COLUMBIA
COLUMBIA UNIVERSITY
DEPARTMENT OF
BIOMEDICAL INFORMATICS

Data source heterogeneity and its influence on phenotyping in distributed data networks

Collecting the knowledge about concept utilization

Background

We can use the rich data in electronic health records and claims data to conduct drug surveillance and drug effectiveness studies, investigate treatment pathways and predict patient outcomes. As observational data is not collected for research purposes and therefore may be inaccurate and sparse, we need to develop executable algorithms to find patients of interest, so called phenotype algorithms. When such algorithms are applied to multiple data sources, we can leverage diverse and large patient populations to generate more reliable evidence. On the other hand, creating reliable and comprehensive phenotype algorithms in distributed data networks is especially hard as differences in patient representation and data source heterogeneity must be taken into account.

Concept Prevalence study

To investigate data source heterogeneity, we collected the clinical codes (condition, procedure codes, lab tests etc.) and their frequency of occurrence from 22 electronic health record and administrative claims datasets from the US, Korea, Australia and Japan. All data sources were mapped to the OMOP Common Data Model, both the structure (data) and the content (mapping of source vocabularies like ICD10-CM to the OMOP Standardized Vocabularies).

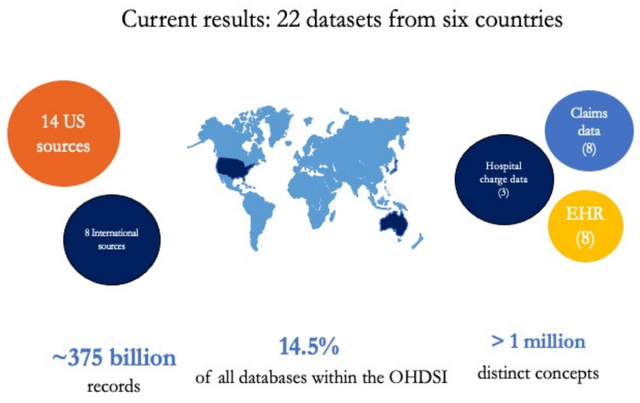


Figure 1. Concept Prevalence study

In general, non-US data sources had less granular (broad) concepts, compared to the US data sources. EHR data sources from primary and secondary care practices appeared to be less granular, while administrative claims data, hospital charge data and EHR data from large tertiary care hospitals were more granular.

Examining data source heterogeneity

We found that the data sources are highly heterogeneous (Figure 2), with most of the concepts appearing only in some of the data sources. A high number of lab test codes, procedure codes and condition codes were unique to one data source and could not be found in the others (red rectangle).
This challenges conventional approaches to phenotyping such as using administrative claims concepts, concepts from existing literature or exploring concepts at a local patient data instance.

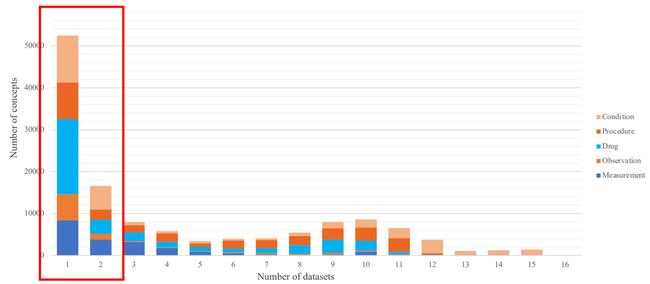


Figure 2. Distribution of the overlapping concepts across the OHDSI network.

For example, a phenotype for attention deficit disorder in kids cannot simply use a SNOMED code 192127007 "Child attention deficit disorder" because it is absent in most of the data sources (Figure 3). On contrary, attention deficit hyperactivity disorder is more commonly used in the network.

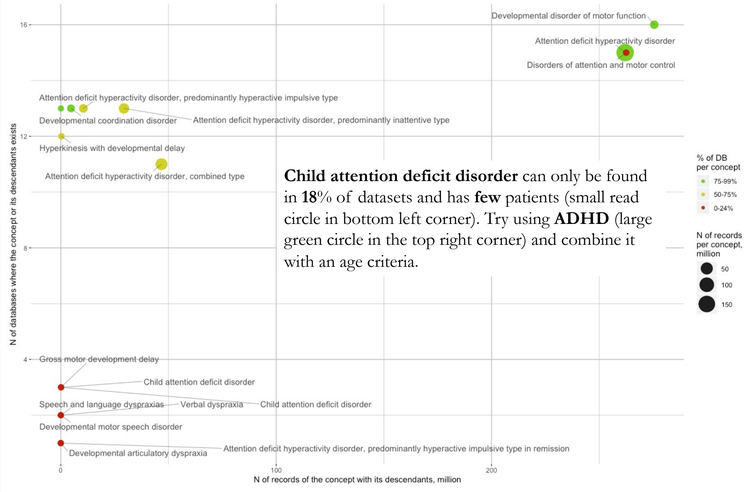


Figure 3. Clinical code utilization for attention deficit disorder in the OHDSI Network.

Using the knowledge in phenotyping

The knowledge about clinical code utilization across the network can guide us in selecting code sets for identifying patients of interest. We developed PHenotype Observed Entity Baseline Endorsements (PHOEBE) - a tool for creating and examining concept sets.

PHOEBE enables researchers select the initial concept for a concept set representing their clinical idea and iteratively create a comprehensive set of codes that would work across the network (Figure 4).

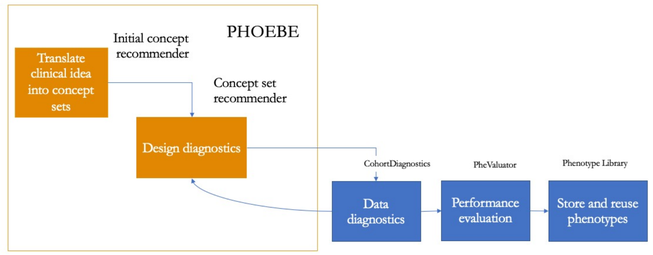
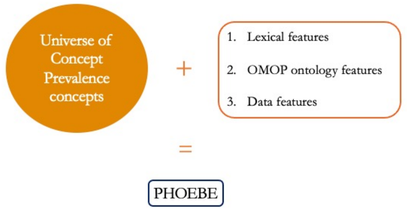


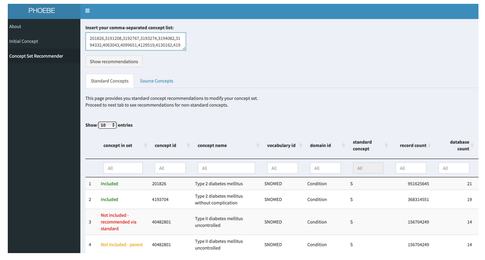
Figure 4. PHOEBE and its place in the OHDSI phenotyping framework



PHOEBE uses similarity features (lexical: substring matching, synonym matching, Levenstein distance; semantic: ontology alignment, path to common ancestor) to find recommendations in the OMOP Vocabularies and organizes them based on their frequency in the network.

Figure 5. Methods used in PHOEBE

When used for studies in the network, cohort definitions constructed with PHOEBE identify more patients and capture them earlier in the course of the disease



It is now used by multiple individuals and organizations in OHDSI and is publicly available at https://data.ohdsi.org/PHOEBE/.

Anna Ostropolets, PhD Candidate
Department of Biomedical Informatics, George Hripcsak lab
and the OHDSI community