HowOften: Next Steps

OHDSI Community Call
Aug. 15, 2023 • 11 am ET

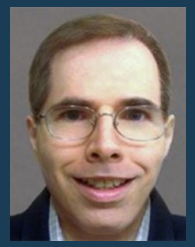
August Community Calls

Date	Topic
Aug. 15	Next Steps for HowOften
Aug. 22	OMOP Supporting Clinical Registries
Aug. 29	Vocabulary Release Update

Aug. 22 — OHDSI and Clinical Registries:
Sanity for Health Systems

Paul Nagy

Program Director for Graduate Training in Biomedical Informatics and Data Science, Deputy Director of the Johns Hopkins Medicine Technology Innovation Center



Lee Evans

Founder, LTS Computing LLC

DuWayne Willett

Chief Medical Informatics Officer, University of Texas Southwestern Health System

Jeff Weaver

Director of Data Solutions for Emory University

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

#OHDSISocialShowcase

ohdsi.org/europe2023-showcase

#OHDSISocialShowcase

2023 Europe Symposium Collaborator Showcase

1	The EHDEN Portal – Simplifying the access to OMOP CDM databases	João Rafael Almeida, Nigel Hughes, Peter Rijnbeek, José Luís Oliveira
2	Privacy-preserving using k-anonymity and l-diversity in OMOP CDM databases	João Rafael Almeida, José Luís Oliveira
3	The Dutch ICU Data Warehouse: towards a standardized multicenter electronic health record database	Ameet Jagesar, Martijn Otten, Tariq Dam, Laurens Biesheuvel, Patrick Thoral, Armand Girbes, Harm-Jan de Grooth, Paul Elbers
4	Community Contribution to the OHDSI Vocabularies, User-Level QA and a New Entity Mapping System SSSOM	Oleg Zhuk, Anna Ostropolets, Nicolas Matentzoglu, Melissa Haendel, Alexander Davydov, Christian Reich
5	Extract, Transform, and Load of the Infectious Disease CDM for Harmonizing and Accessing Data in Real-time Infectious Disease Surveillance	Byungjin Choi, Junhyuk Chang, Soobeen Seol, Seongwon Lee, Rae Woong Park
6	Roadmap and improvement of OHDSI Vocabularies	Christian Reich, Alexander Davydov, Anna Ostropolets
7	Integrating the OMOP CDM into the AI Sandbox of the German Health Data Lab	Elham Taghizadeh, Maxim Moinat
18	Hierarchical clustering of microbial resistance profiles and ventilation protocols using the oncology extension	Jared Houghtaling, Frederic Jung, Ankur Krishnan, Marc Padros Goossens, Frank Leus, Lauren Maxwell, Tom Feusels, Freija Descamps
19	Capture and consolidation of renal specific concepts into a cohesive OMOP dataset	Jared Houghtaling, Jose Antonio Ramírez García, Clémence Le Cornec, Lore Vermeylen, Nir Assaraf, Lars Halvorsen
20	Creation of a reusable OMOP transformation workflow for Belgian electronic health record systems	Jared Houghtaling, Lore Vermeylen, Louise Vandenbroucke, Korneel Bernaert, Brecht Dekeyser, Freija Descamps
21	Construction of a central ontology platform for semantic mapping coordination and vocabulary augmentation across a multi-partner oncology consortium	Jared Houghtaling, Peter Prinsen, Maaike van Swieten, Chiara Attanasio, Lars Halvorsen
22	Application of the R-CDM extension to capture metadata and features extracted from quantitative brain MRI and CT data	Jelle Praet, Jared Houghtaling, Frederic Jung, Steve De Backer, Jeroen Pinxten and Dirk Smeets
23	NNRD-AI: a national neonatal research database for rapid insights with machine learning and artificial intelligence	Julia Lanoue, Kayleigh Ougham, Neena Modi, Sam Greenbury
24	OMOP-CDM Data conversion for the Papageorgiou General Hospital in Greece	Achilleas Chytas, Maria Bigaki, Pantelis Natsiavas
25	Development of a GA4GH Beacon for structured Clinical Data Discovery using the OMOP-CDM	Alberto Labarga, Sergi Aguiló
26	Quality Management System of the OHDSI Standardized Vocabularies	Vlad Korsik, Anna Ostropolets, Christian Reich, Alexander Davydov

