Best practices for creating the standardized content of an entry in the OHDSI phenotype library

James Weaver[1,2], Aaron Potvien[2,3], Joel Swerdel[1,2], Erica A Voss[1,2], Laura Hester[1,2], Azza Shoaibi[1,2], Patrick B Ryan[1,2,4], Jon Duke[2,3]
[1] Janssen Research and Development, LLC, Raritan, NJ, [2] - OHDSI Collaborators, Observational Health Data Sciences and Informatics (OHDSI), New York, NY, [3]Georgia Tech, [4]Columbia University

A primary goal of observational research is to generate reliable evidence from real-world data to inform decision making for patient care, where reliable evidence is repeatable, reproducible, replicable, generalizable, and robust[1, 2]. Generating characterization, patient-level prediction, or population-level effect estimation evidence requires a sequence of analytic processes and inputs, which have been standardized to enable consistent analyses across disparate data sources[3, 4]. Necessary inputs to these processes are the data of patients who share observed clinical states of health. In the context of observational database research and informatics parlance, an observed clinical state of health that persists for a defined duration of time is referred to as a phenotype[5].

The Observational Health Data Sciences and Informatics (OHDSI) phenotype library is intended to catalogue phenotypes that are of common interest to observational research. Given the variety and limitations of observational data, there may be multiple approaches or implementations for identifying members of a phenotype. In the analogy that the phenotype library is a collection of phenotype books, a chapter in a phenotype book is a cohort definition intended to identify patients of the phenotype. We propose that the best practices for an entry in the phenotype library is that each entry adhere to standards for 1) phenotype title section, 2) cohort definition, 3) characterization, and 4) evaluation. Following the analogy, a phenotype book will include a title and 1 or more chapters, and each chapter will contain a cohort definition with characterization and evaluation results for 1 or more databases.

The phenotype title section consists of a simple label of the phenotype and a complete clinical/biological description of what the health state entails. The title section will also include the intended use(s) of the phenotype for generating evidence. It may represent an exposure of interest, a study outcome or endpoint, or otherwise serve as input in a characterization, prediction, or estimation study in a single database or across a disparate database network. In the library analogy, the title section is the title and preface to a single book comprised of 1 or more chapters. The phenotype title is database-agnostic.

Each book must include 1 or more chapters, each chapter is an attempt to represent the phenotype in a database. A chapter must include a) a cohort definition, b) characterization results in 1 or more databases, and c) evaluation results from 1 or more databases. The cohort definition is a computationally transportable heuristic or probabilistic set of instructions that attempts to represent the phenotype given a common data structure and the phenotype's intended use. In addition to computer-readable code, the cohort definition must include a human-readable technical specification. The cohort definition is also database-agnostic.

Implementing the cohort definition against data source returns a cohort, defined as a set of 0 or more patients who satisfy 1 or more inclusion criteria for a duration 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 can be reported as a time series plot of the incidence proportion per 1000 persons of cohort entry by year, further stratified by age, gender, and age by gender. Characterization results to allow face validity assessment will be reported as a univariate summary table, easily interpreted by a human reader. Such a table will include counts and proportions (using the database population as the denominator) for demographics, comorbidities, and past and concomitant medications.

Lastly, 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). These operating characteristics can be computed by evaluation different methods and those used must be described. For example, chart adjudication details such as which and how many charts were sampled, and who reviewed them must be provided. Software for evaluating cohort definition operating characteristics at scale can also be used -available (https://github.com/OHDSI/PheValuator)-. Where these systems are employed, the input specifications for the noisy positive and negative labels for training the probabilistic gold standard must be reported.

The following is an example entry into the phenotype library for a single phenotype (1 book) with a single cohort definition (1 chapter) with characterization and evaluation results for 1 database. Table 1 reports database-agnostic components for a phenotype entry for ischemic stroke. Figure 1 depicts a time series plot of the yearly incidence proportion per 1000 persons of phenotype cohort entry by gender and age in the IBM MarketScan Commercial Database (CCAE) database. Table 3 reports a subset of the summary features for the cohort definition, which includes demographic characteristics and comorbid conditions. Operating characteristics generated using PheValuator report sensitivity = 0.66, specificity = 0.99, and positive predictive value = 0.36.

Adoption of a standardized framework for cataloguing phenotypes has the potential to further advance

observational science by increasing researcher awareness of the operating characteristics of the inputs to analytic methods employed to generate evidence.

