Breakout Conversations + Phinal Phenotype Phebruary Report

OHDSI Community Call

March 1, 2022 • 11 am ET

Future OHDSI Community Calls

Date	Topic
March 8	CDM Workshop (Part 1)
March 15	CDM Workshop (Part 2)
March 22	OHDSI Vocabulary Journey
March 29	Reproducibility
April 5	Name That Result
April 12	OHDSI Coordinating Center
April 19	DARWIN EU
April 26	Open-Source Community

Future OHDSI Community Calls

Date	Topic
March 8	CDM Workshop (Part 1)
March 15	CDM Workshop (Part 2)
March 22	OHDSI Vocabulary Journey
March 29	Reproducibility
April 5	Name That Result
April 12	OHDSI Coordinating Center
April 19	DARWIN EU
April 26	Open-Source Community

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

2022 OHDSI U.S. Symposium

The 2022 OHDSI U.S. Symposium will be held Oct. 14-16. The main symposium day is scheduled to be the 14th, while activities will be held the next two days.

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Wednesday	2 am	Patient-Level Prediction/Population-Level Estimation (Eastern Hemi)
Wednesday	9 am	ATLAS
Wednesday	10 am	FHIR and OMOP Digital Quality Measurements Subgroup (Zoom)
Wednesday	4 pm	FHIR and OMOP Data Model Harmonization Subgroup (Zoom)
Thursday	12 pm	Patient-Level Prediction/Population-Level Estimation (Western Hemi)
Thursday	12 pm	FHIR and OMOP Oncology Subgroup
Thursday	6 pm	FHIR and OMOP Terminologies Subgroup (Zoom)
Friday	10:30 am	Clinical Trials
Monday	10 am	GIS– Geographic Information System
Tuesday	9 am	OMOP CDM Oncology

www.ohdsi.org/upcoming-working-group-calls

Get Access To Different Teams/WGs/Chapters

5. Select the workgroups you want to join (you can refer to the WIKI for work group objectives www.ohdsi.org/web/wiki/doku.php?id=projects:overview)

- [ ] ATLAS
- [ ] Clinical Trials
- [ ] Common Data Model
- [ ] Data Quality Dashboard Development
- [ ] Early-stage Researchers
- [ ] Education Work Group
- [ ] FHIR and OMOP
- [ ] Geographic Information System (GIS)
- [ ] HADES Health Analytics Data-to-Evidence Suite
- [ ] Healthcare Systems Interest Group (formerly EHR)
- [ ] Health Equity
- [ ] Latin America
- [ ] Medical Devices
- [ ] Medical Imaging
- [ ] Natural Language Processing
- [ ] OHDSI APAC
- [ ] OHDSI APAC Steering Committee
- [ ] OHDSI Steering Committee
- [ ] Oncology
- [ ] Open-source Community
- [ ] Phenotype Development and Evaluation
- [ ] Population-Level Effect Estimation / Patient-Level Prediction
- [ ] Psychiatry
- [ ] Registry (formerly UK Biobank)
- [ ] Surgery and Perioperative Medicine
- [ ] Vaccine Evidence
- [ ] Vaccine Vocabulary

