Ontologizing Health Systems Data at Scale: Making Translational Discovery a Reality

SUPPLEMENTARY MATERIAL

Tiffany J. Callahan, Adrianne L. Stefanski, Jordan M. Wyrwa, Chenjie Zeng, Anna Ostropolets, Juan M. Banda, William A. Baumgartner Jr., Richard D. Boyce, Elena Casiraghi, Ben D. Coleman, Janine H. Collins, Sara J. Deakyne-Davies, James A. Feinstein, Melissa A. Haendel, Asiyah Y. Lin, Blake Martin, Nicolas A. Matentzoglu, Daniella Meeker, Justin Reese, Jessica Sinclair, Sanya B. Taneja, Katy E. Trinkley, Nicole A. Vasilevsky, Andrew Williams, Xingman A. Zhang, Peter N. Robinson, Patrick Ryan, George Hripcsak, Tellen D. Bennett, Lawrence E. Hunter, Michael G. Kahn

Table of Contents

Supplementary Table 1: Paper Acronyms and Concept Definitions.

Term	Definition
Acronyms	
ACMG	American College of Medical Genetics and Genomics
AoU	All of Us Research Program
CDM	Common Data Model
ChEBI	Chemical Entities of Biological Interest
CL	Cell Ontology
CUI	Concept Unique Identifier
dbXRef	Database Cross-Reference
EHR	Electronic Health Record
eMERGE	Electronic Medical Records and Genomics
FBN1	Fibrillin 1
HPO	Human Phenotype Ontology
ICD	International Classification of Diseases
LOINC	Logical Observation Identifiers, Names and Codes
MEN1	Menin 1
Mondo	Mondo Disease Ontology
NCBITaxon	National Center for Biotechnology Information Organismal Taxonomy
NF2	Moesin-Ezrin-Radixin Like (MERLIN) Tumor
OHDSI	Observational Health Data Sciences and Informatics
OBO	Open Biological and Biomedical Ontology
OMIM	Online Mendelian Inheritance in Man
OMOP	Observational Medical Outcomes Partnership
PEDSnet	National Pediatric Learning Health System
PheRS	Phenotype Risk Score
PRO	Protein Ontology
RET	Ret Proto-Oncogene
SDHAF2	Succinate Dehydrogenase Complex Assembly Factor 2
SDHB	Succinate Dehydrogenase Complex Subunit B
SDHC	Succinate Dehydrogenase Complex Subunit C
SNOMED-CT	Systematized Nomenclature of Medicine -- Clinical Terms
TGFBR1	Transforming Growth Factor Beta Receptor 1
TSC1	Tuberous Sclerosis Complex Subunit 1

Term	Definition
TSC2	Tuberous Sclerosis Complex Subunit 2
Uberon	Uber-Anatomy Ontology
UMLS	Unified Medical Language System
VO	Vaccine Ontology
Concepts	
Standard Concepts Used in Practice (Data Wave 1)	All standard OMOP concepts used at least once in clinical practice
Standard Concepts Not Used in Practice (Data Wave 2)	All standard OMOP concepts not used in clinical practice
OMOP Standard Condition Occurrence Vocabulary	SnomedCT Release 20180131
OMOP Standard Drug Exposure Ingredient Vocabulary	RxNorm Full 20180507
OMOP Standard Measurement Vocabulary	LOINC 2.64
OBO Foundry Ontologies mapped to OMOP Conditions	HPO, Mondo
OBO Foundry Ontologies mapped to OMOP Drug Ingredients	ChEBI, NCBITaxon, PRO, VO
OBO Foundry Ontologies mapped to OMOP Measurements	ChEBI, CL, HPO, NCBITaxon, PRO, Uberon

Supplementary Table 2: OMOP2OBO Framework and Evaluation Resources.