Open-source analytics development

46	CDMConnector: Cross platform OMOP CDM database queries using dplyr	Adam Black, Edward Burn, Artem Gorbachev, Martí Català
47	Development of an OMOP Ontology Application – PROSA – for creation and maintenance of highly granular source concepts within the OMOP vocabulary structure	Jared Houghtaling, Emma Gesquiere, and Lars Halvorsen
48	A method to facilitate rapid stand up of OMOP research tools from validated libraries for RWE research	Jack Brewster
49	Generating Synthetic Data from OMOP-CDM databases for Health Applications	Alberto Labarga, Sergi Aguiló
50	Performance Improvement of Post-ETL in OMOP CDM	Antonella Delmestri

Clinical applications

51	Drug utilisation of valproate-containing medicinal products in women of childbearing age: a network study part of DARWIN EU®	Albert Prats-Uribe, Martí Català, Katia M Verhamme, Maria de Ridder, Carlen Reyes, Talita Duarte-Salles, Peter Rijnbeek, Edward Burn, Daniel Prieto-Alhambra, Annika M. Jödicke

ohdsi.org/europe2023-showcase

#OHDSISocialShowcase

MONDAY

Interpretable decision rules for patient-level prediction with EXPLORE

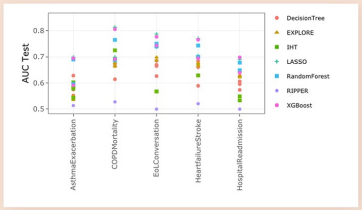
(Aniek F. Markus, Jan A. Kors, Egill A. Fridgeirsson, Katia M.C. Verhamme, Peter R. Rijnbeek)

EXPLORE can achieve similar performance for prediction problems compared to other interpretable models on dataset with 50 covariates

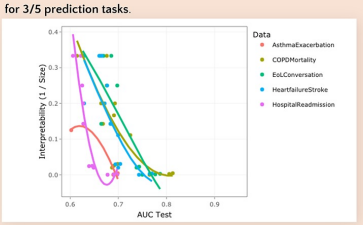
Interpretable decision rules for patient-level prediction with EXPLORE

Background: EXPLORE (Exhaustive Procedure for LOgic-Rule Extraction) is an exhaustive search algorithm designed to find optimal decision rules. This algorithm has several features that make it attractive for patient-level prediction models. Prior work investigating the performance of EXPLORE on standard UCI datasets has shown promising results, but is limited as the studied prediction tasks are much simpler than real-world settings.

Result 1: On the full dataset LASSO, Random forest, and XGBoost using all candidate covariates had the **best predictive performance** across prediction tasks. However, models were **very complex** (221-5315 covariates).



Result 2: On the reduced dataset we observe there exists an interpretable model (3-10 covariates) with **similar or even better** performance than LASSO, Random forest, XGBoost (33-50 covariates) **for 3/5 prediction tasks**.



Methods

1. We investigated the performance (AUC) and complexity of prediction models developed using EXPLORE and other frequently used algorithms across five prediction tasks in the Dutch Integrated Primary Care Information (IPCI) database.
2. We created a reduced dataset with 50 covariates based on the highest (absolute) Pearson's correlation coefficients.
3. Model complexity was measured as the total rule length (EXPLORE, RIPPER), number of nodes (Decision Tree), number of non-zero features (LASSO, Iterative Hard Thresholding), and number of features used in model (Random Forest, XGBoost).

Prediction task	Target population	Outcome of interest	Time-at-risk period
Hospital readmission	Adult patient discharged from hospital	Hospital admission	1 month
End-of-life care	GP visit of older patients (60+) with a prior diagnosis of heart failure, COPD, or cancer	End-of-life conversation	6 months
Asthma exacerbations	Adult patients with new asthma diagnosis receiving medication	Exacerbation (as defined by 3-30 days of systemic corticosteroids)	2 years
Mortality in COPD	Adult patients with new COPD diagnosis	All-cause mortality	2 years
Cardiovascular disease in T2DM	Adult patients with new T2DM diagnosis	Heart failure or stroke	5 years

Limitation: Univariate pre-variable selection was used to make EXPLORE computationally feasible. Future work needs to investigate into how to most effectively reduce data dimensionality for EXPLORE.

