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]Observational Health Data Sciences and Informatics (OHDSI), New York, NY, [3]Georgia Tech, [4]Columbia 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

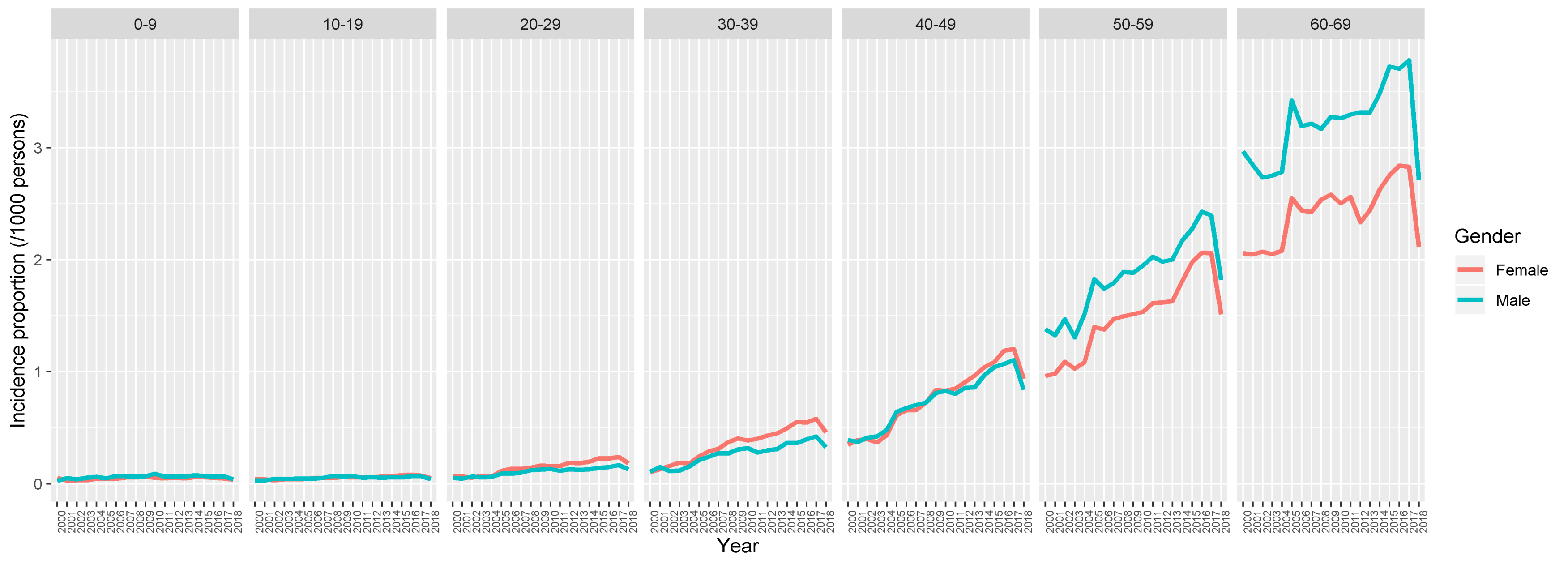
Chapter

- 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
- The 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 concomitant medications (Table 1)

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.