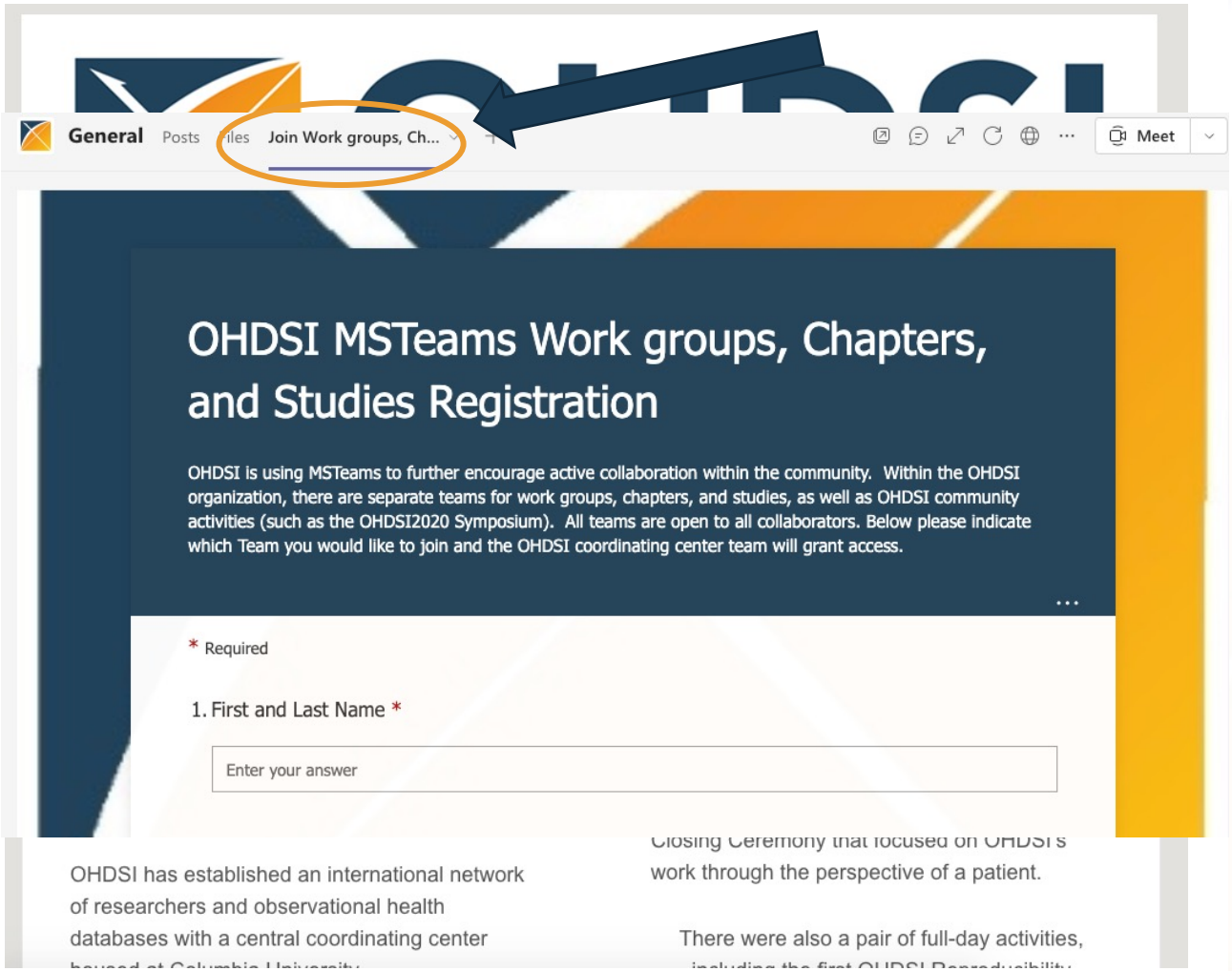
6. Select the chapter(s) you want to join

- [ ] Africa
- [ ] Australia
- [ ] China
- [ ] Europe
- [ ] Japan
- [ ] Korea
- [ ] Singapore
- [ ] Taiwan

7. Select the studies you want to join

- [ ] HERA-Health Equity Research Assessment
- [ ] PIONEER for Prostate Cancer (study-a-thon ended)
- [ ] SCYLLA (SARS-Cov-2 Large-scale Longitudinal Analyses)

Get Access To Different Teams/WGs/Chapters



5. Select the workgroups you want to join (you can refer to the WIKI for work group objectives www.ohdsi.org/web/wiki/doku.php?id=projects:overview)

- ☐ ATLAS
- ☐ Clinical Trials
- ☐ Common Data Model
- ☐ Data Quality Dashboard Development
- ☐ Early-stage Researchers
- ☐ Education Work Group
- ☐ FHIR and OMOP
- ☐ Geographic Information System (GIS)
- ☐ HADES Health Analytics Data-to-Evidence Suite
- ☐ Healthcare Systems Interest Group (formerly EHR)
- ☐ Health Equity
- ☐ Latin America
- ☐ Medical Devices
- ☐ Medical Imaging
- ☐ Natural Language Processing
- ☐ OHDSI APAC
- ☐ OHDSI APAC Steering Committee
- ☐ OHDSI Steering Committee
- ☐ Oncology
- ☐ Open-source Community
- ☐ Phenotype Development and Evaluation
- ☐ Population-Level Effect Estimation / Patient-Level Prediction
- ☐ Psychiatry
- ☐ Registry (formerly UK Biobank)
- ☐ Surgery and Perioperative Medicine
- ☐ Vaccine Evidence
- ☐ Vaccine Vocabulary