OHDSI

Aniek F. Markus, Jan A. Kors, Egill A. Fridgeirsson, Katia M.C. Verhamme, Peter R. Rijnbeek

Erasmus MC University Medical Center Rotterdam
Health Data Science

#OHDSISocialShowcase

TUESDAY

Characteristics and outcomes of over a million inflammatory bowel disease subjects in seven countries: a multinational cohort study

(Chen Yanover, Ramit Magen-Rimon, Erica Voss, Joel Swerdel, Anna Sheahan, Nathan Hall, Jimyung Park, Rae Woong Park, Kwang Jae Lee, Sung Jae Shin, Seung In Seo, Kyung-Joo Lee, Thomas Falconer, Leonard Haas, Paul Nagy, Mary Bowring, Michael Cook, Steven Miller, Tal El-Hay, Maytal Bivas-Benita, Pinchas Akiva, Yehuda Chowers, Roni Weisshof)

Characteristics and Outcomes of >1M **Inflammatory Bowel Disease Patients**

Disease Trajectory of Crohn's Disease and Ulcerative Colitis Patients from Australia, Korea, Japan, the UK, Germany, France, and the USA

Background: Crohn's disease (CD) and ulcerative colitis (UC) are chronic inflammatory bowel diseases (IBD) with consistently increasing incidence rates. These conditions significantly impact the quality of life of patients and families.

Study design: A multinational cohort study using routinely collected healthcare data from 16 OMOPed databases (DBs):

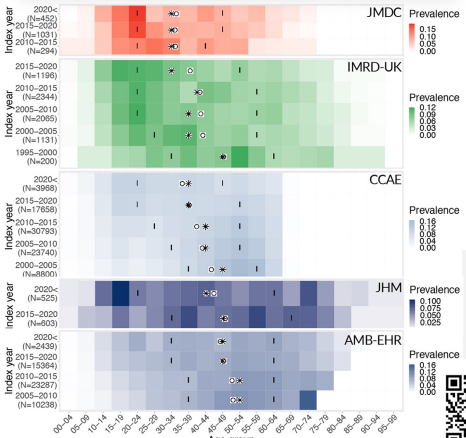
IBM® MarketScan® Commercial Claims DB	CCAE	■	IQVIA™ Adjudicated Health Plan Claims Data	AMB-EHR	◥
IBM® MarketScan® Multi-State Medicaid DB	MDCD	■	IQVIA™ Disease Analyzer – France	France	□
IBM® MarketScan® Medicare Supplemental DB	MDCR	■	IQVIA™ Disease Analyzer – Germany	Germany	□
Optum's Clinformatics® Data Mart – Date of Death	DOD	■	IQVIA™ Medical Research Data – UK	IMRD-UK	□
IQVIA™ Adjudicated Health Plan Claims	PharMetrics+	■	Insurance claims from Japan	JMDC	■
Optum® Pan-Therapeutic Electronic Health Records	Optum EHR	□	Ajou University School of Medicine	AUSOM	□
Columbia University Irving Medical Center	CUIMC	□	Kangdong Sacred Heart Hospital	KDH	□
Johns Hopkins Medicine	JHM	□	IQVIA Australian Longitudinal Patient Data	Australia	□

Geography USA; Europe; Asia; Australia | **Data type** ■ Admin claims; □ EHRs; ◥ Claims + EHRs | **Included visits** Outpatient; Inpatient; ER

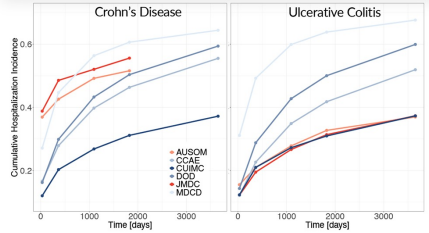
Study population: IBD cohorts include individuals with ≥2 IBD Dx or with IBD Dx + IBD medication Rx; CD and UC cohorts also require at least one diagnosis of the corresponding disease and none of the other.