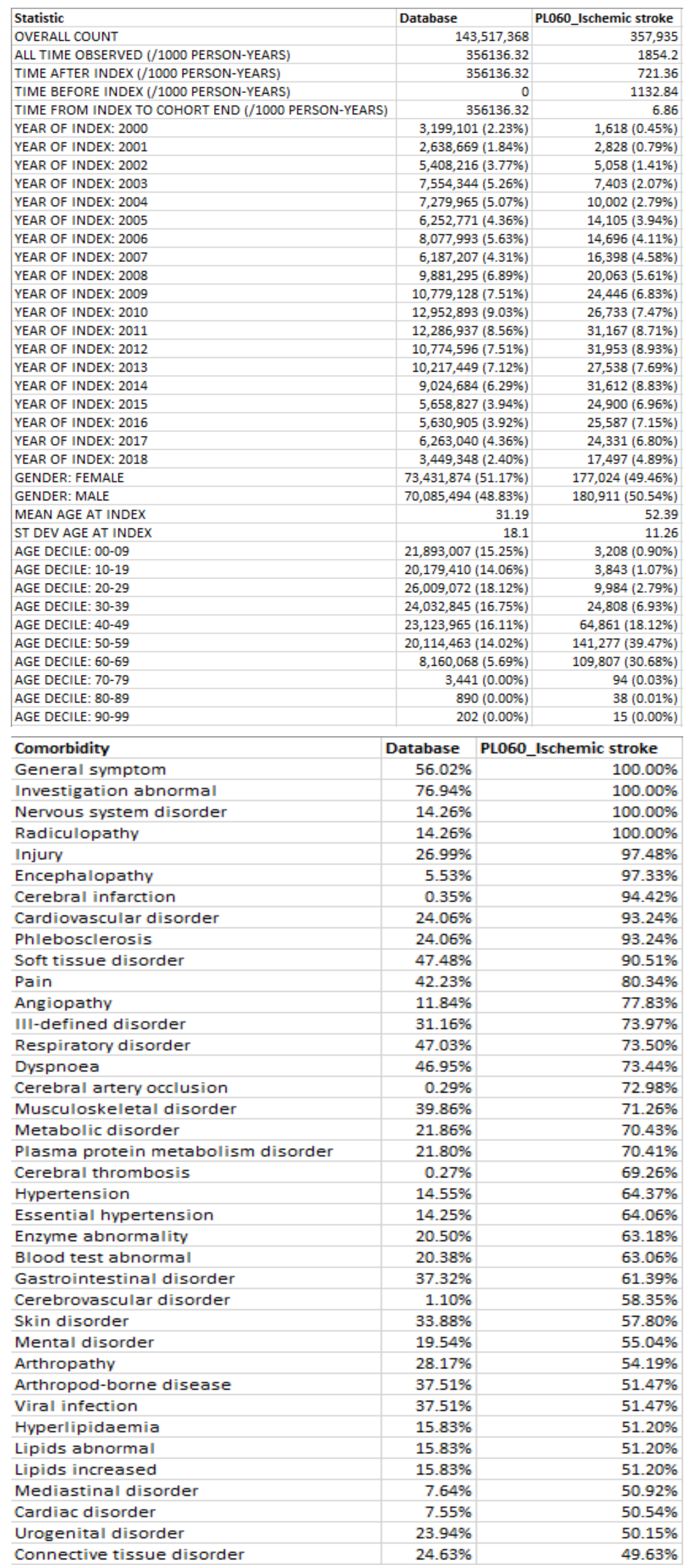
Take a picture, it'll last longer.

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.

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.