Table 1. Database-agnostic components of an entry in the phenotype library

Component	Sub-component	Entry
Phenotype title section	Title	Ischemic stroke
	Clinical/biological definition	Patients with the sudden death of brain cells due to oxygen deprivation caused by reduced or blocked blood flow to the brain
	Research intent	Prediction study outcome, estimation study outcome, indication for exposure cohort definition; for use across database network
Cohort definition	Computer-readable description	LEGEND_ishemic_stroke.json
	Human-readable description	1x ischemic stroke condition diagnosis record on or the day before an inpatient or ER visit

Figure 1. 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 at least 1 days of enrollment in the same year with denominator right censored at time of event.

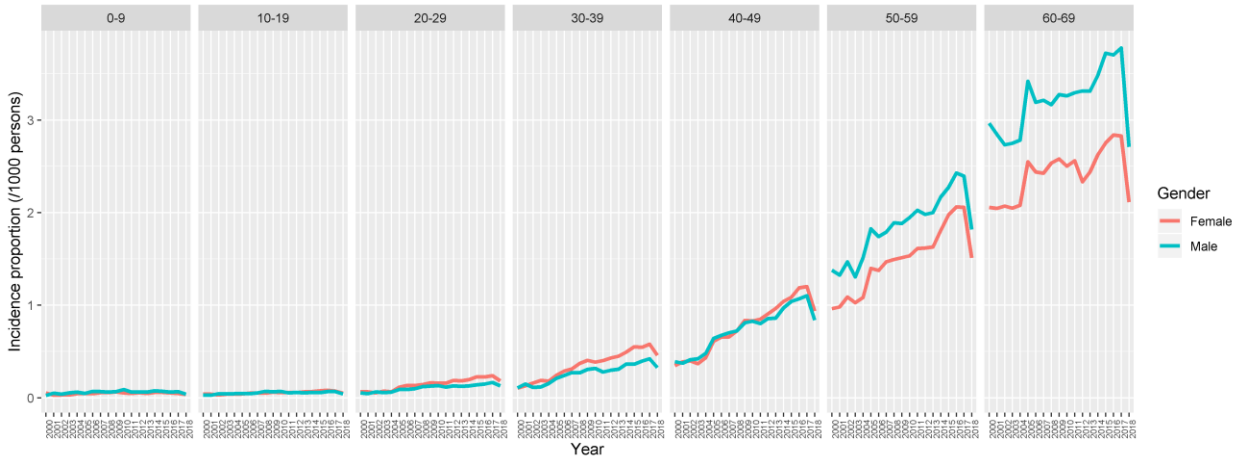

Table 1. Example univariate summary measures for demographic characteristics and conditions prior to or on index date for ischemic stroke patients in the IBM MarketScan Commercial database. Conditions reported as MedDRA preferred terms. Database characterization results apply to 1 chapter of an entry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

