Welcome to OHDSI, pt 2

OHDSI Community Call
Oct. 31, 2023 • 11 am ET

Upcoming Community Calls

Date	Topic
Oct. 31	Welcome to OHDSI, part 2
Nov. 7	Meet The Titans
Nov. 14	Collaborator Showcase Honorees
Nov. 21	Showcase Software Demos
Nov. 28	TBA
Dec. 5	Recent Publications
Dec. 12	How Did OHDSI Do This Year?
Dec. 19	Holiday-Themed Goodbye to 2023!

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 Raventós, Martí Català, Mike Du, Yuchen Guo, Adam Black, Ger Inberg, Xintong Li, Kim López-Güell, Danielle Newby, Maria de Ridder, Cesar Barboza, Talita Duarte-Salles, Katia Verhamme, Peter Rijnbeek, Daniel Prieto Alhambra, and Edward Burn** on the publication of **IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model** in *Pharmacoepidemiology & Drug Safety.*

PDS Pharmacoepidemiology & Drug Safety — ispe

ORIGINAL ARTICLE | Open Access

IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model

Berta Raventós, Martí Català, Mike Du, Yuchen Guo, Adam Black, Ger Inberg, Xintong Li, Kim López-Güell, Danielle Newby, Maria de Ridder, Cesar Barboza, Talita Duarte-Salles ... See all authors

First published: 25 October 2023 | https://doi.org/10.1002/pds.5717

Berta Raventós and Martí Català should be considered as joint first-authors
This work has been presented as a Software Demonstration in the Observational Health Data Sciences and Informatics (OHDSI) Symposium held in Bethesda, USA, on 14–16 October 2022.

SECTIONS PDF TOOLS SHARE

Abstract

Purpose

Real-world data (RWD) offers a valuable resource for generating population-level disease epidemiology metrics. We aimed to develop a well-tested and user-friendly R package to compute incidence rates and prevalence in data mapped to the observational medical outcomes partnership (OMOP) common data model (CDM).

Materials and Methods

We created IncidencePrevalence, an R package to support the analysis of population-level incidence rates and point- and period-prevalence in OMOP-formatted data. On top of unit testing, we assessed the face validity of the package. To do so, we calculated incidence rates of COVID-19 using RWD from Spain (SIDIAP) and the United Kingdom (CPRD Aurum), and replicated two previously published studies using data from the Netherlands (IPCI) and the United Kingdom (CPRD Gold). We compared the obtained results to those previously published, and measured execution times by running a benchmark analysis across databases.

Best Community Contribution Honorees!

Data Standards

Augmenting the National COVID Cohort Collaborative (N3C) Dataset with Medicare and Medicaid (CMS) Data, Secure and Deidentified Clinical Dataset

PRESENTER: Stephanie S. Hong

INTRO:

The National COVID Cohort Collaborative (N3C) data Enclave is a platform that provides researchers access to COVID-related patient EMR data in OMOP CDM format. It is the largest centralized repository of COVID-related Patient EMR data in U.S. CMS claims data is also transformed into OMOP CDM format using code map service. N3C COVID patient cohort is now linked to CMS claims data via Privacy Preserving Record Linkage (PPRL). As a result, N3C EMR datasets in OMOP CDM format are enriched with the following additional CMS claims data.

Inpatient	***Outpatient,***
Part D drug prescription	***Home health***
Part B	***Skilled nursing***
Long term care	***Other services***
Durable medical equipment	

METHODS

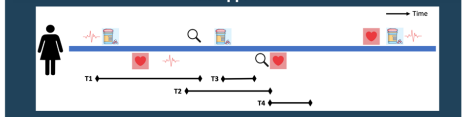
1. CMS claim files in wide format are parsed and pivoted into long format. The clinical concept codes are organized into a condensed format per patient per visit for efficient data transformation.
2. The condensed dataset is then used by the Code Map service to generate the clinical concept translation table. The unified version of the OMOP vocabulary tables are used to perform the translation from the source code to OMOP concept IDs
3. The generated code map service table is used as input in the data pipeline to transform the CMS claims datasets into OMOP CDM format.
4. The data pipeline is built to generate CMS dataset in OMOP CDM format with N3C PPRL linkage.
5. N3C data is enriched with CMS data per PPRL-linked N3C patient. In cases where N3C person_id is duplicated, a Global ID is provided for each.

How much data from Medicare?

How much data from Medicaid?

Link to continuously updated view of N3C enriched dataset via PPRL linkage: https://covid.cd2h.org/dashboard/public-health/pprl/1

A timeline with no gaps, either overlapping or contiguous, is used to construct the *Macro Visit* akin to the N3C macro visit approach.



Take a picture to download the brief report

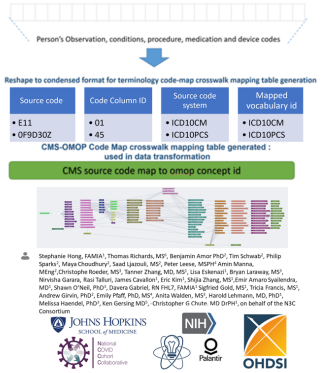
Contact: shong59@jh.edu

RESULT:

82 sites are participating in N3C
30 sites are participating in N3C CMS PPRL linkage
N3C total patients : **20,868,921**
N3C PPRL-linked CMS patients : **359,096**
Total rows of data in N3C: **28.3billion**
Total rows of data in CMS: **653,366,927**
N3C dataset enriched by CMS
Among the PPRL-linked patients, in average, additional concepts are available from CMS:

Claims	Domain	PPRL-Linked patient/ domain	Average # of additional concepts added per patient
Medicare	Condition	71%	78
Medicare	Procedure	60%	6.33
Medicare	Drug_exposure	75%	21.83
Medicare	Measurement	60%	16.48
Medicare	Observation	66.9%	8.6
Medicare	Device	47.6%	6.8
Medicaid	Condition	20.8%	33.9
Medicaid	Procedure	20.2%	23
Medicaid	Drug_exposure	21.9%	20.9
Medicaid	Measurement	17.8%	17.44
Medicaid	Observation	18.3%	6.68
Medicaid	Device	13.9%	6.3