6. Select the chapter(s) you want to join

- ☐ Africa
- ☐ Australia
- ☐ China
- ☐ Europe
- ☐ Japan
- ☐ Korea
- ☐ Singapore
- ☐ Taiwan

7. Select the studies you want to join

- ☐ HERA-Health Equity Research Assessment
- ☐ PIONEER for Prostate Cancer (study-a-thon ended)
- ☐ SCYLLA (SARS-Cov-2 Large-scale Longitudinal Analyses)

#OHDSISocialShowcase This Week

OMOP-CDM ETLs with Dask and Prefect

Freija Descamps, Roberto Guarnieri, Lars Halvorsen, Jared Houghtaling
edenceHealth NV, Belgium
Contact: freija@edence.health

Background

For each OHDSI data harmonization project, a dedicated ETL is typically designed that transforms the data from the original source format into the OMOP-CDM model. The design choices and tools used depend on many factors, like for example the format and size of the source data, the available resources on the host where the ETL is run, the available expertise and preference of the data-source partner, the ETL execution frequency and whether the ETL performs an update or an overwrite of the target OMOP-CDM tables.

A typical ETL will broadly consist of a set of tasks that need to be performed in a well-defined order and an executor that handles that workflow. For ETLs that are written in Python, a data-wrangling library like Pandas[1] can perform easy and fast operations on small datasets that comfortably fit in memory. For larger datasets and computation-intensive, but parallelizable, transformations, open-source Python libraries like Dask[2] (parallel computing library) and Prefect[3] (automation and scheduling engine) are a powerful option to further optimize these Python ETLs.

Methods

Dask is an open-source Python library for parallel computing. Among other things, it provides an extension to the Pandas DataFrame (2D structured data collection). The Dask DataFrame consists of a collection of smaller Pandas DataFrames which can be located on disk. This allows Dask to handle operations on datasets that are larger than the available memory as well as to parallelize computations across the constituent Pandas DataFrames, see Figure 1.

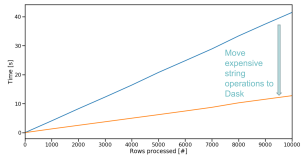


Figure 1. An example plot used for performance profiling illustrating the performance gain by moving a parallelizable computation from Pandas (blue) to Dask (orange).

The Prefect Python library provides tools for building and running data workflows. At a high level, it allows to define a series of interdependent discreet tasks. It also includes a task-library that covers common tasks like for example tasks for executing queries against a Postgres, Redis or other databases. Simply by defining all task dependencies, Prefect can generate and execute the workflow, launching tasks in parallel where possible.

We tested the tools discussed here by implementing a Dask+Prefect ETL that partially transforms large synthetic dataset (Synthea[4]) to the OMOP CDM model. A small toy example workflow for Synthea is shown in Figure 2.

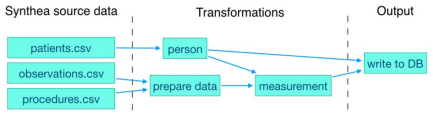


Figure 2. A toy workflow example for a small part of the Synthea[4] dataset. The blue arrows indicate directional dependencies. In this example, the measurement transformation requires the person transformation to have finished. By implementing a Prefect Flow, the 'person' and 'prepare data' tasks will automatically be scheduled in parallel if resources allow.

By combining Prefect and Dask, an ETL can be transformed from a set of tasks that are performed in sequence and on a single core (slow and limited by available memory) to a workflow where tasks as well as computations can be run in parallel on datasets that do not fit in the available memory.

Example use-case

We describe here a real-world example of a Dask OMOPCD-CDM ETL. The original ETL was designed for 500Mb of source data which was made available in a set of csv files. As the data fit comfortably in memory, the ETL was designed to use Pandas and it included expensive and slow row-wise operations. This original ETL ran within 2 hours inside an 8Gb Docker container. The original performance requirements were met. When the data-size was increased to 5GB, a faster and more memory-efficient implementation was required. In this case, Dask was used both for memory-optimization and parallelization of computations, both were needed in order to meet the performance requirements.