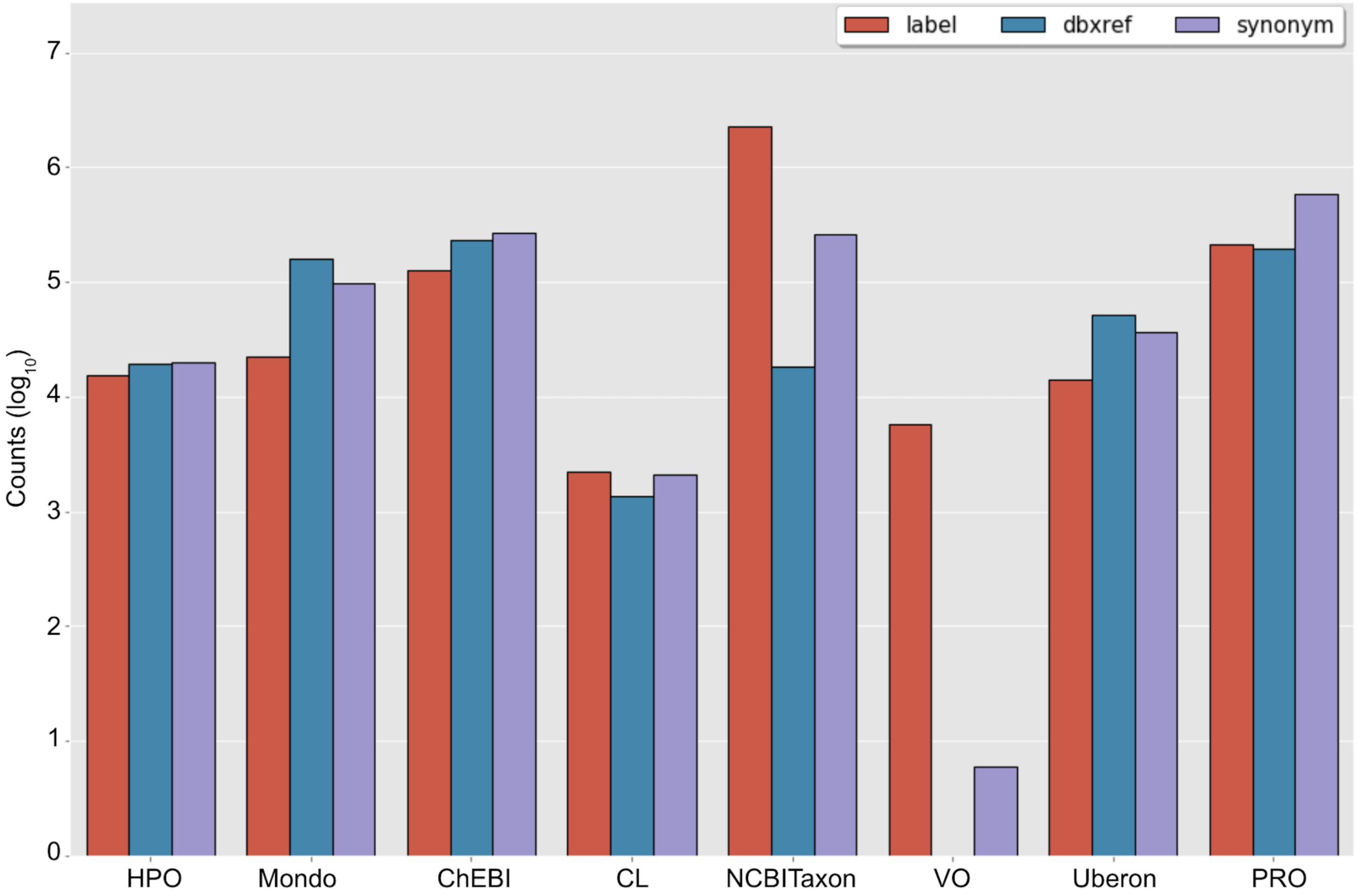
Resource	URL
OMOP2OBO Resources	
PyPI Package	https://pypi.org/project/omop2obo/
GitHub Repository	https://github.com/callahantiff/OMOP2OBO
Project Wiki	https://github.com/callahantiff/OMOP2OBO/wiki
Mapping Dashboard	http://tiffanycallahan.com/OMOP2OBO_Dashboard/
Zenodo Community	https://zenodo.org/communities/omop2obo
Condition Occurrence Mappings (v1)	https://doi.org/10.5281/zenodo.6774363
Drug Exposure Ingredient Mappings (v1)	https://doi.org/10.5281/zenodo.6774401
Measurement Mappings (v1)	https://doi.org/10.5281/zenodo.6774443
Accuracy Evaluation	https://github.com/callahantiff/OMOP2OBO/wiki/Accuracy
Generalizability Evaluation	https://github.com/callahantiff/OMOP2OBO/wiki/Generalizability
Mapping Resources	
OMOP CDM V5.3	https://ohdsi.github.io/CommonDataModel/cdm53.html
OHDSI Athena	https://athena.ohdsi.org/
UMLS 2020AA Release Date	https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html#2020AA
LOINC2HPO Annotations	https://github.com/monarch-initiative/loinc2hpo/annotations.tsv
OHDSI Concept Prevalence Study	https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ConceptPrevalence
OBO Foundry Ontologies	
ChEBI	http://purl.obolibrary.org/obo/chebi.owl
CL	http://purl.obolibrary.org/obo/cl.owl
HPO	http://purl.obolibrary.org/obo/hp.owl
Mondo	http://purl.obolibrary.org/obo/mondo.owl
NCBITaxon	http://purl.obolibrary.org/obo/ncbitaxon.owl
PRO	http://purl.obolibrary.org/obo/pr.owl
Uberon	http://purl.obolibrary.org/obo/uberon/ext.owl
VO	http://purl.obolibrary.org/obo/vo.owl
Project Notebooks	
[a]OMOP2OBO	https://github.com/callahantiff/OMOP2OBO/blob/master/omop2obo_notebook.ipynb
Mapping Analysis	https://github.com/callahantiff/OMOP2OBO/blob/master/resources/analyses/omop2obo_manuscript_analyses-checkpoint.ipynb
Mapping Evaluation	https://github.com/callahantiff/OMOP2OBO/blob/master/resources/analyses/omop2obo_mapping_validation-checkpoint.ipynb