Characteristics and outcomes: Predefined features (demographics, condition groups, drug era groups), +100 IBD-specific features during subjects' entire history, 1Y, 1M before index date; 1M, 1, 3, 5, 10Y and all-time following index date.

CD diagnosis age decreases over time



Higher CD hospitalization rate in Asia



Limitations:

- Potential differences in coding, reporting across DBs
- Vocabulary updates render concept sets outdated
- HUGE amounts of data (>2G), challenging to view, handle
- Only binary attributes; no cross-strata info

* median age group; | interquartile range; O estimated average

Analysis package | Extended Table 1 | Full results

OHDSI

KI Research Institute	Chen Yanover, Tal El-Hay, Maytal Bivas-Benita, Pinchas Akiva	**Kandong Sacred Heart Hospital**	Seung In Seo, Kyung-Joo Lee
Janssen R&D, LLC	Erica A Voss, Joel Swerdel, Anna Sheahan, Nathan Hall	**Johns Hopkins**	Leonard Haas, Paul Nagy, Mary Bowring, Michael Cook, Steven Miller
Columbia University	Jimyung Park, Thomas Falconer	**Rambam Medical Center**	Ramit Magen-Rimon, Yehuda Chowers, Roni Weisshof
Ajou University	Rae Woong Park, Kwang Jae Lee, Sung Jae Shin		

#OHDSISocialShowcase

WEDNESDAY

Generating Synthetic Data from OMOP-CDM databases for Health Applications

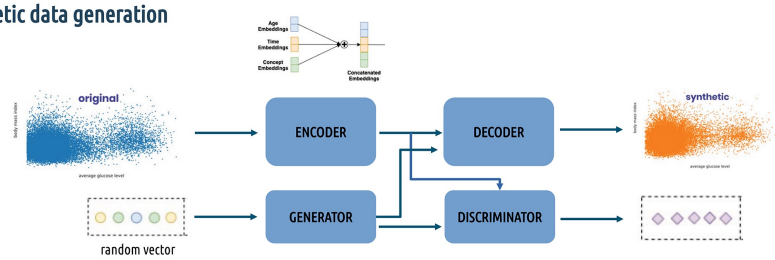
(Alberto Labarga, Sergi Aguiló, S. Capella-Gutierrez)

Generating Synthetic Data from OMOP-CDM Databases for Health Applications

A. Labarga[1,2] , S. Aguiló-Castillo[1,2], S. Capella-Gutierrez[1,2]
[1] Barcelona Supercomputing Center (BSC), Barcelona Spain.
[2] Spanish National Bioinformatics Institute (INB/ELIXIR-ES).

Analysis of Electronic Health Records (EHR) has a tremendous potential for enhancing patient care, quantitatively measuring performance of clinical practices, and facilitating clinical research. Statistical estimation and machine learning (ML) models trained on EHR data can be used to predict the probability of various diseases (such as diabetes), track patient wellness, and predict how patients respond to specific drugs. For such models, researchers and practitioners need access to EHR data. However, it can be challenging to leverage EHR data while ensuring data privacy and conforming to patient confidentiality regulations. Here we present an approach fo generating synthetic health data from an OMOP-CDM. The goal of this study was to develop and evaluate a model for simulating longitudinal healthcare data that adequately captures clinical data temporal and conditional complexities.

Synthetic data generation

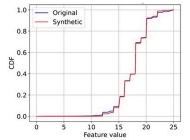


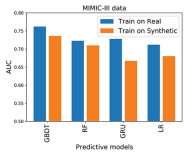
Generating synthetic data comes down to learning the joint probability distribution in an original, real dataset to generate a new dataset with the same distribution. Deep learning models such as **generative adversarial networks (GAN)** and **variational autoencoders (VAE)** are well suited for synthetic data generation but have problems capturing temporal and causal dependencies in the data or generating categorical variables common in clinical data.