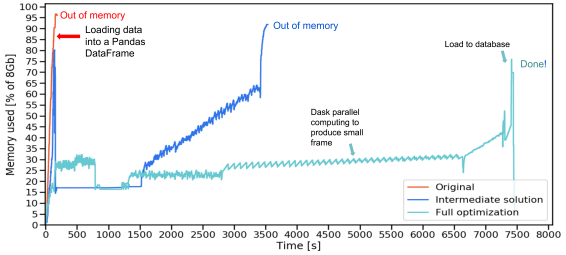


Figure 2. Percentage of memory used (8Gb Docker container) versus time for the original Pandas ETL (orange, runs out of memory when loading data), an intermediate optimization (dark blue, runs out of memory when computing an intermediate frame) and the full optimization (light blue, executes).

Conclusions

The specific design and implementation of an ETL can take many forms. When using Python, open-source parallel processing libraries like Dask and scheduling engines like Prefect can increase the performance by accommodating parallel execution of tasks as well as computations.

References

1. Wes McKinney. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference. 2010;51-56.
2. Matthew Rocklin. Dask: Parallel computation with blocked algorithms and task scheduling. Proceedings of the 14th Python in Science Conference. 2015;130-136.
3. Prefect Delopment Team. PrefectCore: an open-source automation and scheduling engine. https://www.prefect.io
4. Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association, Volume 25, Issue 3, 2018;230–238.

MONDAY

OMOP-CDM ETLs with Dask and Prefect

Authors: Freija Descamps, Roberto Guarnieri, Lars Halvorsen, Jared Houghtaling

#OHDSISocialShowcase This Week

Learning under constraints with EXPLORE

PRESENTER: **Aniek Markus**

INTRO:
The rule induction algorithm EXPLORE has several features that make it attractive for patient-level prediction models:

- The resulting model is a (short) decision rule and can be considered interpretable
- It is possible to specify additional constraints, e.g. minimum specificity or mandatory features that should be included

METHODS:

1. We investigated the performance of EXPLORE in comparison to LASSO logistic regression and RandomForest.
2. As default setting for EXPLORE we use a maximum rule length of 3 and maximize accuracy.
3. We also investigate learning with EXPLORE under two types of constraints: minimum specificity 0.9 and selecting the best predictor in LASSO logistic regression as mandatory feature.

RESULTS:

1. The methods perform roughly similar on the same prediction problems, even though some prediction problems seem to be more difficult than others (see Figure 1).
2. RandomForest uses most features, followed by LASSO logistic regression, and EXPLORE with the lowest number for 11/12 prediction problems.
3. Imposing a minimum specificity of 0.9 leads to a slightly lower AUC.
4. Adding one mandatory feature led to a different model in 4/13 prediction problems. In these cases, the AUC decreased only marginally (0.005-0.016) and increased once (0.01).

Results on standard UCI datasets show that the rule induction algorithm EXPLORE can achieve similar performance as LASSO logistic regression and RandomForest, with substantially smaller models.

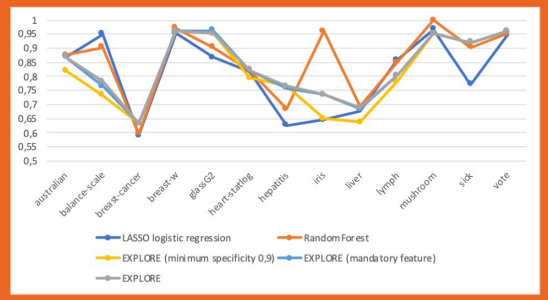


Figure 1: Comparison performance of AUC on standard UCI datasets between LASSO logistic regression, RandomForest, and EXPLORE (maximum rule length 3).

R package to run EXPLORE can be downloaded from GitHub

About EXPLORE:

1. EXPLORE (**Exhaustive Procedure for LOgic-Rule Extraction**) is an exhaustive search algorithm designed to find optimal decision rules.