Statistic	Database	PL060_Ischemic stroke	Comorbidity	Database	PL060_Ischemic stroke
OVERALL COUNT	143,517,368	357,935	General symptom	56.02%	100.00%
ALL TIME OBSERVED (/1000 PERSON-YEARS)	356136.32	1854.2	Investigation abnormal	76.94%	100.00%
TIME AFTER INDEX (/1000 PERSON-YEARS)	356136.32	721.36	Nervous system disorder	14.26%	100.00%
TIME BEFORE INDEX (/1000 PERSON-YEARS)	0	1132.84	Radiculopathy	14.26%	100.00%
TIME FROM INDEX TO COHORT END (/1000 PERSON-YEARS)	356136.32	6.86	Injury	26.99%	97.48%
YEAR OF INDEX: 2000	3,199,101 (2.23%)	1,618 (0.45%)	Encephalopathy	5.53%	97.33%
YEAR OF INDEX: 2001	2,638,669 (1.84%)	2,828 (0.79%)	Cerebral infarction	0.35%	94.42%
YEAR OF INDEX: 2002	5,408,216 (3.77%)	5,058 (1.41%)	Cardiovascular disorder	24.06%	93.24%
YEAR OF INDEX: 2003	7,554,344 (5.26%)	7,403 (2.07%)	Phlebosclerosis	24.06%	93.24%
YEAR OF INDEX: 2004	7,279,965 (5.07%)	10,002 (2.79%)	Soft tissue disorder	47.48%	90.51%
YEAR OF INDEX: 2005	6,252,771 (4.36%)	14,105 (3.94%)	Pain	42.23%	80.34%
YEAR OF INDEX: 2006	8,077,993 (5.63%)	14,696 (4.11%)	Angiopathy	11.84%	77.83%
YEAR OF INDEX: 2007	6,187,207 (4.31%)	16,398 (4.58%)	Ill-defined disorder	31.16%	73.97%
YEAR OF INDEX: 2008	9,881,295 (6.89%)	20,063 (5.61%)	Respiratory disorder	47.03%	73.50%
YEAR OF INDEX: 2009	10,779,128 (7.51%)	24,446 (6.83%)	Dyspnoea	46.95%	73.44%
YEAR OF INDEX: 2010	12,952,893 (9.03%)	26,733 (7.47%)	Cerebral artery occlusion	0.29%	72.98%
YEAR OF INDEX: 2011	12,286,937 (8.56%)	31,167 (8.71%)	Musculoskeletal disorder	39.86%	71.26%
YEAR OF INDEX: 2012	10,774,596 (7.51%)	31,953 (8.93%)	Metabolic disorder	21.86%	70.43%
YEAR OF INDEX: 2013	10,217,449 (7.12%)	27,538 (7.69%)	Plasma protein metabolism disorder	21.80%	70.41%
YEAR OF INDEX: 2014	9,024,684 (6.29%)	31,612 (8.83%)	Cerebral thrombosis	0.27%	69.26%
YEAR OF INDEX: 2015	5,658,827 (3.94%)	24,900 (6.96%)	Hypertension	14.55%	64.37%
YEAR OF INDEX: 2016	5,630,905 (3.92%)	25,587 (7.15%)	Essential hypertension	14.25%	64.06%
YEAR OF INDEX: 2017	6,263,040 (4.36%)	24,331 (6.80%)	Enzyme abnormality	20.50%	63.18%
YEAR OF INDEX: 2018	3,449,348 (2.40%)	17,497 (4.89%)	Blood test abnormal	20.38%	63.06%
GENDER: FEMALE	73,431,874 (51.17%)	177,024 (49.46%)	Gastrointestinal disorder	37.32%	61.39%
GENDER: MALE	70,085,494 (48.83%)	180,911 (50.54%)	Cerebrovascular disorder	1.10%	58.35%
MEAN AGE AT INDEX	31.19	52.39	Skin disorder	33.88%	57.80%
ST DEV AGE AT INDEX	18.1	11.26	Mental disorder	19.54%	55.04%
AGE DECILE: 00-09	21,893,007 (15.25%)	3,208 (0.90%)	Arthropathy	28.17%	54.19%
AGE DECILE: 10-19	20,179,410 (14.06%)	3,843 (1.07%)	Arthropod-borne disease	37.51%	51.47%
AGE DECILE: 20-29	26,009,072 (18.12%)	9,984 (2.79%)	Viral infection	37.51%	51.47%
AGE DECILE: 30-39	24,032,845 (16.75%)	24,808 (6.93%)	Hyperlipidaemia	15.83%	51.20%
AGE DECILE: 40-49	23,123,965 (16.11%)	64,861 (18.12%)	Lipids abnormal	15.83%	51.20%
AGE DECILE: 50-59	20,114,463 (14.02%)	141,277 (39.47%)	Lipids increased	15.83%	51.20%
AGE DECILE: 60-69	8,160,068 (5.69%)	109,807 (30.68%)	Mediastinal disorder	7.64%	50.92%
AGE DECILE: 70-79	3,441 (0.00%)	94 (0.03%)	Cardiac disorder	7.55%	50.54%
AGE DECILE: 80-89	890 (0.00%)	38 (0.01%)	Urogenital disorder	23.94%	50.15%
AGE DECILE: 90-99	202 (0.00%)	15 (0.00%)	Connective tissue disorder	24.63%	49.63%

References

1. Plesser, H.E., *Reproducibility vs. Replicability: A Brief History of a Confused Terminology.* Front Neuroinform, 2017. **11**: p. 76.
2. Goodman, S.N., D. Fanelli, and J.P.A. Ioannidis, *What does research reproducibility mean?* Science Translational Medicine, 2016. **8**(341): p. 341ps12-341ps12.
3. Voss, E.A., et al., *Feasibility and utility of applications of the common data model to multiple, disparate observational health databases.* J Am Med Inform Assoc, 2015. **22**(3): p. 553-64.
4. Platt, R.W., et al., *How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias.* Pharmacoepidemiol Drug Saf, 2019.
5. Hripcsak, G. and D.J. Albers, *High-fidelity phenotyping: richness and freedom from bias.* J Am Med Inform Assoc, 2017.