**Concept Heterogeneity in the OHDSI Network**

**Anna Ostropolets, MD1, Anthony Molinaro2, Frank DeFalco2, Jitendra Jonnagaddala, MIS, PhD3, Siaw-Teng Liaw, PhD, FACMI3, Hokyun Jeon4, Rae Woong Park, MD, PhD4,5, Matthew E. Spotnitz, MD, MPH1, Karthik Natarajan, PhD1, George Argyriou6, Kristin Kostka, MPH6, Christian Reich, MD, PhD6, Chunhua Weng, PhD1, Patrick Ryan, PhD1,2**

***1 Columbia University, New York, USA; 2 Janssen Pharmaceuticals, New Jersey, USA; 3 University of New South Wales Sydney, Australia; 4 Ajou University Graduate School of Medicine, Suwon, Gyeonggi-do, Republic of Korea; 5Ajou University School of Medicine, Suwon, Gyeonggi-do, Republic of Korea, 6 IQVIA, Cambridge, MA, USA***

**Background**

When conducting studies within the OHDSI network, researchers usually operate those clinical concepts that are available at their instances without assessing 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, complicates initial feasibility assessment and lengthens the process**.** Moreover, there is little knowledge about the heterogeneity of the data across partner data sites and how it may impact the results of the studies. With disparate coding practices, vocabulary granularity and utilization, it is unclear to what extent a unified approach to cohort creation can produce the same results in different datasets. This study that will investigate heterogeneity, diversity, and granularity of clinical concept across OHDSI partner data sites to ensure the possibility of consistent clinical data representation across OMOP CDM databases, leverage existing data, and facilitate cross-dataset research.

**Materials and methods**

We collected standard and source concepts (\*\_concept\_id and \*\_source\_concept\_id) along with their frequency from the main OMOP CDM version 5 tables (condition\_occurrence, procedure\_occurrence, drug\_exposure, device\_exposure, measurement, observation) and group them according to the OMOP CDM table. We obtained the data from seventeen datasets (IBM\_MDCR, MDCD and CCAE; NHIS, MIMIC3, PREMIER, JMDC, CUMC, Ajou university, OPTUM EXTENDED SES and DOD, OPTUM PANTHER, OpenClaims, HCUP, Hospital, ePBRN and AmbEMR). To study the utilization of the concepts in the OHDSI network, we analyzed the distribution of the unique and overlapping concepts from the main OMOP Vocabulary domains (Condition, Measurement, Observation, Procedure, Drug) across different datasets. We obtained only condition concepts from Ajou University dataset did not contain measurements, observations, procedures and drugs and NHIS Korea dataset – drugs and procedures.

**Results**

We collected the data from the US, Korea, Australia, and Japan. The data came mainly from electronic health records collected in large teaching hospitals and administrative claims data. The size of the datasets varied greatly, with the average number of 13 383±8 734 unique concepts per table and total number of unique concepts varying from 2 274 in MIMIC3 to 156 499 in OPTUM PANTHER.

We found high variability of the concepts utilized in the OHDSI network with the significant number of concepts unique to one dataset (Figure 1). Condition was the least heterogeneous domain with the highest number of overlapping concepts across datasets, followed by Procedure and Drug domains. While conditions and procedures are usually coded using a limited number of vocabularies, measurements and observations do not have established coding practices, which may explain high variability of the concepts in these domains.

**Figure 1. Distribution of the overlapping concepts across OHDSI partner datasets.**

Surprisingly, some of the frequent concepts associated with higher number of records in OMOP CDM can be found only in individual datasets (Figure 2). Nevertheless, most of them were found in six and more datasets, which shows that the data sites share the most common conditions, procedures and drugs.

****

**Figure 2. Distribution of frequency of the concepts across OHDSI partner datasets.**

When analyzing the granularity of concepts in Condition domain (Supplementary materials), we found that concepts related to diabetes type II, urinary tract infections, ischemic heart disease, depressive disorder, attention deficit disorder, rheumatoid arthritis and chronic kidney disorder were mostly found in 50-75% of datasets.

**Conclusion**

In this work, we present our preliminary results from the study of concept heterogeneity and variability across the OHDSI network. We found that datasets share only a limited number of concepts, while a significant part of them remains unique to a dataset. Conditions and procedures appeared to be the most conservative domains with the least heterogeneity, while measurements and observations differed significantly across sites. We found most concepts don't appear in most databases, so all network studies require use of comprehensive concept sets when defining cohorts to represent the diversity of data across the community. Along with the growing knowledge base, it will have the potential to facilitate network studies, make them more transparent, short, and effective.

**References**

1. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, Van Der Lei J. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics. 2015;216:574.