[a]This Jupyter Notebook serves the same purpose as the main.py script and provides users with a more interactive interface to use when running the algorithm.

Supplementary Table 3: OMOP Data Mapping Data by Clinical Domain.

CONCEPT LEVEL	CONCEPTS	LABELS	SYNONYMS
CONDITIONS			
Standard Concepts Used In Practice			
Concept	29129	29129	86630
Ancestor	1421104	1389525	N/A
Standard Concepts Not Used In Practice			
Concept	80590	80590	194264
Ancestor	3458072	3393343	N/A
DRUG INGREDIENTS			
Standard Concepts Used In Practice			
Concept	1697	1696	1868
Ancestor	1697	1696	N/A
Standard Concepts Not Used In Practice			
Concept	10110	10110	11235
Ancestor	10578	10578	N/A
MEASUREMENTS			
Standard Concepts Used In Practice			
Concept	1606	1606	41891
Ancestor	20781	21191	N/A
Standard Concepts Not Used In Practice			
Concept	2477	2477	73612
Ancestor	23457	24306	N/A

Note. All concepts were from a standard OMOP vocabulary except for one measurement concept which was from a pediatric-specific local source vocabulary.



Supplementary Figure 1: Available Mapping Metadata by Ontology.

Figure provides a visual illustration of the counts, in natural log scale, of labels, external database references, and synonyms available for mappings by OBO Foundry ontology. Acronyms: dbxref (database cross-reference); HPO (Human Phenotype Ontology); Mondo (Mondo Disease Ontology); ChEBI (Chemical Entities of Biological Interest); CL (Cell Ontology); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); VO (Vaccine Ontology); Uberon (Uber-Anatomy Ontology); PRO (Protein Ontology); OBO (Open Biological and Biomedical Ontology).

Supplementary Table 4: OBO Foundry Ontology Mapping Data.

Ontology	Classes	Labels	Synonyms	Database Cross-References
ChEBI	126169	126169	269798	231247
CL	2238	2238	2124	1376
HPO	15247	15247	19860	19569
Mondo	22288	22288	98181	19569
NCBITaxon	2241110	2241110	263571	18246
PRO	215624	215624	590190	195671
Uberon	13898	13898	36771	51322
VO	5789	5789	6	0

Acronyms: ChEBI (Chemical Entities of Biological Interest); CL (Cell Ontology); HPO (Human Phenotype Ontology); Mondo (Mondo Disease Ontology); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); PRO (Protein Ontology); Uberon (Uber-Anatomy Ontology); VO (Vaccine Ontology).