Code Map Service: *Terminology codes appear in multiple columns, i.e.col01 to col45. And some claim source files were over 4000 columns wide. The dataset is pivoted to condense format to generate the clinical concept translation table using OMOP vocab.*



Stephanie Hong, FAMIA[1], Thomas Richards, MS[2], Benjamin Amor PhD[2], Tim Schwab[2], Philip Sparks[2], Maya Choudhury[2], Saad Ljazouli, MS[2], Peter Leese, MSPH[2] Amin Manna, MEng[2],Christophe Roeder, MS[2], Tanner Zhang, MD, MS[2], Lisa Eskenazi[2], Bryan Laraway, MS[2], Nirvisha Garara, Rasi Talluri, James Cavallon[2], Eric Kim[2], Shijia Zhang, MS[2] Emir Amaro Syailendra, MD[2], Shawn O'Neil, PhD[2], Davera Gabriel, RN FHL7, FAMIA[2] Sigfried Gold, MS[2], Tricia Francis, MS[2], Andrew Girvin, PhD[2], Emily Pfaff, PhD, MS[2], Anita Walden, MS[2], Harold Lehmann, MD, PhD[2], Melissa Haendel, PhD[2], Ken Gersing MD[2], Christopher G Chute. MD DrPH[2], on behalf of the N3C Consortium

Augmenting the National COVID Cohort Collaborative (N3C) Dataset with Medicare and Medicaid (CMS) Data, Secure and Deidentified Clinical Dataset

(Stephanie Hong, Thomas Richards, Benjamin Amor, Tim Schwab, Philip Sparks, Maya Choudhury, Saad Ljazouli, Peter Leese, Amin Manna, Christophe Roeder, Tanner Zhang, Lisa Eskenazi, Bryan Laraway, James Cavallon, Eric Kim, Shijia Zhang, Emir Amaro Syailendra, Shawn O'Neil, Davera Gabriel, Sigfried Gold, Tricia Francis, Andrew Girvin, Emily Pfaff, Anita Walden, Harold Lehmann, Melissa Haendel, Ken Gersing, Christopher G Chute)

Best Community Contribution Honorees!

Methods Research

Generating Synthetic Electronic Health Records in OMOP using GPT

Chao Pang[1], Xinzhuo Jiang[1], Nishanth Parameshwar Pavinkurve[1], Krishna S. Kalluri[1], Elise L. Minto[1], Karthik Natarajan[1]
[1]Columbia University Irving Medical Center, Department of Biomedical Informatics

OHDSI

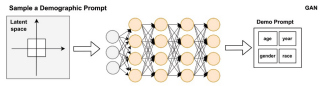
Background

This work focuses on synthetic data generation and demonstrate the capability of training a GPT model using a patient representation derived from CEHR-BERT, enabling the generation of patient sequences that can be seamlessly converted to the OMOP data format bi-direction.

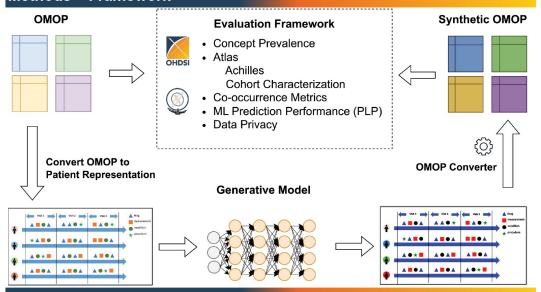
Current approach: Bag of Word + GAN Model

Use cases of synthetic EHR data:

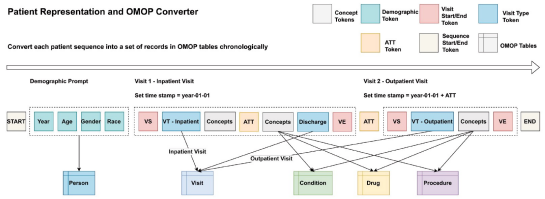
- Phenotype algorithm validation
- Prediction research
- Tool development
- External validation
- Training and education
- Debiasing the the source data
- Counterfactual dataset



Methods – Framework

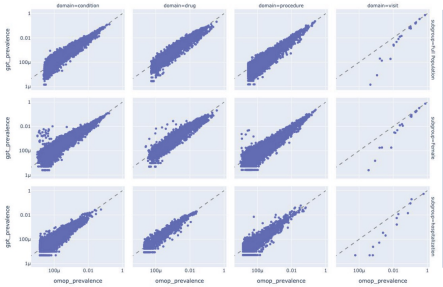


Methods – Patient Representation and OMOP Converter

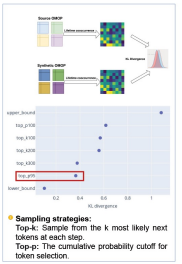


Results

Concept Prevalence



Co-occurrence Metrics



Sampling strategies:
Top-k: Sample from the k most likely next tokens at each step.
Top-p: The cumulative probability cutoff for token selection.

Machine Learning Performance Metrics

Target Cohorts	Real data	Top P cumulative probability cutoff P (%)		Top K K concepts with the highest probabilities	
		Top P = 95%	Top P = 100%	Top K = 100	Top K = 300
HF Readmission	Pre = 25.7 AUC = 65.7 PR = 39.3	Pre = 27.6 AUC = 69.2 PR = 45.7	Pre = 28.4 AUC = 65.9 PR = 41.8	Pre = 30.7 AUC = 68.1 PR = 47.8	Pre = 26.5 AUC = 64.9 PR = 39.3
Hospitalization	Pre = 5.6 AUC = 75.3 PR = 19.5	Pre = 5.2 AUC = 77.1 PR = 21.4	Pre = 7.3 AUC = 68.3 PR = 16.5	Pre = 2.8 AUC = 87.0 PR = 22.1	Pre = 6.3 AUC = 78.7 PR = 24.6
COPD Readmission	Pre = 34.5 AUC = 74.2 PR = 83.8	Pre = 37.8 AUC = 76.4 PR = 84.4	Pre = 47.2 AUC = 74.1 PR = 67.2	Pre = 26.4 AUC = 75.9 PR = 90.3	Pre = 34.5 AUC = 68.8 PR = 80.2
Afib Ischemic Stroke	Pre = 8.7 AUC = 84.0 PR = 48.5	Pre = 10.2 AUC = 78.9 PR = 41.2	Pre = 10.4 AUC = 70.7 PR = 39.1	Pre = 16.6 AUC = 77.1 PR = 50.5	Pre = 10.8 AUC = 76.8 PR = 38.5
CAD CABG	Pre = 7.1 AUC = 88.4 PR = 55.9	Pre = 4.1 AUC = 81.5 PR = 25.2	Pre = 4.4 AUC = 52.9 PR = 4.3	Pre = 7.2 AUC = 75.6 PR = 38.5	Pre = 4.0 AUC = 79.0 PR = 24.1

Conclusions

- First framework generated longitudinal synthetic EHR data using OMOP CDM.
- Designed an innovative patient representation by incorporating temporal information which allowed for an accurate reconstruction of patient medical timeline as compared to state of art methods.
- Comprehensive evaluation procedures showed that the synthetic data preserved the fundamental patient characteristics of the real population.