Statistic	Database	PL060_Ischemic stroke
OVERALL COUNT	143,517,368	357,935
ALL TIME OBSERVED (/1000 PERSON-YEARS)	356136.32	1854.2
TIME AFTER INDEX (/1000 PERSON-YEARS)	356136.32	721.36
TIME BEFORE INDEX (/1000 PERSON-YEARS)	0	1132.84
TIME FROM INDEX TO COHORT END (/1000 PERSON-YEARS)	356136.32	6.86
YEAR OF INDEX: 2000	3,199,101 (2.23%)	1,618 (0.45%)
YEAR OF INDEX: 2001	2,638,669 (1.84%)	2,828 (0.79%)
YEAR OF INDEX: 2002	5,408,216 (3.77%)	5,058 (1.41%)
YEAR OF INDEX: 2003	7,554,344 (5.26%)	7,403 (2.07%)
YEAR OF INDEX: 2004	7,279,965 (5.07%)	10,002 (2.79%)
YEAR OF INDEX: 2005	6,252,771 (4.36%)	14,105 (3.94%)
YEAR OF INDEX: 2006	8,077,993 (5.63%)	14,696 (4.11%)
YEAR OF INDEX: 2007	6,187,207 (4.31%)	16,398 (4.58%)
YEAR OF INDEX: 2008	9,881,295 (6.89%)	20,063 (5.61%)
YEAR OF INDEX: 2009	10,779,128 (7.51%)	24,446 (6.83%)
YEAR OF INDEX: 2010	12,952,893 (9.03%)	26,733 (7.47%)
YEAR OF INDEX: 2011	12,286,937 (8.56%)	31,167 (8.71%)
YEAR OF INDEX: 2012	10,774,596 (7.51%)	31,953 (8.93%)
YEAR OF INDEX: 2013	10,217,449 (7.12%)	27,538 (7.69%)
YEAR OF INDEX: 2014	9,024,684 (6.29%)	31,612 (8.83%)
YEAR OF INDEX: 2015	5,658,827 (3.94%)	24,900 (6.96%)
YEAR OF INDEX: 2016	5,630,905 (3.92%)	25,587 (7.15%)
YEAR OF INDEX: 2017	6,263,040 (4.36%)	24,331 (6.80%)
YEAR OF INDEX: 2018	3,449,348 (2.40%)	17,497 (4.89%)
GENDER: FEMALE	73,431,874 (51.17%)	177,024 (49.46%)
GENDER: MALE	70,085,494 (48.83%)	180,911 (50.54%)
MEAN AGE AT INDEX	31.19	52.39
ST DEV AGE AT INDEX	18.1	11.26
AGE DECILE: 00-09	21,893,007 (15.25%)	3,208 (0.90%)
AGE DECILE: 10-19	20,179,410 (14.06%)	3,843 (1.07%)
AGE DECILE: 20-29	26,009,072 (18.12%)	9,984 (2.79%)
AGE DECILE: 30-39	24,032,845 (16.75%)	24,808 (6.93%)
AGE DECILE: 40-49	23,123,965 (16.11%)	64,861 (18.12%)
AGE DECILE: 50-59	20,114,463 (14.02%)	141,277 (39.47%)
AGE DECILE: 60-69	8,160,068 (5.69%)	109,807 (30.68%)
AGE DECILE: 70-79	3,441 (0.00%)	94 (0.03%)
AGE DECILE: 80-89	890 (0.00%)	38 (0.01%)
AGE DECILE: 90-99	202 (0.00%)	15 (0.00%)

Comorbidity	Database	PL060_Ischemic stroke
General symptom	56.02%	100.00%
Investigation abnormal	76.94%	100.00%
Nervous system disorder	14.26%	100.00%
Radiculopathy	14.26%	100.00%
Injury	26.99%	97.48%
Encephalopathy	5.53%	97.33%
Cerebral infarction	0.35%	94.42%
Cardiovascular disorder	24.06%	93.24%
Phlebosclerosis	24.06%	93.24%
Soft tissue disorder	47.48%	90.51%
Pain	42.23%	80.34%
Angiopathy	11.84%	77.83%
Ill-defined disorder	31.16%	73.97%
Respiratory disorder	47.03%	73.50%
Dyspnoea	46.95%	73.44%
Cerebral artery occlusion	0.29%	72.98%
Musculoskeletal disorder	39.86%	71.26%
Metabolic disorder	21.86%	70.43%
Plasma protein metabolism disorder	21.80%	70.41%
Cerebral thrombosis	0.27%	69.26%
Hypertension	14.55%	64.37%
Essential hypertension	14.25%	64.06%
Enzyme abnormality	20.50%	63.18%
Blood test abnormal	20.38%	63.06%
Gastrointestinal disorder	37.32%	61.39%
Cerebrovascular disorder	1.10%	58.35%
Skin disorder	33.88%	57.80%
Mental disorder	19.54%	55.04%
Arthropathy	28.17%	54.19%
Arthropod-borne disease	37.51%	51.47%
Viral infection	37.51%	51.47%
Hyperlipidaemia	15.83%	51.20%
Lipids abnormal	15.83%	51.20%
Lipids increased	15.83%	51.20%
Mediastinal disorder	7.64%	50.92%
Cardiac disorder	7.55%	50.54%
Urogenital disorder	23.94%	50.15%
Connective tissue disorder	24.63%	49.63%

Conclusion

- A standardized framework for cataloguing phenotypes has the potential to advance observational science by increasing researcher awareness of the operating characteristics of the inputs to analytic methods employed to generate evidence

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS