Version: 1.0

Authors:

Clair Blacketer, Janssen Research & Development

Gowtham Rao, Janssen Research & Development

Paul Nagy, Johns Hopkins University

Date: 06-10-2024

Acknowledgement: The analysis is based in part on work from the Observational Health Data Sciences and Informatics collaborative. OHDSI (http://ohdsi.org) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics.

1 Table of contents

2 List of abbreviations

CDM – Common Data Model
OMOP – Observational Medical Outcomes Partnership
OHDSI – Observational Health Data Sciences and Informatics

3 Abstract

This study aims to investigate the data and concept distribution patterns across different OHDSI datasets and their impact on study feasibility to facilitate cross-institutional and network studies. The aggregate statistics collected will inform the similarities between federated network data partners, what a "typical" OHDSI dataset looks like, and allow for rapid feasibility assessments for across the network.

4 Amendments and Milestones

Protocol version	Planned / Estimated Date	Brief description
1.0		Initial version

Milestone	Planned / Estimated Date
Finalize Protocol	By 6/10
Data collection from OHDSI sites	
End of analysis	
Posting of results	
Co-authors review and approval	
Draft manuscript w co-authors	
Submission of manuscript	

5 Rationale and Background

The Observational Health Data Sciences and Informatics (OHDSI) federated network is a collaborative effort aimed at leveraging healthcare data from multiple institutions for large-scale federated observational research. In its current state there are over 500 data sources from over 49 countries mapped to the OMOP Common Data Model, the standard that enables such ambitious evidence generation. One major challenge of federated network studies is the assessment of network data quality, study feasibility and data fitness-for-use across these data sources in such a way that does not strain the time and resources of data holders while still supporting rigorous evidence generation that engenders trust and buy-in from the larger research community.

To facilitate collaborative research efforts and ensure the quality and integrity of the data across the OHDSI network, it is imperative to understand the characteristics and variability of the databases within the network. This study aims to collect summary statistics from participating sites to describe the databases and learn about the network as a whole. The output of the study will inform and enhance the research capabilities of the OHDSI community by enabling rapid data quality and fitness-for-use assessments.

5.1 Research Questions

The main research question of this study is:

What are the population-level characteristics of the databases within the OHDSI federated network?

The specific aims of this study are as follows:

- To create an open public resource comprised of summary statistics of the databases within the OHDSI network (that the data owners are able to provide in compliance with IRB, GDPR, HIPAA) to support research.
    - To collect population-level summary statistics of databases within the OHDSI federated network to inform study feasibility for network research.
    - To generate network-based benchmarks based on the collected statistics to support observational research and analysis. These will be used to describe the network and inform data owners about the quality of their data by learning what a “typical” OMOP CDM standardized databases looks like. This will be done by characterizing the heterogeneity, granularity, timeliness, and domain coverage of the participating databases.

6 Research methods

6.1 Study Design

1. **Data Collection:** Participating sites will execute pre-defined queries by running an R package against their local databases to generate summary statistics. Sites will also share high-level descriptive information about their data to aid in identification. The list of statistics and descriptive information to be collected is available in section 6.1.1.

1. The R package is available at GitHub: https://github.com/ohdsi/DbDiagnostics

2. **Data Submission:** Participating sites will submit the summary statistics to the research team who will review the results for quality.
3. **Data Aggregation:** The research team will develop network-based benchmarks and characterizations by comparing the summaries of contributing databases.
4. **Data Storage:** The collected statistics will be stored as an open public resource

6.1.1 Collected Summary Statistics

We will use the DbDiagnostics R package that will run SQL queries on each site to produce these descriptive statistics. The summary will consist of:

- Number of persons
- Number of persons by gender
- Number of persons by year of birth
- Number of persons by race
- Number of persons by ethnicity
- Number of persons with at least one day of observation in each month
- Number of persons by observation period start month
- Number of persons by number of observation periods
- Number of persons by length of observation period, in 30d increments
- Number of persons with at least one visit occurrence, by visit_concept_id
- Number of distinct patients that overlap between specific domains - including death
- Number of persons with at least one concept_id, by measurement_concept_id
- Number of measurement occurrence records, by measurement_concept_id
- Number of measurement occurrence records, by measurement_source_concept_id
- Number of measurement records, by measurement_concept_id and value_as_concept_id
- Number of measurement records with no value (numeric, string, or concept)
- Number of persons with at least one concept_id, by condition_concept_id
- Number of condition occurrence records, by condition_concept_id
- Number of condition occurrence records, by condition_source_concept_id
- Number of persons with at least one concept_id, by procedure_concept_id
- Number of procedure occurrence records, by procedure_concept_id
- Number of procedure occurrence records, by procedure_source_concept_id
- Number of persons with at least one concept_id, by drug_concept_id
- Number of drug exposure records, by drug_concept_id
- Number of drug exposure records, by drug_source_concept_id
- Number of persons with at least one concept_id, by observation_concept_id
- Number of observation occurrence records, by observation_concept_id
- Number of observation occurrence records, by observation_source_concept_id
- Number of observation records, by observation_concept_id and value_as_concept_id
- Number of persons with at least one concept_id, by device_concept_id
- Number of device exposure records, by device_concept_id
- Number of device exposure records, by device_source_concept_id
- Distribution of numeric values, by measurement_concept_id and unit_concept_id

In addition, descriptive information about the database will be collected, which includes:

- Dataset name
- Name of the owner or licensee of the dataset
- Dataset DOI, if applicable
- Type of data in the database (EHR, administrative claims, clinical registry, etc.)
- OHDSI Standardized Vocabularies version
- Name of contact person responsible for network studies on the database
- Email address of contact person responsible for network studies on the database
- If the site has participated in an OHDSI network study before
- If there is someone at the site who can run an OHDSI study package
- How long in weeks it takes to get approval to run a study using the dataset
- How often the data are refreshed

6.1.2 Excluded Events

Each participating site will have the option to exclude any concepts necessary to adhere to their individual data governance guidelines.

6.2 Data Sources

The analyses will be performed across a network of observational healthcare databases that have been transformed into OMOP CDM, version 5 series.

7 Protection of Human Subjects

The collected statistics are aggregate and pose no privacy risk to individual patients. To reduce the risk of reidentification, any concepts with a record count < 10 will be rounded up to 10. Any concepts chosen by each individual site to be excluded will not be available in the results, as described in section 6.1.2.

8 Study Results Dissemination

The aggregated study results will be posted on the OHDSI website after completion of the study. The summary statistics generated will be made available as an open public resource by the OHDSI Coordinating Center for reuse to support network research. At least one paper will be written and submitted for publication to a peer-reviewed scientific journal with an appendix containing the summary statistics from each contributing organization.