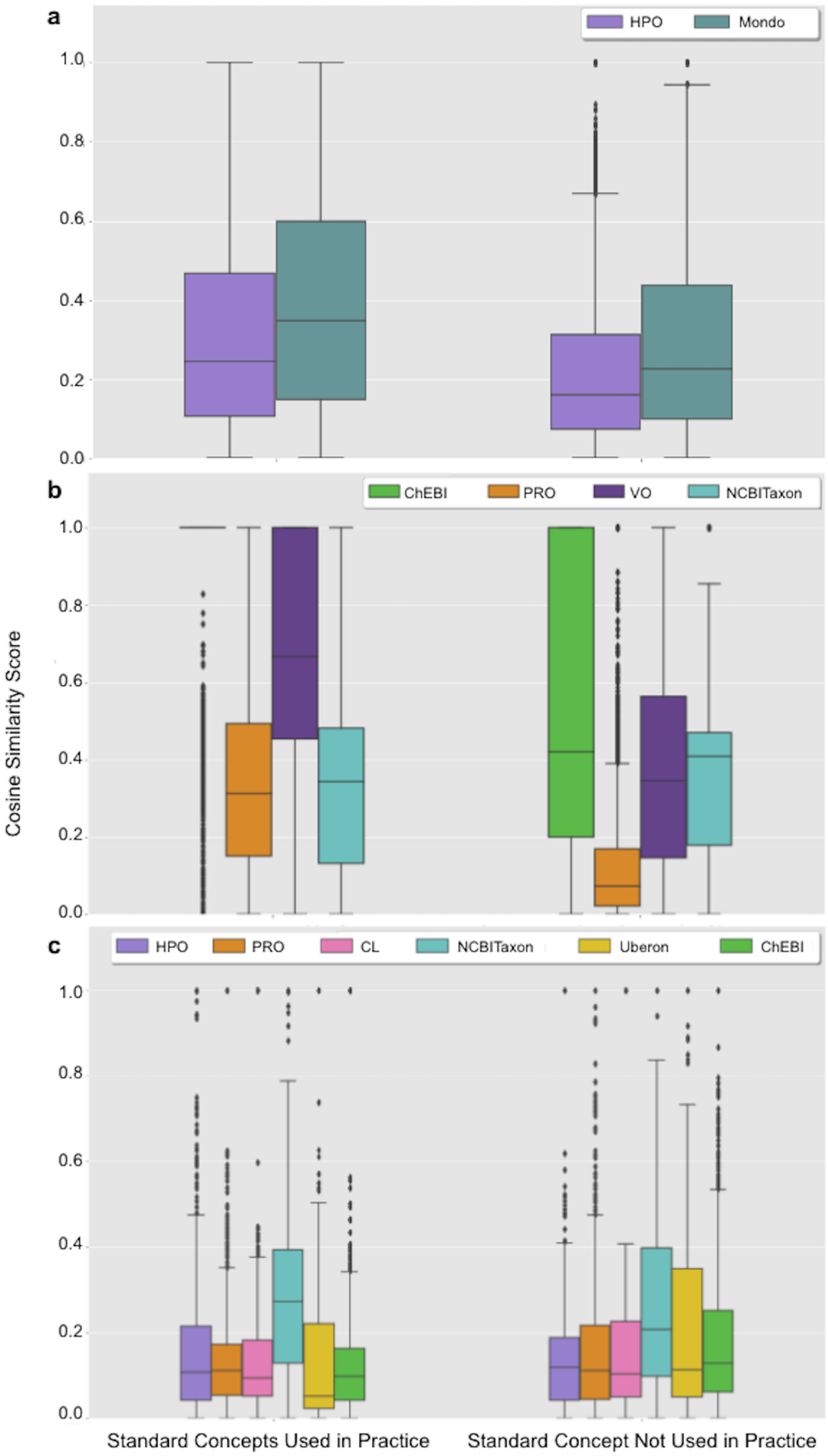
Supplementary Table 5: OMOP2OBO Mapping Categories.

