Best Practices for Creating the Standardized Content of an Entry in the OHDSI Phenotype Library

James Weaver¹,², Aaron Potvien³, Joel Swerdel¹,², Erica A Voss¹,², Laura Hester¹,², Azza Shoaibi¹,², Patrick B Ryan¹,², Jon Duke²,³

¹ Janssen Research and Development, LLC, Raritan, NJ, ²Observational Health Data Sciences and Informatics (OHDSI), New York, NY, ³Georgia Tech, ⁴Columbia University

Background

- Reliable evidence generated from observational data should be repeatable, reproducible, replicable, generalizable, and robust.
- Standardizing the evidence generation process to enable consistent analyses across disparate data sources advances these aims.
- A necessary input to an evidence generating process is a set of patients of who share an observable, clinical state of health—patients of the same phenotype.
- The goal of the OHDSI phenotype library is to catalogue computable phenotypes and provide the necessary information for researchers to make decisions regarding their appropriate use.

Book

- Each phenotype is represented as a book in the phenotype library.
- The book title is a simple label of the phenotype.
- The introductory section must include a complete biological/clinical description of what the health state entails, disease etiology, and the intended use of the phenotype for research purposes.
- A complete book must include one or more chapters.

Chapter

- A chapter documents an attempt to identify and represent the phenotype in an observational database for a specific purpose.
- Given the variety and limitations of observational data, multiple approaches to identifying members of a phenotype are often necessary—each approach is documented in a chapter.
- A chapter must include a cohort definition, characterization results in 1 or more databases, and performance evaluation results from 1 or more databases.
- A cohort definition is a computationally transportable heuristic or probabilistic set of instructions for patient identification.

Characterization

- Implementing the cohort definition returns a cohort, a set of 0 or more patients who satisfy the definition for a period of time.
- For each database in which a cohort is built, characterization results will be generated as a set of artifacts for assessing occurrence and face validity.
- Occurrence is reported as a time series plot of the incidence proportion per 1000 persons of cohort entry by year further stratified by age and gender (Figure 1).
- Characterization results allow face validity assessment will be reported as a univariate summary table, easily interpreted by a human reader.
- The summary table will include counts and proportions (using the database population and/or cohort population as the denominator) for demographics, comorbidities, and past and concurrent medications (Table 1).

Evaluation

- For each database in which a cohort is built, misclassification representing the difference between true phenotype membership and those identified by the cohort definition must be reported.
- The evaluation will include standard diagnostic counts (i.e. true-positives, false-positives, true-negatives, false-negatives) and performance metrics (e.g. sensitivity, specificity, positive predictive value).
- Performance metrics can be computed by various evaluation methods which must be fully described.

Table 1. Example univariate summary measures for demographic characteristics and conditions prior to or on index date for ischemic stroke patients in the IBM MarketScan Commercial database. Conditions reported as MedDRA preferred terms. Database characterization results apply to 1 chapter of an entry in the phenotype library.

A book in the phenotype library includes 1 or more chapters.

Each chapter includes 1 cohort definition with characterization and evaluation results from 1 or more databases.

Figure 1. Characterization example reporting temporal stability of one cohort definition in one database. Incidence proportion of ischemic stroke/1000 persons by year stratified by gender and age in the IBM MarketScan Commercial database. The yearly incidence proportion value was calculated as number of first ischemic stroke events in a given year divided by the number of patients with ≥1 days of enrollment in the same year with denominator right censored at time of event.