We propose a novel generative modeling framework that combines GANs with a **bidirectional encoder representations from transformers (BERT)** architecture. We first train the encoder-decoder model using a reconstruction loss. Then, we use the trained encoder to transform the original inputs into latent space (encoder states). Lastly, we train the GAN framework using an adversarial loss in the latent space to incorporate temporal data across multiple clinical domains. We use a hybrid approach by augmenting the input to BERT using artificial time tokens, incorporating time, age, and concept embeddings, and introducing a new second learning objective for visit type.

Quality evaluation

FIDELITY	UTILITY	PRIVACY
How similar is this synthetic data as compared to the original training sets	How useful is this synthetic data for our downstream machine learning applications	Has any sensitive data been inadvertently synthesized by our model
Kullback-Leibler (KL) divergence, pairwise correlation difference	Accuracy, F1-score, ROC, and AUC-ROC	Membership inference, re-identification and attribute inference attacks





References

1. Murray Reet al. Design and validation of a data simulation model for longitudinal healthcare data. AMIA Annu Symp Proc. 2011
2. Pang etal. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. Proc. of Machine Learning for Health, 158, 2021.
3. Yoon et al. EHR-Safe: Generating High-Fidelity and Privacy-Preserving Synthetic Electronic Health Records https://doi.org/10.21203/rs.3.rs-2347130/v1

Contact
Alberto Labarga
alberto.labarga@bsc.es
Barcelona Supercomputing Center (BSC)
Plaça Eusebi Güell, 1-3. 08034 Barcelona, Spain

Funding
This work has been funded by the Institute of Health Carlos III (project IMPaCT Data, exp. IMP/00019), co-funded by the European Union, European Regional Development Fund (ERDF, "A way to make Europe") and from the European Union's Horizon Europe Programme under GA 101058427 — EOSC4Cancer — HORIZON-INFRA-2021-EOSC-01

#OHDSISocialShowcase

THURSDAY

Building a European cancer OMOP network in hospitals with limited OMOP experience using tiered and modular protocolised research

(Piers Mahon, Atif Adam)

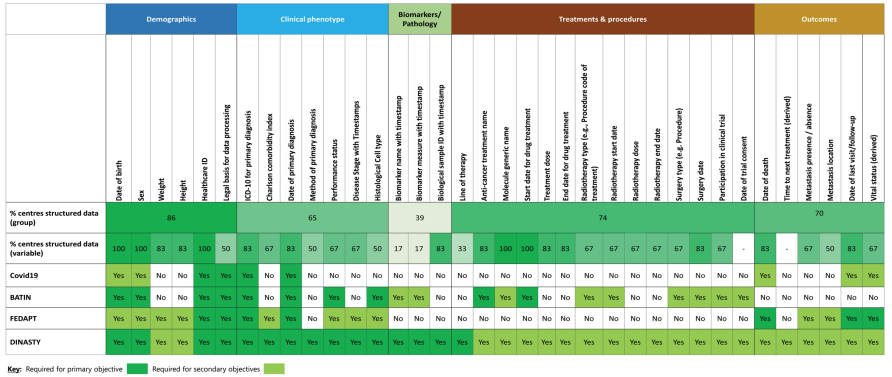
Given OMOP knowledge in most hospitals, DigiONE is normalising cancer data using a protocol driven ETL approach

*Building a European cancer OMOP network in hospitals with limited OMOP experience using tiered and modular protocolised research in **Digi**tal **O**ncology **N**etwork for **E**urope (DigiONE)*

Result 1: Proposed repeatable model to allow multiple hospitals to normalize their data together to allow not only sematic interoperability but also clinical interoperability in concept definition.

	Dx volumes during Covid19 lockdowns	Benchmarking Access To Innovation (BATIN)	Federated Advanced Prognostics (FEDAPT)	Disease Natural history, outcomes with care quality (DINASTY)
Description	Change in volume of cancer diagnoses and 12-month survival by cancer type, sex and age	Examine whether access to innovative drugs, tests or procedures varies by ECOG, sex or age	Predicting 2-year survival; Testing multiple models with increasing # data concepts	Natural history and treatment outcomes studies with care quality assessment
# Data concepts	**9**	**19**	**17**	**36**
Examples	Age, sex, ICD10, Dx date, date of death, date of last visit	+ ECOG, drugs, procedures, trial access	+ Disease stage, location of metastases, comorbidities, histological cell type	All MEDOC: Biomarkers, TTnT, Tx dose
Coverage	All cancers excl. non-melanoma skin	Wave 1: NSCLC Wave 2: breast, ovarian, prostate		
Complexity	Very simple	Simple	Moderate	Hard

Result 2: These 4 protocols get us stepwise to all 36 MEDOC data concepts in OMOP. Below table illustrates which data concepts are used in each of the protocols including the data format by variable and by variable category.