Mapping Category	Definition
Automatic One-to-One Concept	**Definition:** A one-to-one mapping that is automatically generated at the concept-level through exact string mappings to labels or synonyms or exact mappings between codes. **Example:** - `OMOP:22945` (Horizontal overbite) - `HP:0011095` (Overjet) This mapping was created through an exact string mapping on “overjet”, which is the HP concept label and an OMOP concept synonym. This mapping is also supported through exact mappings between database cross-references to SNOMED-CT 70305005 and UMLS C0596028.
Automatic One-to-One Ancestor	**Definition:** A one-to-one mapping that is automatically generated for a concept’s ancestor through exact string mappings to labels or synonyms or exact mappings between codes. **Example:** - `OMOP:22722` (Accessory salivary gland) - `HP:0010286` (abnormal salivary gland morphology) This mapping was created through exact mappings to one of the OMOP concept’s ancestors on the database cross-references to SNOMED-CT 10890000 and UMLS C0036093.
Automatic One-to-Many Concept	**Definition:** A one-to-many mapping that is automatically generated at the concept-level through exact string mappings to labels or synonyms or exact mappings between codes. For release 1.0, one-to-many mappings indicate that one OMOP concept was mapped to one or more OBO ontology concepts. Example: - `OMOP:78854` (Osteopoikilosis) - `MONDO:0001414` (Osteopoikilosis (disease)) AND `MONDO:0008157` (Duschke-Ollendorff Syndrome) This mapping was created through 2 exact string mappings on “osteopoikilosis”, which is a Mondo concept exact synonym and an OMOP concept label and synonym and “duschke-ollendorff syndrome”, which is a Mondo concept exact synonym and label and an OMOP concept synonym. This mapping is also supported through exact mappings between database cross-references to SNOMED-CT 9147009.

Mapping Category	Definition
Automatic One-to-Many Ancestor	**Definition:** A one-to-many mapping that is automatically generated for a concept's ancestor through exact string mappings to labels or synonyms or exact mappings between codes. For release 1.0, one-to-many mappings indicate that one OMOP concept was mapped to one or more OBO ontology concepts. **Example:** - `OMOP:74185` (Open fracture of cuboid bone of foot) - `MONDO:0005315` (bone fracture) AND `MONDO:0044989` (foot disease) This mapping was created through 3 exact string mappings on "fracture", "fracture of bone", and "disorder of foot", which are all Mondo exact synonyms and labels of the OMOP concept's ancestors. This mapping is also supported by exact mappings to one or more of the OMOP concept's ancestors on the database cross-references to SNOMED-CT 125605004 and 118932009.
Manual One-to-One Concept	**Definition:** A one-to-one mapping that is manually generated at the concept-level and usually requires the use of external resources. **Example:** - `OMOP:4070954` (Mesiodens) - `MONDO:0008533` (Teeth, supernumeracy) This mapping was manually created through external evidence from a PubMed article, which stated "Mesiodens is a supernumerary tooth present in the midline between the two central incisors" (`PMID:21998774`).

