**Version:** 0.1

**Authors:**

Anna Ostropolets, Department of Biomedical Informatics, Columbia University

Chunhua Weng, PhD, Department of Biomedical Informatics, Columbia University

Casey Ta, PhD, Department of Biomedical Informatics, Columbia University

Patrick Ryan, PhD, Janssen Research and Development

**Date:** 03-15-2019

**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.

Table of contents

[2 List of abbreviations 2](#_Toc3474981)

[3 Abstract 2](#_Toc3474982)

[4 Amendments and Milestones 3](#_Toc3474983)

[5 Rationale and Background 3](#_Toc3474984)

[5.1 Research Questions 3](#_Toc3474988)

[6 Research methods 4](#_Toc3474993)

[6.1 Study Design 4](#_Toc3474994)

[6.2 Data Sources 6](#_Toc3474995)

[6.3 Use Cases 6](#_Toc3474996)

[7 Protection of Human Subjects 6](#_Toc3474998)

[8 Study Results Dissemination 7](#_Toc3474999)

# 

# List of abbreviations

CDM – Common Data Model

OMOP – Observational Medical Outcomes Partnership

OHDSI – Observational Health Data Sciences and Informatics

# Abstract

This study aims to investigate the concept usage patterns across different OHDSI datasets and their impact on cohort identification to facilitate cross-institutional and network studies. For every standard concept in the OMOP CDM, we will investigate if it is used at each site, and if used, which source code(s) are used at individual sites, and how differences in corresponding source codes impact the portability of cohort identification algorithms.

# Amendments and Milestones

|  |  |  |
| --- | --- | --- |
| Protocol version | Planned / Estimated Date | Brief description |
| 0.1 |  | Initial draft |

|  |  |
| --- | --- |
| Milestone | Planned / Estimated Date |
| Finalize Protocol | By 3/25 |
| Data collection from OHDSI sites |  |
| End of analysis |  |
| Posting of results |  |
| Co-authors review and approval |  |
| Draft manuscript w co-authors |  |
| Submission of manuscript |  |

# Rationale and Background

## The goal of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is to create a standardized approach at many levels: (1) standard representation of source data, (2) unified standardized vocabularies, and (3) standardized statistical methods and tools to enable accurate, reproducible and efficient research. As the community drives its efforts towards those goals, there has been an increasing number of network studies initiated across different sites. Nevertheless, an exploratory stage of these studies may lack transparency. While the researchers operate those clinical concepts that are available at their instances, they cannot assess the availability of these concepts at other sites. This makes it hard to evaluate the number of data instances that can participate in a network study without modifying the study design, complicates initial feasibility assessment and lengthens the process.

## There is no established practice to share concept prevalence, which can be a part of data site characterization and can be coupled with the Themis initiative. The latter aims at creating the system of CDM conversion and maintenance policies as well as data sites certification process.

## In order to ensure interoperability and consistency of network study designs across all OMOP CDM databases, leverage existing data and facilitate cross-dataset research, we suggest launching a study that will investigate prevalence, heterogeneity, diversity, and granularity of clinical concept utilization across OHDSI partner data sites.

## Research Questions

The main research question of this study is:

*What are the prevalence, heterogeneity, diversity, and granularity of clinical concepts across different OMOP CDM datasets?*

The specific aims of this study are as follows:

1. To investigate the real-world terminology utilization across datasets with different provenance of the data (administrative claim data, electronic health records, clinical registry data, clinical trial data), country of origin and focus of clinical care (generic, oncology, surgery, etc).

2. To investigate the data loss related to OHDSI Vocabulary long-term changes related to different vocabulary version mappings.

3. To explore the patterns of clinical concept granularity and specificity that are specific to particular domains, the provenance of the data, country of origin and focus of clinical care.

# Research methods

## Study Design

We will accumulate standard clinical concepts along with their prevalence from the main OMOP CDM tables and break them into the groups based on their domain (Table 1). Here, we refer to standard concepts as the OMOP standard concepts, i.e., the concepts belonging to OMOP CDM Standardized Vocabularies with standard\_concept = ‘S’ (Full description is available at: <http://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary>). To address Aim 2, we will separately collect source\_concept\_ids along with frequency for the respective fields. The list can be extended as new tables or fields are introduced. The full up-to-date specification is available at GitHub: <https://github.com/OHDSI/CommonDataModel> .

|  |  |  |
| --- | --- | --- |
| Type of data | Group | Source field |
| Demographic data | Race | Person.race\_concept\_id |
| Ethnicity | Person.ethnicity\_concept\_id |
| Clinical data | Condition | Condition\_occurrence.condition\_concept\_id  Condition\_occurrence.condition\_type\_concept\_id |
| Drug | Drug\_exposure.drug\_concept\_id  Drug\_exposure.drug\_type\_concept\_id  Drug\_exposure. route\_concept\_id |
| Device | Device\_exposure.device\_concept\_id  Device\_exposure.device\_type\_concept\_id |
| Measurement | Measurement.measurement\_concept\_id  Measurement.measurement\_type\_concept\_id |
| Observation | Observation.observation\_concept\_id & observation.value\_as\_concept\_id (if exists)  Obervation.observation\_type\_concept\_id |
| Procedure | Procedure\_occurrence.procedure\_concept\_id  Procedure\_occurrence.procedure\_type\_concept\_id |
| Specimen | Specimen.specimen\_concept\_id  Specimen.specimen\_type\_concept\_id |
| Health economics data | Cost | Cost.cost\_concept\_id  Cost.cost\_type\_concept\_id |
| Payer plan period | Payer.payer\_concept\_id |
| Health system data | Visit | Visit\_occurrence.visit\_type\_concept\_id  Visit\_detail.visit\_detail\_type\_concept\_id |
| Care site | Care\_site.place\_of\_service\_concept\_id |
| Provider | Provider.specialty\_concept\_id |

We will generate and disseminate an R package that will run SQL queries on each site and produce descriptive statistics. The summary will consist of:

* Dataset identifier (name of the dataset)
* Database span (minimum observation\_period\_start\_date – maximum observation\_period\_end\_date)
* Number of patients in the dataset
* OMOP Vocabulary version
* Concept ID
* Count of occurrences of concepts in the dataset and the table where the concept was used

To protect patient privacy and reduce the risk of re-identification, each site will exclude data for any concept where the record count ≤ 10. Furthermore, each count will be randomized from the true record count by sampling from the Poisson distribution with λ = count. The Poisson is a probability distribution describing the probability of observing a given number of events occurring in a time interval given that the events occur at a given rate (λ) and independently of each other. By sampling from the Poisson distribution with λ = count, we perturb the true record count while preserving the mean occurrence rate as the true count.

Pseudo results tables 1,2 and 3:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Standard Concept | Site 1 | Site 2 |  | Site n |
| C1 | Pois(CountC1,Site1) | Pois(CountC1,Site2) | … | Pois(CountC1,SiteN) |
| C2 | … | … | … | … |
| … | … | … | … | … |
| Cn | Pois(CountCn,Site1) | Pois(CountCn,Site2) | … | Pois(CountCn,SiteN) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source Concept | Site 1 | Site 2 |  | Site n |
| C1 | Pois(CountC1,Site1) | Pois(CountC1,Site2) | … | Pois(CountC1,SiteN) |
| C2 | … | … | … | … |
| … | … | … | … | … |
| Cn | Pois(CountCn,Site1) | Pois(CountCn,Site2) | … | Pois(CountCn,SiteN) |

|  |  |  |
| --- | --- | --- |
| Standard Concept | Relationship\_id | Source Concept |
| StdC1 | Maps to | SrcC1 |
| StdC1 | Maps to value | SrcC1 |
| … | … | … |
| StdCn | Maps to | SrcCn |

The R package will be available at GitHub: <https://github.com/OHDSI/StudyProtocolSandbox/ConceptPrevalence> and the results will be available through the public instance of OHDSI Atlas (<http://www.ohdsi.org/web/atlas/#/datasources>) on Data Sources page.

## Data Sources

The analyses will be performed across a network of observational healthcare databases that have been transformed into OMOP CDM, version 5. The following databases will be included in this analysis:

* Columbia University Medical Center

## Use Cases

Besides our original question, the data generated by this study can serve multiple research purposes:

**1. Preliminary estimate of study feasibility. Table 1 can serve as the lookup table for identifying the number of patients available for each concept at each site. This information can guide investigators of recruitment the adequate number of patients with the condition of interest at each site.** The data generated by this study can be leveraged to perform an initial step of feasibility studies, particularly those that involve rare outcomes or exposures, demographic differences, etc. We expect that the generated data will facilitate network studies and reduce the time required to prepare network studies.

**2. Data quality improvement.** Table 2 shows the mapping from source concepts to standard concepts and their usage patterns. Information from this table can be used to understand the discrepancy in granularity and specificity of chosen concepts. Coupled with Themis initiative, the data can be used for data sources characterization and certification.

# Protection of Human Subjects

The study is collecting only de-identified data. Only aggregated results will be shared across the community. To protect patient privacy and reduce the risk of re-identification, each site will exclude data for any concept where the record count ≤ 10, and the true counts will be randomized by the Poisson distribution. Confidentiality of patient records will be maintained at all times.

# Study Results Dissemination

The aggregated study results will be posted on the OHDSI website after completion of the study. We plan to make the results available through public instance of OHDSI Atlas (<http://www.ohdsi.org/web/atlas/#/datasources>) on Data Sources page. At least one paper will be written and submitted for publication to a peer-reviewed scientific journal.