Conclusion: These four protocols are expected to be ready for Ethics Committee submission in summer 2023. Post that, hospitals can perform more advanced protocols with the well characterized cohorts.

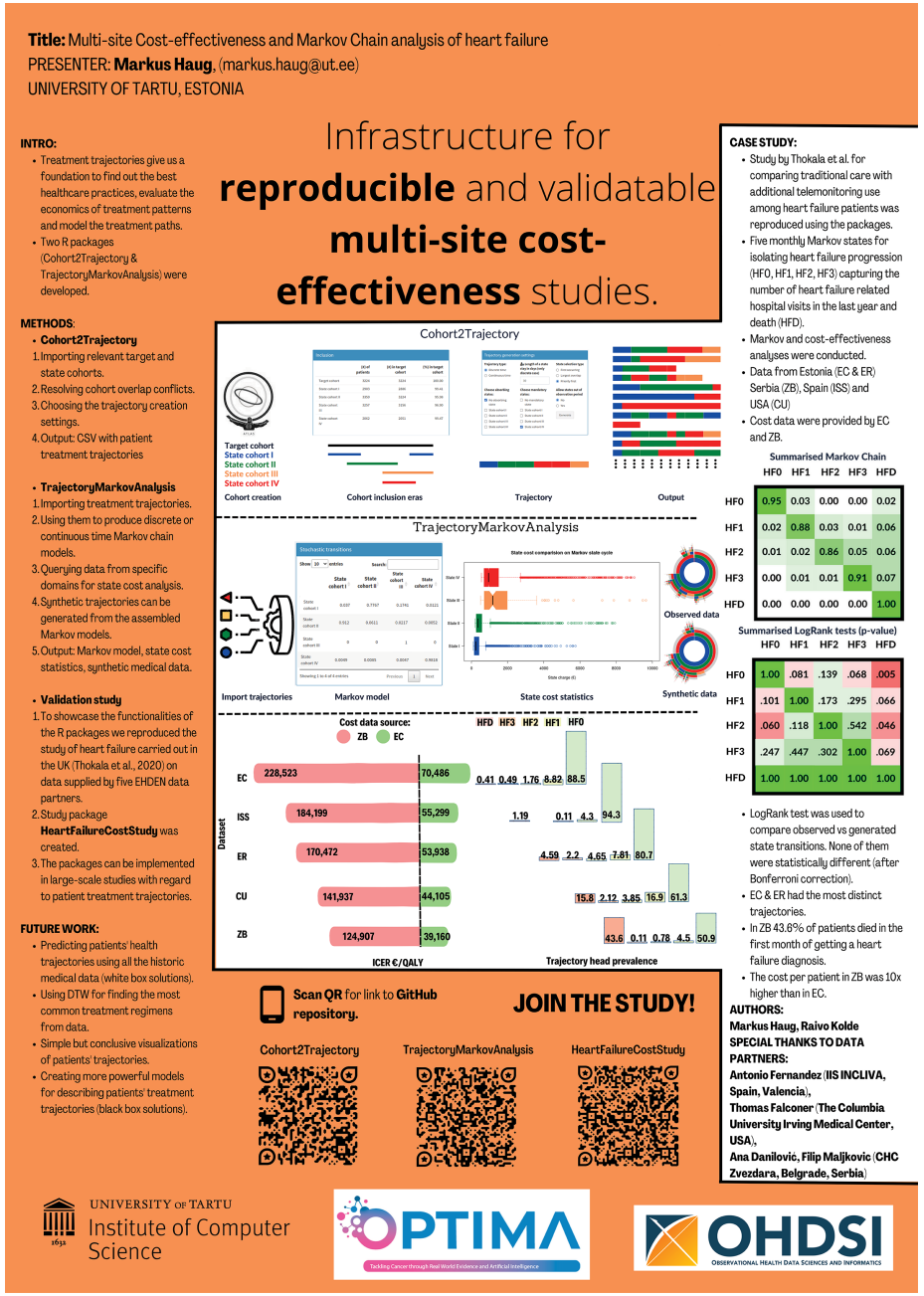
Piers Mahon and Atif Adam

#OHDSISocialShowcase

FRIDAY

Multi-site Cost-effectiveness and Markov Chain analysis of heart failure

(Markus Haug, Raivo Kolde)



OHDSI Shoutouts!

Any shoutouts from the community? Please share and help promote and celebrate OHDSI work!

Do you have anything you want to share? Please send to sachson@ohdsi.org so we can highlight during this call and on our social channels.
Let's work together to promote the collaborative work happening in OHDSI!

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	11 am	Perinatal & Reproductive Health
Wednesday	12 pm	Health Equity Journal Club
Thursday	9 am	OMOP CDM Oncology Vocabulary/Development Subgroup
Thursday	9:30 am	Themis
Thursday	12 pm	Medical Devices
Thursday	12 pm	HADES
Thursday	7 pm	Dentistry
Friday	9 am	GIS – Geographic Information Systems Development
Friday	1 pm	Clinical Trials
Monday	9 am	Vaccine Vocabulary
Monday	10 am	Africa Chapter
Monday	11 am	Data Bricks User Group
Monday	6 pm	OMOP & FHIR
Tuesday	9 am	OMOP CDM Oncology Genomic Subgroup

OHDSI HADES releases: SqlRender 1.15.2

SqlRender 1.15.2 | Reference | Articles | SqlDeveloper | Changelog | HADES

SqlRender 1.15.2

Bugfixes:

1. Fixing translation of `DATEADD()` for DuckDB when number to add is not an integer.

SqlRender 1.15.1 2023-06-29

Bugfixes:

1. Fixed translation of `DATEADD()` for DuckDB when number to add is an expression instead of a verbatim number.
2. Fixed Synapse option in the SqlDeveloper Shiny app.

SqlRender 1.15.0 2023-05-08

Changes:

1. Adding translation of `FROM (VALUES ...) AS drvd(...)` for PostgreSQL, SQL Server, Oracle, RedShift, SQLite, DuckDb, BigQuery, and Spark.

Bugfixes:

1. Correct translation when referring to temp table field for DBMSs that don't support temp tables (e.g. `SELECT #tmp.name FROM #tmp;`).

Contents

Global Symposium

Oct. 20-22 • East Brunswick, NJ, USA

Hilton East Brunswick Hotel & Executive Meeting Center

ohdsi.org/OHDSI2023

Global Symposium Weekend Agenda

	Friday, Oct. 20	Saturday, Oct. 21	Sunday, Oct. 22
7:00 am	Registration/Lite Breakfast	Lite Breakfast	Lite Breakfast
8:00 am	Welcome to OHDSI2023!	Intro to OHDSI Tutorial & OHDSI Workgroup Activities	OHDSI collaborative workshop: HowOften (part 2)
9:00 am	State of the Community		
10:00 am	Community Networking		
11:00	Plenary Session		
12:00 pm	Buffet Lunch	Buffet Lunch + Collaborator Showcase: Posters & Demos	Buffet Lunch + Collaborator Showcase: Posters & Demos
1:00 pm	Panel: Network Studies	OHDSI collaborative workshop: HowOften (part 1)	OHDSI workgroup activites
2:00 pm	Collaborator Showcase: Posters & Demos		
3:00 pm	Collaborator Showcase: Lightning Talks		
4:00 pm	Collaborator Showcase: Posters & Demos		
5:00 pm	Closing Talk & Titan Awards	Free time	We'll see you again in 2024!
6:00 pm	Networking Reception		
7:00 pm	OHDSI Got Talent!		

** this agenda is tentative and subject to change*

OHDSI Got Talent!

