners - 9 countries)

- Applies machine learning (ML) to 9 large-scale patient datasets.
- Uses federated approach for real-world, routinely collected data.
- Develops a platform for sharing model results, based on the OMOP CDM (Observational Medical Outcomes Partnership Common Data Model) standard through the OHDSI collaborative.
- Creates prediction and stratification Machine Learning strategies for rehabilitation data.
- Validates models through nine clinical demonstration pilots.
- Investigates research questions on: changes in clinical decisions, clinician adoption and patient experiences with AI-based CDSSs

Nine pilot cases

PREPARE-Rehab leverages the OMOP-CDM standard to develop and validate AI-driven CDSSs tailored to the complexities of rehabilitation care.

By addressing gaps in existing vocabularies and fostering collaboration within the OHDSI community, this project advances standardized data mapping and machine learning applications in rehabilitation.

The initiative lays a foundation for scalable, interoperable solutions that enhance personalized patient stratification and prediction models to ultimately improve clinical outcomes across diverse rehabilitation settings.

This may be of interest for other fields dealing with chronic conditions.

Logos: IRCCS Ospedale Galeazzi - Sant'Ambrogio, UNIVERSITA DEGLI STUDI DI MILANO, isico, Radboudumc, Erasmus MC, prepare-rehab.eu, PREPARE Rehab, PREPARE REHAB

Contact: Carlotte Kiekens*, Esther Janssen, Lisa Hoogendam, Ruud Selles, Philip van der Wees, Alberto Negrini, Stefano Negrini and the PREPARE group *IRCCS Galeazzi-Sant'Ambrogio Hospital/ISICO - Milan, Italy; carlotte.kiekens@isico.it



#OHDSISocialShowcase This Week

Tuesday

Transforming clinical data from innovative European Cancer Precision Medicine projects: The PRIME-ROSE Example

(**Maria Martin Agudo**, Henk van der Pol, Gabriel Bratseth Stav, Tina Kringelbach, Katarina Puco, Åsmund Flobak, Hans Gelderblom, Kjetil Taskén, Eivind Hovig and Gro Live Fagereng)

OMOP CDM enables transforming aggregated data in a **European clinical trial network** in precision cancer medicine

Transforming clinical data from innovative European Precision Cancer Medicine projects: The PRIME-ROSE Example

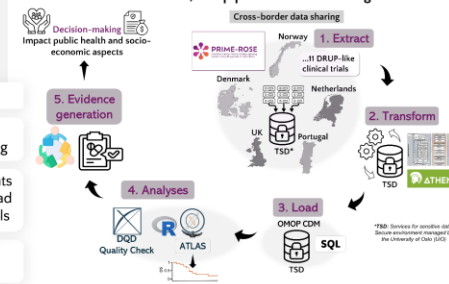
Background: PRIME-ROSE has established a European Precision Cancer Medicine Trial network, where partners share clinical trial data. Merging this data results in **aggregated patient cohorts** defined by tumor type, biomarker and treatment. The aim is to **generate evidence** linked to the feasibility of indication expansion and efficient treatments in terms of clinical outcome and cost-effectiveness in precision cancer medicine. Cross-border data sharing **accelerates the completion of patient enrollment** and subsequent data analysis. **Standardizing** the data to OMOP CDM enables clear cohort analysis and facilitates further cross-border data sharing.

Methods

The data aggregation group is working on the implementation of a pipeline to enable fast and accurate standardization of aggregated data

- Define common variable set
- Trials deposit data in a secure environment. Data pre-processing
- Create reproducible environments for the extract, transform and load (ETL) tools, including OHDSI tools
- Cohort data analysis for evidence generation

Methods 1: Data flow, ETL pipeline and evidence generation



Results



Conclusions: This work will contribute to demonstrate how standardized data can improve cross-border data sharing. As an example, PRIME-ROSE seeks for additional partnerships inside and outside Europe, and the project will benefit from transforming clinical data into OMOP CDM standards. Data generated in this project may become an invaluable resource for producing evidence to inform public health efforts, leading eventually to improved patient care.