Contact: CEHR-BERT@lists.cumc.columbia.edu

Generating Synthetic Electronic Health Records in OMOP using GPT

(Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Karthik Natarajan)

Best Community Contribution Honorees!

Open-Source Development

GUSTO Data Vault:
Laying the foundations for an open science system with OMOP Data Catalogue

PRESENTER: Cindy Ho, Mukkesh Kumar

INTRO:
- Growing Up in Singapore Towards healthy Outcomes (GUSTO) aims to understand how conditions in pregnancy and early childhood influence the subsequent health and development of women and children.
- The A*STAR/GUSTO Data Vault platform have advanced data exploration capabilities for research data, biospecimens and publications asset management.
- The OMOP Data Catalogue was created in GUSTO Data Vault to showcase the GUSTO data which have been converted into OMOP CDM format.

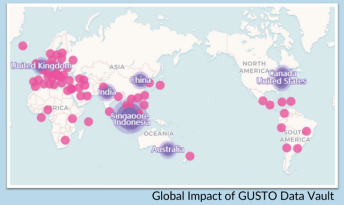
METHODS
- Data Vault (containerized web application with Docker) was built using PostgreSQL database and Django.
- Tools used: HTML, CSS, jQuery, Ajax, Python, Plotly Dash, Dashboard engine in Dash Enterprise, AWS Cloud Platform.
- OMOP fields were mapped using Athena and customized R programming scripts.

RESULTS
- OMOP Data Catalogue makes GUSTO cohort-specific CDM fields to be discovered across the Person, Condition, Observation and Measurement tables by the global research community.
- Metadata is described with relevant attributes such as CDM Field, Concept ID, Name, Subject Type, Visit Timepoint, Description and Domain.
- Data profiling of the OMOP Concept IDs enables GUSTO data to be reused, described, discovered, and identified by researchers (FAIR data principles).
- OMOPed data from incremental OMOP conversions can be seamlessly integrated in OMOP Data Catalogue by GUSTO data curators.
- This enables database level characterizations for GUSTO study.

GUSTO OMOP Data Catalogue lays the foundations for developing cross-study OMOP Data Catalogues expanded across **APAC and global OHDSI data partners**, enabling **database level characterizations.**

Scan to visit GUSTO Data Vault (https://gustodatavault.sg)

Scan to download the abstract



Global Impact of GUSTO Data Vault

Our future work includes the optimization of GUSTO OMOP data conversion journey using advanced OMOP conversion tools such as the IQVIA OMOP Converter.

Snippets of OMOP Data Catalogue

Landing Page

Person table

Measurement table

Observation table

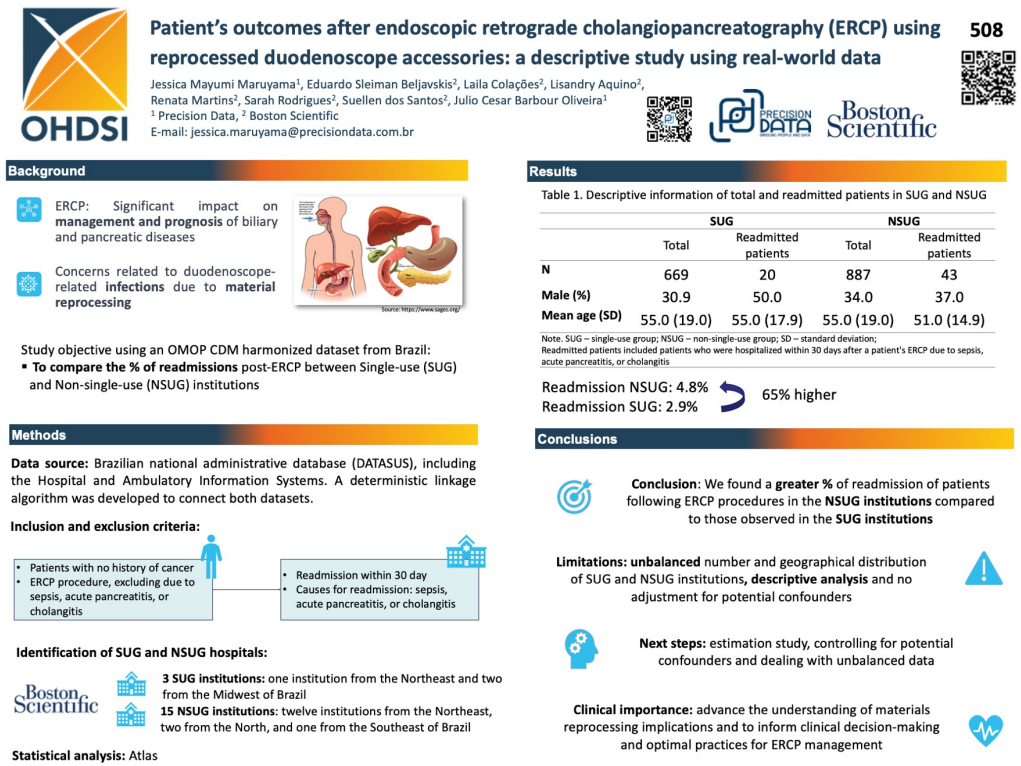
Cindy Ho, Li Ting Ang, Maisie Ng, Hang Png, Shuen Lin Tan, Estella Ye, Sunil Kumar Raja, Mengling Feng, Johan G Eriksson, Mukkesh Kumar

Agency for Science, Technology and Research SINGAPORE
NUS Saw Swee Hock School of Public Health
OHDSI

GUSTO Data Vault: Laying the foundations for an open science system with OMOP Data Catalogue (Cindy Ho, Li Ting Ang, Maisie Ng, Hang Png, Shuen Lin Tan, Estella Ye, Sunil Kumar Raja, Mengling Feng, Johan G Eriksson, Mukkesh Kumar)

Best Community Contribution Honorees!

Clinical Applications



Patient's outcomes after endoscopic retrograde cholangiopancreatography (ERCP) using reprocessed duodenoscope accessories: a descriptive study using real-world data

508

Jessica Mayumi Maruyama[1], Eduardo Sleiman Beljavskis[2], Laila Colações[2], Lisandry Aquino[2], Renata Martins[2], Sarah Rodrigues[2], Suellen dos Santos[2], Julio Cesar Barbour Oliveira[1]
[1] Precision Data, [2] Boston Scientific
E-mail: jessica.maruyama@precisiondata.com.br

OHDSI · PRECISION DATA · Boston Scientific

Background

- ERCP: Significant impact on **management and prognosis** of biliary and pancreatic diseases
- Concerns related to duodenoscope-related **infections** due to **material reprocessing**

Study objective using an OMOP CDM harmonized dataset from Brazil:
- **To compare the % of readmissions** post-ERCP between Single-use (SUG) and Non-single-use (NSUG) institutions

Methods

Data source: Brazilian national administrative database (DATASUS), including the Hospital and Ambulatory Information Systems. A deterministic linkage algorithm was developed to connect both datasets.

Inclusion and exclusion criteria:

- Patients with no history of cancer
- ERCP procedure, excluding due to sepsis, acute pancreatitis, or cholangitis

- Readmission within 30 day
- Causes for readmission: sepsis, acute pancreatitis, or cholangitis

Identification of SUG and NSUG hospitals:

Boston Scientific

3 SUG institutions: one institution from the Northeast and two from the Midwest of Brazil
15 NSUG institutions: twelve institutions from the Northeast, two from the North, and one from the Southeast of Brazil

Statistical analysis: Atlas

Results

Table 1. Descriptive information of total and readmitted patients in SUG and NSUG

	SUG		NSUG	
	Total	Readmitted patients	Total	Readmitted patients
N	669	20	887	43
Male (%)	30.9	50.0	34.0	37.0
Mean age (SD)	55.0 (19.0)	55.0 (17.9)	55.0 (19.0)	51.0 (14.9)

Note. SUG – single-use group; NSUG – non-single-use group; SD – standard deviation;
Readmitted patients included patients who were hospitalized within 30 days after a patient's ERCP due to sepsis, acute pancreatitis, or cholangitis

Readmission NSUG: 4.8%
Readmission SUG: 2.9%
65% higher

Conclusions

Conclusion: We found a **greater %** of readmission of patients following ERCP procedures in the **NSUG institutions** compared to those observed in the **SUG institutions**

Limitations: unbalanced number and geographical distribution of SUG and NSUG institutions**, descriptive analysis** and no adjustment for potential confounders

Next steps: estimation study, controlling for potential confounders and dealing with unbalanced data

Clinical importance: advance the understanding of materials reprocessing implications and to inform clinical decision-making and optimal practices for ERCP management

Patient's outcomes after endoscopic retrograde cholangiopancreatography (ERCP) using reprocessed duodenoscope accessories: a descriptive study using real-world data

(Jessica Mayumi Maruyama, Eduardo Sleiman Beljavskis, Laila Colações, Lisandry Aquino, Renata Martins, Sarah Rodrigues, Suellen dos Santos, Julio Cesar Barbour Oliveira)

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
Tuesday	12 pm	Common Data Model Vocabulary Subgroup
Wednesday	2 am	Methods Research
Wednesday	7 am	Medical Imaging
Wednesday	8 am	Psychiatry
Thursday	1 pm	OMOP CDM Oncology Vocabulary/Development Subgroup
Thursday	7 pm	Dentistry
Friday	9 am	GIS – Geographic Information System General
Friday	11 am	Clinical Trials
Monday	10 am	Healthcare Systems Interest Group
Monday	6 pm	OMOP & FHIR
Tuesday	9 am	ATLAS
Tuesday	10 am	Common Data Model

#OHDSISocialShowcase

MONDAY

Automated Concept Mapping System for OMOP using Vector Representations and Cross-hospital Mapping

(Martina Carres, Gabriel Maeztu, Mónica Arrúe)

Efficient automated mapping of internal source codes to OMOP CDM concepts

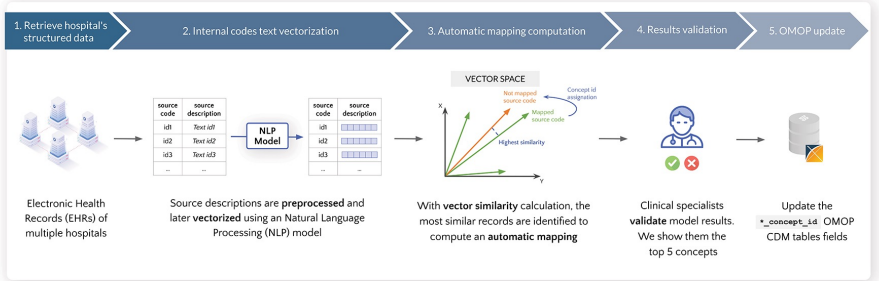
IOMED Medical Solutions

Automated Concept Mapping System for OMOP using Vector Representations and Cross-hospital Mapping

The searching based manual assignment of standard concepts in Observational Medical Outcomes Partnership (OMOP) to local source codes is a time-consuming and error-prone process. Although the OHDSI-developed tool USAGI was designed for this procedure, the non-English language data and OMOP integration limitations hinder efficient mapping

Methods

We propose a method that leverages existing mappings from other hospitals, enabling the **efficient scaling of a single mapping across all others**. The mapping process is automated by computing **vector representations of source code texts**, which capture the relevant syntactic and semantic features, ensuring that similar records are grouped closely in the vector space. Consequently, assigning concept IDs becomes a matter of performing **similarity searches** within the vector space.



Example

source description	vector
Creatinine nmol/dL in serum	*[0.23, 0.45, 0.32, ..., 0.67]*
Leukocytes x10E9/L (serum)	*[0.80, 0.15, 0.20, ..., 0.80]*
White blood cells x10E9/l in serum	*[0.78, 0.21, 0.18, ..., 0.82]*

Similar text vector representation

Close in the vector space

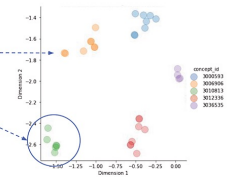


Figure. Text vector representations mapped to 5 concept IDs.

Results

For model evaluation, a comprehensive dataset has been constructed:

- Data sourced from **11 distinct hospitals**
- A total of **79.928 unique source codes**

78.07% reduction in mapping validation time achieved

Accuracy: **72%**
Number of correct mappings executed by the model

Top-5 Accuracy: **88%**
In our use case, shows how likely it is for the correct mapping to be amongst the ones provided to the clinical specialist

OHDSI

Martina Carrés martina.carres@iomed.es
Gabriel Maeztu gabriel.maeztu@iomed.es
Mónica Arrúe monica.arrue@iomed.es
IOMED

#OHDSISocialShowcase This Week

TUESDAY

Mapping Dental Use Cases to the OMOP-CDM: Vocabulary and Common Data Model Evaluation

(Robert Koski, Gopikrishnan Chandrasekharan, William D. Duncan)

Mapping Dental Use Cases to the OMOP-CDM

Vocabulary and Common Data Model Evaluation

PRESENTER: **Robert Koski**

INTRO

Anyone who wants to conduct observational research using dental data with the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) should understand the benefits and limitations of the data model.

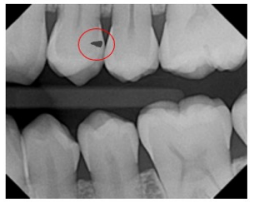
Our hypothetical use case illustrated the challenges of conducting observational research in dentistry.

METHODS

1. Develop a use case
2. Created synthetic data with Synthea
3. Map the data to the OMOP-CDM
4. Explore the mapping and gain key insights

Amongst patients that received a posterior composite restoration, how many patients experienced restoration failure within five years?

Appointment 1: Periodic Evaluation



Appointment 2: Operative Dentistry

The OMOP-CDM has the potential to elevate observational research in dentistry.

The Dentistry Workgroup is leading the effort.

SITE

SITE_EVENT

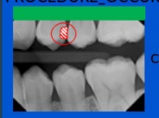
site_id	event_id	site_event_field_concept_id
1	1	1147127
1	1	1147082
1	1	1147115
1	2	1147115

Take a picture to **download** the **full paper**

CONDITION_OCCURRENCE

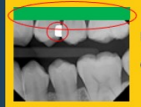
Dental caries

PROCEDURE_OCCURRENCE



Caries excavation

DEVICE_EXPOSURE



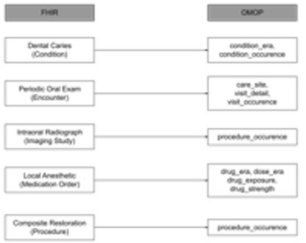
Rubber Dam

Composite resin

RESULTS

- The OMOP-CDM schema can accommodate many dentistry-specific concepts, but is **NOT tooth centric**
- Synthea can create synthetic dental patient data with limitations
- Terminology does exist for dental concepts, but many terms are ambiguous
- Real world data does not typically use diagnostic codes

Example of Mapped Concepts



Additional concepts for the use case are currently being mapped

- Billing codes for dental procedures
- Caries risk
- Dental devices
- Tooth-centric (anatomic site) mapping
- Survival of tooth

Future Use Cases

- Characterization of Obstructive Sleep Apnea treatments
- Care pathways for special needs dental patients
- Periodontal disease management

Robert Koski, Gopikrishnan Chandrasekharan, William D. Duncan

JOHNS HOPKINS SCHOOL of MEDICINE
UF UNIVERSITY of FLORIDA College of Dentistry
OHDSI

#OHDSISocialShowcase This Week

WEDNESDAY

Bayesian sparse logistic models in patient-level predictive studies with the R package PatientLevelPrediction

(Kelly Li, Jenna Reps, Marc Suchard)

Bayesian sparse logistic models in patient-level predictive studies with the R package PatientLevelPrediction

PRESENTER: Kelly Li
Jenna Reps, Marc Suchard

INTRODUCTION

- We implement Bayesian sparse logistic models for advantages over L1-regularized logistic regression:
  - Models uncertainty via posterior distributions of parameter estimates
  - Flexibility for choosing prior distribution leading to sparsity
- Nishimura and Suchard (2022)'s sampler incorporates the **Bayesian Bridge prior** p_0 on regression coefficients β_j:

$$p_0(\beta_j|\tau) \propto \tau^{-1} \exp\left(-\left|\frac{\beta_j}{\tau}\right|^{\alpha}\right)$$

where τ controls the overall shrinkage and α controls the shape of the prior

- We boost informativeness using a **mixture prior** of a normal distribution and the Bayesian Bridge:

$$\beta_j|\tau, \mu_j, \sigma_j^2 \sim \gamma N(\mu_j, \sigma_j^2) + (1-\gamma)p_0(\beta_j|\tau),$$

where each μ_j, σ_j are the prior mean and standard deviations for β_j, and γ_j is a random indicator variable with a Bernoulli prior.

METHODS

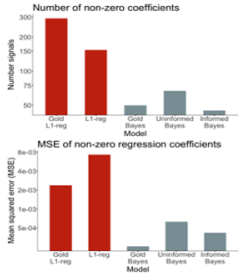
- For simulated and real-world data, we use primary and secondary datasets in the following models for comparison:
  - Uninformed Bayesian model: fits Bayesian regression using the **Bayesian Bridge prior** on primary set
  - Informed Bayesian model: fits an Uninformed Bayesian model on secondary set to inform the following **mixture prior** for the primary set
  - "Gold" standards: fit models on a combined primary/secondary set
- Simulated data: 10^4 x 10^3 design matrix with first 25 coefficients known, rest 0
- Real-world example: Combined 50,000 x 30,000 design matrix to investigate hypothyroidism cases in patients with pharmaceutically-treated depression following the start of the depressive episode

Bayesian sparse logistic regression models exhibit smaller bias and sparser models than L1-regularized models by incorporating prior information on our parameters of interest.

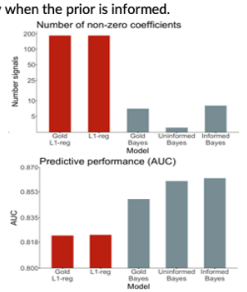
Take a picture to **download** the **full paper**

RESULTS

- Simulated data: Results indicate that Bayesian models consistently exhibit lower mean-squared error (MSE) and higher sparsity than L1-regularized models. Further informing the prior tends to decrease bias while increasing sparsity.



- Real-world data models are fit using the existing package **PatientLevelPrediction**.
- Real-world data: Results indicate that Bayesian models have a slight improvement on predictive performance measured by area under the curve (AUC), along with much sparser models. There is a small improvement in predictive performance and decreased trade-off in sparsity when the prior is informed.



APPLICATIONS

- Population-level studies condition on the propensity-score estimate (logistic regression model): any problems solved for patient-level studies yields an impact for population-level studies too
  - Strength in able to model uncertainty helps to assess reliability and make informed decisions
- Small datasets often lack sufficient information to construct a robust model: being able to inject prior information via Bayesian analyses may bridge the gap

UCLA Fielding School of Public Health

OHDSI

#OHDSISocialShowcase This Week

THURSDAY

Comparing Penalization Methods for Linear Models on Large Observational Health Data

(Egill A. Fridgeirsson, Ross D. Williams, Peter Rijnbeek, Marc Suchard, Jenna Reps)

Title: Comparing Penalization Methods for Linear Models on Large Observational Health Data

PRESENTER: **Egill A. Fridgeirsson**

INTRO:

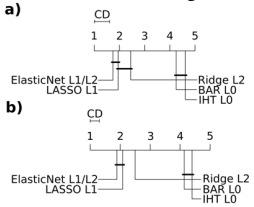
- **We explore the impact of various penalty techniques on predictive model performance aiming to create more robust and generalizable models for better healthcare outcomes**

METHODS

1. We study L1 (LASSO), L2 (Ridge), L1/L2 (ElasticNet) and L0 (iterative hard thresholding and broken adaptive ridge) penalized logistic regression models.
2. We study 21 outcomes occurring in the year after patients start pharmaceutical treatment for major depressive disorder
3. We use five databases (IBM CCAE, IBM MDCR, IBM MDCD, Optum EHR and Optums Claims).
4. We evaluate discrimination (AUC) and calibration (Eavg) and use critical difference diagrams to investigate significant difference in ranks between methods

RESULTS

Critical difference diagram (AUC)

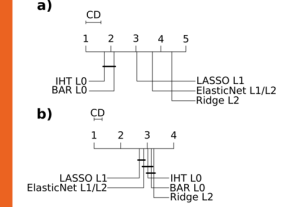


LASSO is the best penalization method for **discrimination.** For **parsimony** and **calibration L0** penalties are best.

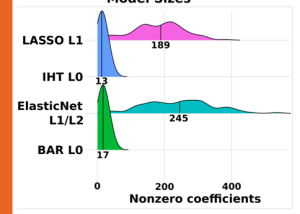
Take a picture to **download** the **full paper**

Calibration:

Critical difference diagram (Eavg)



Model Sizes



IHT: Iterative hard thresholding
BAR: Broken adaptive ridge

Egill A. Fridgeirsson, R. Williams, P. Rijnbeek, M. Suchard, J. Reps

Erasmus MC
Health Data Science
OHDSI

#OHDSISocialShowcase This Week

FRIDAY

Prediction of End Stage Renal Disease in Patients with Type 2 Diabetes Mellitus Patients Using Common Data Model and Machine Learning Algorithm

(Hyuna Yoon, Kyungseon Choi, Sang Youl Rhee, Hae Sun Suh)

Prediction of End Stage Renal Disease in Patients with Type 2 Diabetes Mellitus

PRESENTER: Hyuna Yoon
Contact: hyuna.yoon@khu.ac.kr

INTRO

- **End Stage Renal Disease (ESRD)** leads to substantial expenditure for the country and greatly diminishes individual's quality of life.
- Type 2 Diabetes Mellitus (T2DM) is one of leading causes of T2DM.
- Therefore, preventing ESRD in T2DM patients is particularly important.
- To achieve this, ESRD risk prediction model in T2DM patients is needed.
- Our aim was to develop 10-year ESRD risk prediction model among Korean T2DM patients and

METHODS

1. **Data source**: Electronic Health Record (EHR) data from five secondary or tertiary hospitals standardized to Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5.3.1
2. **Included Patients**: Newly diagnosed with T2DM at a single hospital and had no history of ESRD
3. **Outcome**: Occurrence of ESRD defined as: eGFR<15ml/min or transplantation or dialysis
   (eGFR: estimated Glomerular Filtration Rate)

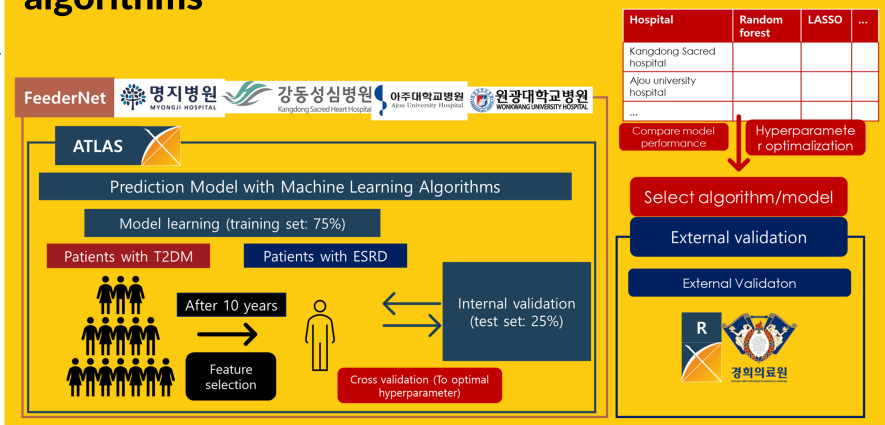
4. Model development

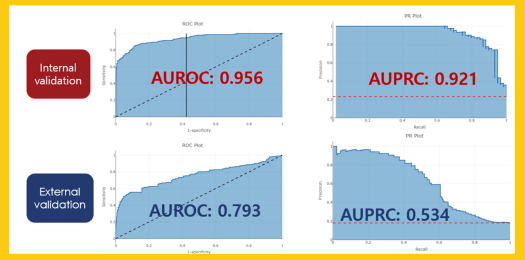
1) Generate several models using several machine learning algorithm
2) Select 3 algorithms which showed best performance
3) Hyperparameter tuning of each three algorithm