Please join us for the first **OHDSI Got Talent!** competition at our 2023 Global Symposium.

We are looking for anybody with a special talent – singing, dancing, playing an instrument, comedy, magic, etc. – to join us for this fun event in October. Please use the link below to share your interest in participation!

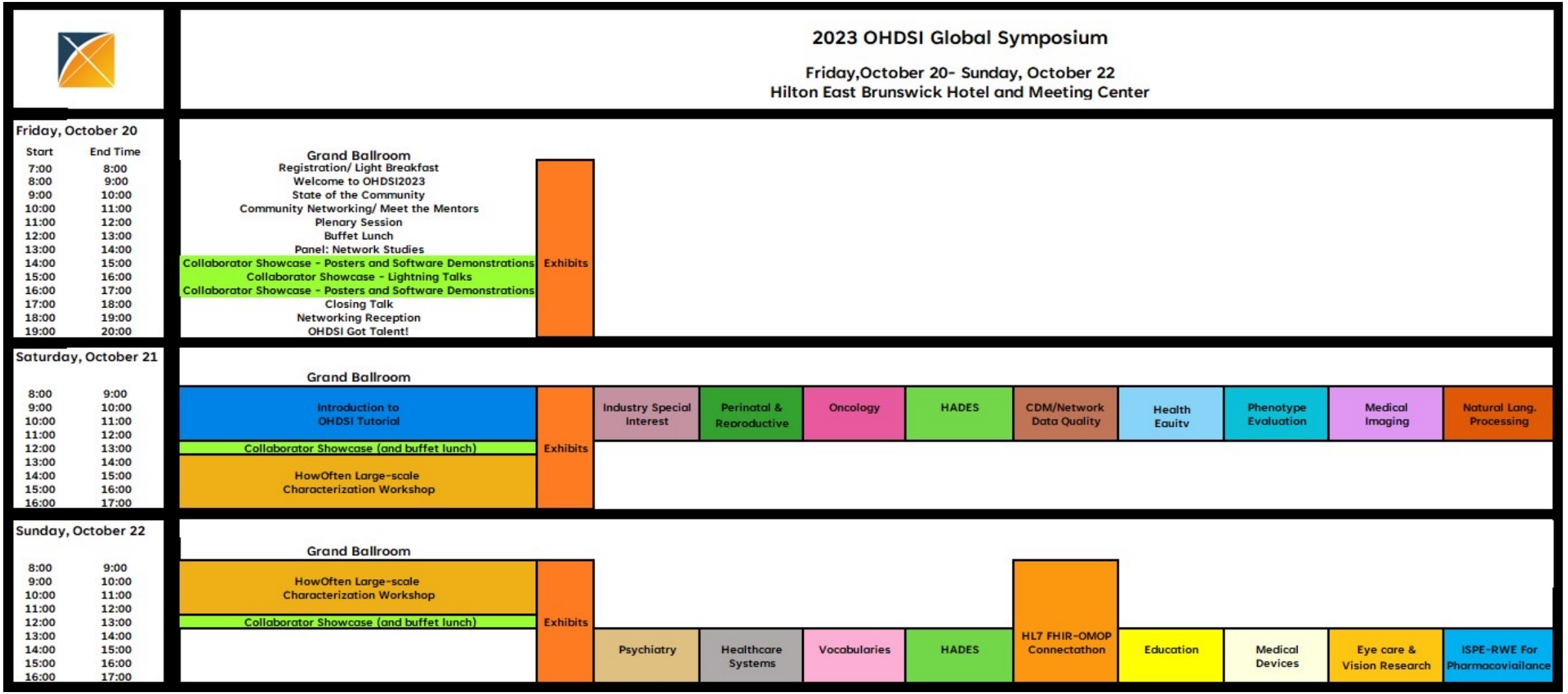
bit.ly/OHDSIGotTalent2023

Titan Award Nominations Are Open!

To recognize OHDSI collaborators (or collaborating institutions) for their contributions towards OHDSI's mission, the OHDSI Titan Awards were introduced at the 2018 Symposium and have been handed out at the Global Symposium each year since.

bit.ly/2023TitanNominations

Global Symposium



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?