Maria Martin Agudo, Henk van der Pol, Gabriel Bratseth Stav, Tina Kringelbach, Katarina Puco, Åsmund Flobak, Hans Gelderblom, Kjetil Taskén, Eivind Hovig and Gro Live Fagereng





#OHDSISocialShowcase This Week

Wednesday

Advancing Epidemiological Research in Africa: Federated Infrastructure, Data Harmonization, and Knowledge Transfer for Scalable Public Health Insights – Technical contribution of the BRIDGE NETWORK project

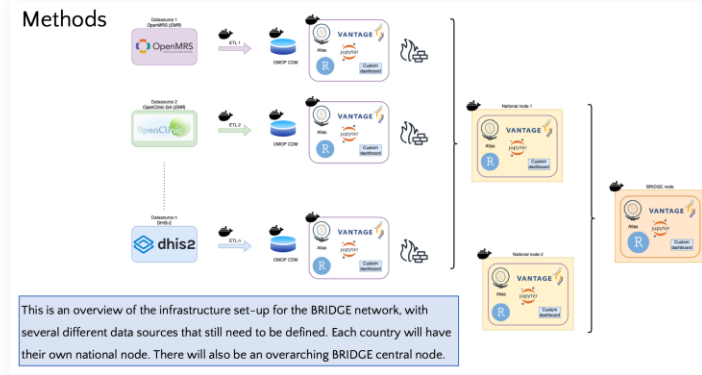
(**Emma Gesquiere**, Lars Halvorsen, Claude Mambo Muvunyi, Marc Twagirimukiza, Pascal Coorevits)

Technical contribution to a more robust and accessible evidence-based data network for public health and epidemiology in Africa

Advancing Epidemiological Research in Africa: Federated Infrastructure, Data Harmonization, and Knowledge Transfer for Scalable Public Health Insights – Technical contribution of the BRIDGE NETWORK

Background: Epidemiological research relies heavily on high-quality, standardized data to analyse disease patterns and inform public health policies and interventions. However, variability in health information systems and data formats hinders interoperability and large-scale (inter)national studies. To address these challenges, the BRIDGE Network aims to empower infectious disease experts to drive research from and for sub-Saharan Africa through creation of an innovative scalable training program that leverages data harmonisation and federated research infrastructure. Knowledge transfer of the technical processes across institutions and countries will ensure the network is sustainable for future research studies.

Methods



Conclusion: This work aims to enhance the efficiency and reproducibility of public health research in Africa by establishing a robust research framework. The integration of OMOP CDM with federated analysis enables scalable, privacy-preserving studies, fostering international collaboration and accelerating evidence generation. These findings will provide valuable insights to the broader OHDSI community and demonstrate effective knowledge transfer within a large-scale international consortium.





#OHDSISocialShowcase This Week

Thursday

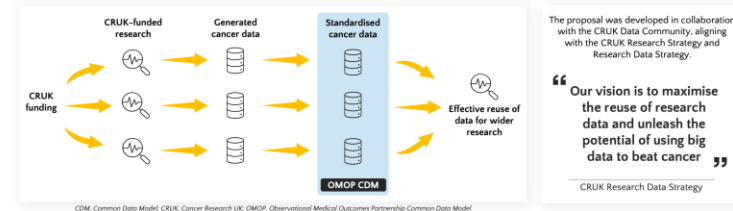
Developing a strategy for the standardisation of Cancer Research UK funded data

(**Jasmine Handford**, Charlotte Moss, Joseph Day, Gemma Codner, Andrew Blake, Mieke Van Hemelrijck)

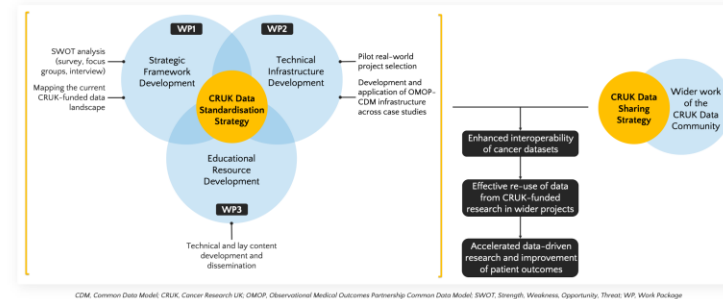
Developing a strategy for the standardisation of Cancer Research UK funded data

The opportunity: Cancer Research UK (CRUK) funds ~£400 million of research per year, with each project using or generating valuable cancer data. However, the collective impact of this research could be amplified by standardising the data produced, e.g. by using a common data model (CDM), facilitating its re-use in a wide range of future projects.