2. EXPLORE generates decision rules of pre-specified length in **disjunctive normal form** (DNF). A formula in DNF is a disjunction of terms (OR), where the terms are conjunctions (AND) of literals, and the literals are feature-operator-value triples (A > a). An example of a DNF formula is (A > a AND B = b) OR C >= c, the resulting decision rule has the form: if (DNF formula) then class = 1 else class = 0.

3. To find the best decision rule EXPLORE performs an exhaustive search of all possible rules of pre-specified length using smart techniques to reduce the search space. For example, by reducing the number of values for each feature that need to be checked (**subsumption pruning**) and disregarding subspaces that cannot contain the optimal decision rule (**branch-and-bound**).

Aniek F. Markus, MSc,
Jan A. Kors, PhD,
Peter R. Rijnbeek, PhD

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

TUESDAY

Learning under constraints with EXPLORE

Authors: Aniek F. Markus, Jan A. Kors, Peter R. Rijnbeek

#OHDSISocialShowcase This Week

Estimating Model Performance on External Datasets from Their Limited Statistical Characteristics: Application to 3-Year Surgery Risk in Ulcerative Colitis

Tal El Hay and Chen Yanover
KI Research Institute

The Israeli Institute for Applied Research in Computational Health

Background

External validation, that is performance evaluation of a model trained using an "internal" data source on other datasets, is increasingly recognized as an essential step in demonstrating model robustness[1]. A more proactive approach may seek a model that performs well on multiple datasets. A mechanism that estimates the performance of a given model on external sources from their statistical characteristics could support such a model optimization procedure.

Here, we propose an algorithm which reweighs samples in the internal dataset according to external dataset statistics, potentially reported in a preceding publication (as "Table 1") or a characterization study. The algorithm then uses the reweighted samples to estimate model performance on the external source. We validate our approach using a prediction model for 3-year risk of intestinal surgery in ulcerative colitis patients.

Methods

Data. We used primary care electronic medical records from the UK (IQVIA Medical Research Data incorporates data from THIN, A Cegedim Database; reference made to THIN is intended to be descriptive of the data asset licensed by IQVIA), which covers approximately 6% of the UK population, and is representative of the population in terms of demographics and major condition prevalence.

Ulcerative colitis use-case. For each patient in the ulcerative colitis cohort (identified based on diagnostic codes and prescriptions; see https://github.com/ohdsi-studies/IbdCharacterization for more details), we extracted a set of features, previously observed as associated with increased intestinal surgery risk[2]. The outcome considers procedure codes for colostomy, colectomy, ileostomy, small intestinal resection, stricturoplasty, balloon dilation, drainage of perianal abscess, drainage of intra-abdominal abscess, or death, within 3 years following index date (first diagnosis or related prescription). We excluded subjects with insufficient follow-up. All concept sets and cohorts have been defined as part of the IBD Characterization project.

Reweighing algorithm. Our goal is to obtain a (weighted) sample of the internal population with statistical properties that are similar to the ones available, e.g., as Table 1, for the external datasets. To avoid overfitting, we also require the set of weights to be close to uniform (i.e., maximal entropy). To derive an optimal set of weights, we define the following optimization problem:

$$\text{minimize}_{\,w} \left\| X^T_{internal} \cdot \vec{w} - \vec{\mu}_{external} \right\|_2 - \lambda \cdot \mathcal{H}(\vec{w}),$$
$$\text{such that } \sum_{i=1}^{n} w_i = 1, w_i \geq 0, \forall_i$$

where $X \in \mathbb{R}^{n \times p}$ is the feature matrix, $\vec{\mu} = \frac{1}{n}\sum_i \vec{x}_i$ is a p-dimensional vector of feature means, $\vec{w} \in \mathbb{R}^n$ are the inferred patient-specific weights, $\mathcal{H}(\vec{w})$ is the weight entropy and λ is a tunable parameter. The inferred weights induce a weighted sample $\{\vec{x}_i, y_i, w_i\}_{i=1}^n$ whose properties approximate the statistics of the external sample. In this study, we used this sample to approximate the area under the receiver operating characteristic curve (AUC) using R's *WeightedROC* library.

Algorithm evaluation. We trained, tuned, and validated the proposed approach on England-based patient cohorts and tested it on three external datasets, including patients residing in Scotland, Wales, and Northern Ireland.