5. Model Validation

- Internal validation: using test set of the data that was used to develop model (75%: training set / 25%: test set)
- External validation: Using external data set
- Model Performance Evaluation
  - Area Under the Receiver Operation Characteristic Curve (AUROC)
  - Area Under the Precision-Recall Curve (AUPRC)
  - Positive Predictive Value and Accuracy at different threshold level
  - Calibration Curve

Prediction of end stage renal disease in patients with type 2 diabetes mellitus patients using common data model and machine learning algorithms



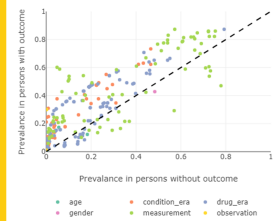


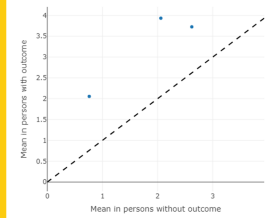
Out of a total of 4,240 variables, 270 were included as predictors

- Measurement above creatinine and urea and nitrogen in serum plasma / Condition of benign hypertension, renal impairment history were considered to be important.

Result

- Considering the model performance and operational principles of algorithm, Lasso logistic regression, random forest, gradient boosting machine was selected.
- Final model: Data from **Myoungji hospital** and **random forest algorithm** was selected as final model
- Feature of final selected model





- Charlson comorbidity index
- Diabetes Complication Severity Index
- CHADS2VASc

Conclusion

- In our study, we developed a prediction model using the random forest algorithm to predict ESRD in patients with T2DM.
- The internal validation of our model demonstrated outstanding performance, and external validation result showed acceptable performance.
- If further external validation is conducted, it can be applied in real clinical practice to assist preventing ESRD in T2DM patients.

Acknowledgement

- This research was supported a grant (21153MFDS601) from ministry of Food and Drug Safety in 2023.

Hyuna Yoon, Kyungseon Choi, Sang Youl Rhee, Hae Sun Suh

IRIS KYUNG HEE UNIVERSITY Institute of Regulatory Innovation through Science

Collaborator Spotlight: Atif Adam

Collaborator Spotlight: Atif Adam

Dr. Atif Adam is a systems scientist and researcher boasting over a decade of diversified experience spanning academia, industry, and entrepreneurial ventures. He attained his doctorate in Health Systems Science and Computational Epidemiology. In addition, Dr. Adam completed his clinical training in Internal Medicine and secured master's degrees in Health Policy and Spatial Epidemiology.

His research probes the nuanced relationships between chronic cardiometabolic diseases, mental health, cognitive aging, and health disparities. During his academic appointments at institutions such as Johns Hopkins and Harvard, Dr. Adam pioneered innovative simulation frameworks for cardiometabolic diseases and rigorously evaluated care pathways for the most vulnerable populations. To this end, he employs advanced statistical, geospatial, and systems modeling methodologies. Transitioning into the digital health space, Dr. Adam co-founded and assumed the role of Chief R&D Officer for the digital mental health startup, Rose Health. In this capacity, he harnessed large-scale data and sensor-based models to curate evidence-based digital solutions, primed for proactive patient monitoring.

In his present role as the Associate Director of Epidemiology at IQVIA, Dr. Adam channels his expertise to spearhead transformative real-world evidence (RWE) initiatives. Within the OMOP team at IQVIA, he merges his deep understanding of health systems, an unwavering commitment to health equity, and knowledge in data science to develop and deliver robust RWE studies at scale. Beyond mere discovery, Dr. Adam is ardently devoted to mentorship, nurturing, and guiding the forthcoming generation of health scientists towards a more informed and equitable healthcare horizon. He discusses his career journey, challenges in health equity and how OHDSI is dealing with them, advice for newcomers in OHDSI, and plenty more in the latest Collaborator Spotlight.

You joined the scientific review committee for the global symposium this year, so what stood out about the variety of research you reviewed for the collaborator showcase?

Joining the scientific review committee for this year's global symposium was a profound journey through cutting-edge developments in health research. As I sifted through the collaborator showcase submissions, it wasn't just the breadth of topics that stood out but also the depth and diversity of approaches employed.

Beyond addressing intriguing questions and hypotheses, there was a noticeable emphasis on methodological innovations. Abstracts showcased a wide range of methods-oriented work, shedding light on enhancements to traditional models and introducing avant-garde techniques. This focus on refining and redefining methods signifies a maturing field, one that's continuously introspecting and evolving.

Atif Adam (right) was a member of the 2023 Scientific Review Committee, and he moderated a series of lightning talks at the Global Symposium.

Several submissions highlighted new collaborations, bringing together multifaceted teams with varied expertise. These collaborations spanned regions and bridged disciplines, underscoring the interdisciplinary nature of modern health research (including Generative AI). Adding new data assets enriched the research landscape, allowing for multifaceted analyses and richer insights.

For me, the collaborator showcase was a microcosm of the future of OHDSI and RWE research. From innovative hypotheses to pioneering methods and new collaborations to the application of advanced models, it was a vivid testament to the dynamism and promise of the global health OHDSI community.

ohdsi.org/spotlight-atif-adam

Global Symposium Homepage

2023 OHDSI Symposium

Oct. 20-22 · East Brunswick, New Jersey

The 2023 OHDSI Global Symposium welcomed more than 440 of our global collaborators together for three days of sharing research, forging new connections and pushing forward together the OHDSI mission of improving health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.