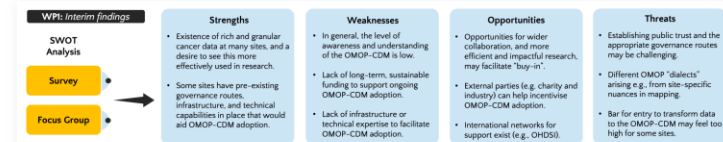
We aim to inform a strategy for OMOP-CDM adoption across CRUK-funded research



Our planned programme of work comprises three interconnected work packages



Interim findings highlight a variety of barriers and enablers for OMOP-CDM adoption



The impact: Alongside wider work to facilitate effective data sharing, this strategy for OMOP-CDM adoption across CRUK-funded projects will drive progress against the CRUK Research Strategy and Research Data Strategy, enabling large scale data integration, accelerating data-driven cancer research, and improving patient outcomes.



Jasmine Handford, Charlotte Moss, Joseph Day, Gemma Codner, Andrew Blake, Mieke Van Hemelrijck
For queries and collaborations, please contact jasmine.handford@kcl.ac.uk





#OHDSISocialShowcase This Week

Friday

RAG-Enhanced LLM Pipeline for Semantic Mapping of Context-based Features to OMOP Vocabulary

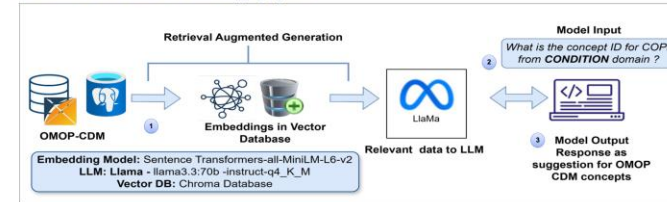
(Sariga Kakkamani, Frederic Jung, Joeri Verbiest, Liesbet Peeters)

Accelerating Feature Extraction with AI-Powered RAG-LLM: Automated Concept Mapping to OMOP-CDM Vocabulary.

Title: RAG-Enhanced LLM Pipeline for Semantic Mapping of Context-based Features to OMOP Vocabulary

Background: Observational health data are often standardized to the commonly used OMOP-CDM standards. This enables us to carry out efficient analyses that can generate reliable evidence. However, understanding these standards and vocabulary terms requires medical knowledge, along with OMOP-CDM expertise. This makes feature extraction crucial, particularly for users without domain expertise.

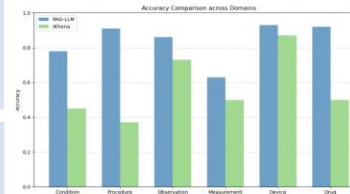
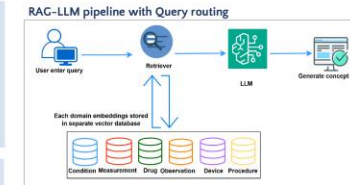
Overview of the RAG-LLM Semantic Mapping Pipeline



Method: In the end-to-end pipeline: (1) OMOP concepts stored in a vector database. (2) User input is encoded and compared against pre-generated embeddings (3) The top-k most semantically similar matches are retrieved and LLM generates context-aware concepts as suggestions.

Results: The pipeline achieved improved performance over the OHDSI tool Athena. With the proposed approach in addition to the suggestions on matching concepts we get explanation on which concepts are more appropriate as per the input query.

Conclusion: The proposed tool mainly focuses on aiding the AI model developers to evaluate their software with a focus on safety, efficacy, and usability, for the direct benefit of patients and healthcare practitioners.



Mapping Accuracy for RAG-LLM pipeline versus Athena at top k = 10 searches



Sariga Kakkamani¹, Frédéric Jung², Joeri Verbiest^{1,3,4}, Liesbet Peeters^{1,3,4}

¹Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium
²VITO, Vlaamse Instelling voor Technologisch Onderzoek, Miel, Belgium
³Biomedical Research Institute, Hasselt University, Diepenbeek, Belgium
⁴University MS Center (UMSC), Hasselt-Paris, Belgium





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?



Nov. 18: DARWIN EU Update





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

Everybody is invited!

Links are sent out weekly and available at:
ohdsi.org/community-calls-2025