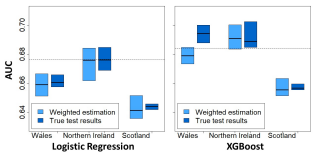
Contact: {talelh, chen}@kinstitute.org.il

Results

The following table summarizes the characteristics of the various location-based cohorts. Note the difference in various features, e.g., the younger age and greater smoking percentage in the Northern Ireland cohort; and the more prevalent use of steroids in Scotland.

	England train	England test	Wales	Northern Ireland	Scotland
No. of subjects	7559	1915	1255	772	1772
Age	48.9 ±18.9	48.0 ±18.9	48.3 ±19.1	46.0 ±18.2	47.0 ±18.5
Female	3700 (48.9%)	938 (49.0%)	602 (48.0%)	382 (49.5%)	909 (51.3%)
Smoking	1770 (23.4%)	461 (24.0%)	313 (24.9%)	221 (28.6%)	484 (27.3%)
Steroids use	2279 (30.1%)	555 (29.0%)	408 (32.5%)	224 (29.0%)	668 (37.7%)
Underweight	202 (2.7%)	46 (2.4%)	30 (2.4%)	24 (3.1%)	37 (2.1%)
Overweight	1839 (24.3%)	440 (23.0%)	343 (27.3%)	200 (25.9%)	442 (24.9%)
Perianal disease	126 (1.7%)	18 (0.9%)	16 (1.3%)	12 (1.6%)	11 (0.6%)
Gastrointestinal procedures	450 (6.0%)	104 (5.4%)	95 (7.6%)	48 (6.2%)	121 (6.8%)

The boxplots on the right show the distribution of actual (dark blue) and estimated performance (light blue) using regularized logistic regression and XGBoost. The dashed line shows the AUC for England's test set (as a baseline). In all instances (except Wales with XGBoost), AUC estimations are more similar to the actual values than the internal calculated performance (dashed line).



Conclusions

Our proposed algorithm can help in identifying models that perform well across multiple clinical settings and geographies, even when detailed test data from such settings is not available.

Notably, this is a work-in-progress, currently having several limitations. First, the divergence between internal and external validation sets in this study may be low and not representative of typical studies. Second, approximation quality depends on the level of detail of the statistical information. Third, the algorithm requires tuning a hyper parameter (λ) that may be crucial to avoid overfitting.

We believe that the proposed algorithm can serve as a building block in network studies that aim to construct robust models across datasets, using OHDSI's tools for extracting and sharing population-level statistics.

References

1. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. Published online June 21, 2021. doi:10.1001/jamainternmed.2021.2626
2. Koliani-Pace JL, Siegel CA. Prognosticating the Course of Inflammatory Bowel Disease. Gastrointest Endosc Clin N Am. 2019;29(3):395-404. doi:10.1016/j.giec.2019.02.003

WEDNESDAY Estimating Model Performance on External Datasets from Their Limited Statistical Characteristics: Application to 3-Year Surgery Risk in Ulcerative Colitis

Authors: Tal El Hay, Chen Yanover

Where Are We Going?

Any other announcements of upcoming work, events, deadlines, etc?

Three Stages of The Journey

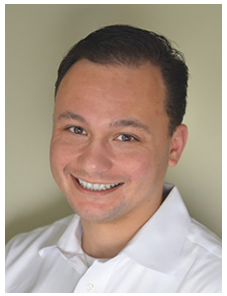
Where Have We Been?

Where Are We Now?

Where Are We Going?

March 1 OHDSI Community Call

Breakout Discussions: What Is Happening In OHDSI, And What Comes Next?

Characterization

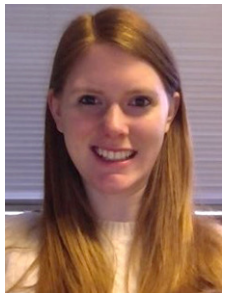
Aniek Markus and Anthony Sena

(stay here)


Estimation

Martijn Schuemie and Marc Suchard

(new link)




Prediction

Jenna Reps and Ross Williams

(new link)

Also, we will have our phinal Phenotype Phebruary Report, pheaturing phenotype leaders from the past week!