This page will be home to all materials from the global symposium. Check back in the coming days for all video presentations from the event!
#JoinTheJourney #OHDSI2023

State of the Community

Various leaders within OHDSI shared a presentation on the state of the community, with specific focuses on data standards, vocabulary enhancements and open-source development. **Speakers included**:

George Hripcsak, Columbia University
Clair Blacketer, Johnson & Johnson
Alexander Davydov, Odysseus Data Services
Katy Sadowski, Boehringer Ingelheim
Peter Rijnbeek, Erasmus MC
Mengling 'Mornin' Feng, National University of Singapore

video coming soon

State of the Community Slides

2023 Global Collaborator Showcase

Observational Data Standards & Management

2 – FinOMOP – a population-based data network (Javier Gracia-Tabuenca, Perttu Koskenvesa, Pia Tajanen, Sampo Kukkurainen, Gustav Klingstedt, Anna Hammais, Persephone Doupi, Oscar Brück, Leena Hakkarainen, Annu Kaila, Marco Hautalahti, Toni Mikkola, Marianna Niemi, Pasi Rikala, Simo Ryhänen, Anna Virtanen, Arto Mannermaa, Arto Vuori, Joanne Demmler, Eric Fey, Terhi Kilpi, Arho Virkki, Tarja Laitinen, Kimmo Porkka)
3 – From OMOP to CDISC SDTM: Successes, Challenges, and Future Opportunities of using EHR Data for Drug Repurposing in COVID-19 (Wesley Anderson, Ruth Kurtycz, Tahsin Farid, Shermarke Hassan, Kalynn Kennon, Pam Dasher, Danielle Boyce, Will Roddy, Smith F. Heavner)
4 – Augmenting the National COVID Cohort Collaborative (N3C) Dataset with Medicare and Medicaid (CMS) Data, Secure and Deidentified Clinical Dataset (Stephanie Hong, Thomas Richards, Benjamin Amor, Tim Schwab, Philip Sparks, Maya Choudhury, Saad Ljazouli, Peter Leese, Amin Manna, Christophe Roeder, Tanner Zhang, Lisa Eskenazi, Bryan Laraway, James Cavallon, Eric Kim, Shijia Zhang, Emir Amaro Syailendra, Shawn O'Neil, Davera Gabriel, Sigfried Gold, Tricia Francis, Andrew Girvin, Emily Pfaff, Anita Walden, Harold Lehmann, Melissa Haendel, Ken Gersing, Christopher G Chute)
5 – Integrating clinical and laboratory research data using the OMOP CDM (Edward A. Frankenberger, Chun Yang, Vamsidhar Reddy Meda Venkata, Alyssa Goodson)
6 – Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension (Woo Yeon Park, Kyulee Jeon, Teri Sippel Schmidt, Haridimos Kondylakis, Seng Chan You, Paul Nagy)
7 – Conversion of a Myositis Precision Medicine Center into a Common Data Model: A Case Study (Zachary Wang, Will Kelly, Paul Nagy, Christopher A Mecoli)
8 – Implementing a common data model in ophthalmology: Comparison of general eye examination mapping to standard OMOP concepts across two major EHR systems (Justin C. Quon, William Halfpenny, Cindy X. Cai, Sally L. Baxter, Brian C. Toy)
9 – Enhancing Data Quality Management: Introducing Capture and Cleanse Modes to the Data Quality Dashboard (Frank DeFalco, Clair Blacketer)
10 – "OMOP Anywhere": Daily Updates from EHR Data Leveraging Epic's Native Tools (Mujeeb A Basit, Mereeja Varghese, Aamirah Vadsariya, Bhavini Nayee, Margaret Langley, Ashley Huynh, Jennifer Cai, Donglu Xie, Cindy Kao, Eric Nguyen, Todd Boutte, Shiby Antony, Tammye Garrett, Christoph U Lehmann, Duwayne L Willett)
11 – A Toxin Vocabulary for the OMOP CDM (Maksym Trofymenko, Polina Talapova, Tetiana Nesmiian, Andrew Williams, Denys Kaduk, Max Ved, Inna Ageeva)
12 – Challenges and opportunities in adopting OMOP-CDM in Brazilian healthcare: a report from Hospital Israelita Albert Einstein (Maria Abrahao, Uri Adrian Prync Flato, Mateus de Lima Freitas, Diogo Patrão, Amanda Gomes Rabelo, Cesar Augusto Madid Truyts, Gabriela Chiuffa Tunes, Etienne Duin, Gabriel Mesquita de Souza, Soraya Yukari Aashiro, Adriano José Pereira, Edson Amaro)
13 – Transforming the Optum® Enriched Oncology module to OMOP CDM (Dmitry Dymshyts, Clair Blacketer)
14 – Mapping Multi-layered Oncology Data in OMOP (John Methot, Sherry Lee)
15 – Development of psychiatric common data model (P-CDM) leveraging psychiatric scales (Dong Yun Lee, Chungsoo Kim, Rae Woong Park)
16 – Brazilian administrative data for real-world research: a deterministic linkage procedure and OMOP CDM harmonization (Jessica Mayumi Maruyama, Julio Cesar Barbour Oliveira)
17 – Integration of Clinical and Genomic Data Mapped to the OMOP Common Data Model in a Federated Data Network in Belgium (Tatjana Jatsenko, Murat Akand, Joris Robert Vermeesch, Dries Rombaut, Michel Van Speybroeck, Martine Lewi , Valerie Vandeweerd)

ohdsi.org/OHDSI2023

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?

What Questions Did The Community Ask?

- What is the sustainability model?
- How do I learn more about the open problems, especially around methods development?
- Is their a dictionary of Greek named tools and what they are used for?
- Can you clarify the roles of the coordinating center(s) and whether and how members are to engage with them? For example writing grants with them?
- how to troubleshoot the install and understand the OHDSI tools
- Would be awesome if there was a Roadmap/Guide for CDM implementation with tools associated with each step along the way. This may be a lot to ask.
- Wondering if there is collaboration that exists between OHDSI and the military (i.e., DHA - Defense Health Agency)?
- Does OHDSI have a workgroup 'need' level? Like one workgroup may have a tons of contribution and other not so much, where could one help most.
- Any logistic recommendations for members from Institutions that don't have microsoft accounts? It was a challenge to get connected to teams with my institutional email without that.
- Is the list of members of the OHDSI network posted anywhere?
- How can someone have a presentation for more than 8 minutes in the OHDSI US conference like Martijn and Patrick have every year?