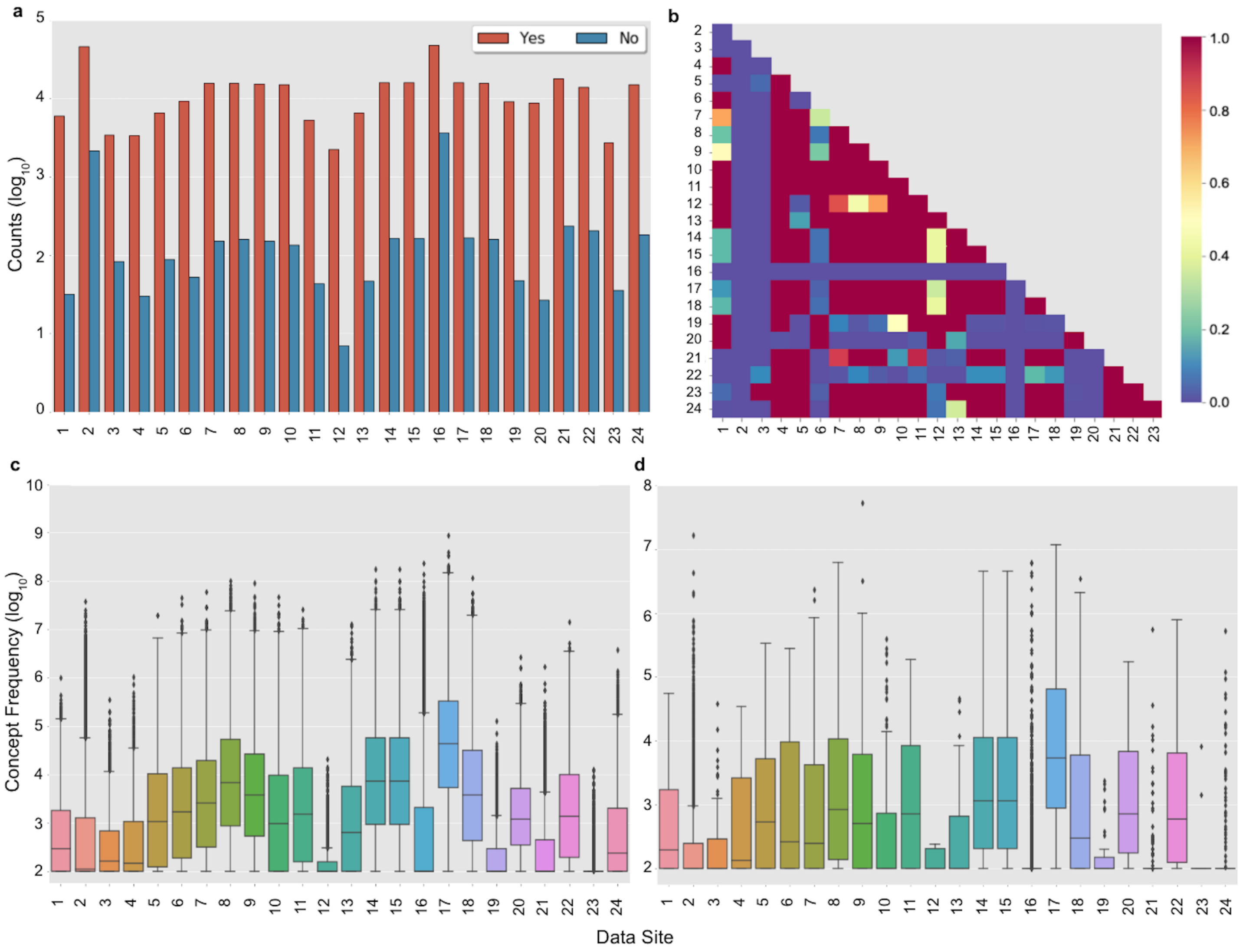
Mapping Category	Definition
Manual One-to-Many Concept	**Definition:** A one-to-many mapping that is manually generated at the concept-level and usually requires the use of external resources. For release 1.0, one-to-many mappings indicate that one OMOP concept was mapped to one or more OBO ontology concepts. **Example:** - `OMOP:439140` (Neonatal polycythemia) - `HP:0003623` (Neonatal onset) AND `HP:0001901` (Polycythemia) This mapping was created through an exact string mappings on “erythrocytosis”, which is a HP concept exact synonym and a OMOP concept ancestor label. This mapping is also supported through exact mappings between database cross-references to SNOMED-CT 127062003 and UMLS C1527405 and C0032461.
Cosine Similarity One-to-One Concept	**Definition:** A one-to-one mapping that is automatically generated at the concept-level using cosine similarity scores. For release 1.0, the cosine similarity scores were applied to concept embeddings learned from a Bag-of-Words model with TF-IDF, which was applied to all available labels and synonyms at the concept- and ancestor-level. **Example:** - `OMOP:4147326` (Sore throat symptom) - `HP:0033050` (Throat pain) This mapping received a cosine similarity score of 0.66.

Mapping Category	Definition
Unmapped	This concept is used when no suitable mapping is possible, for concepts which have not yet been mapped, and for concepts which are purposefully not mapped. **Examples:** *No Suitable Mondo Mapping* - `OMOP:4235440` (Genetic alleles) *Not Yet Mapped to HP or Mondo* - `OMOP:4174055` (Athetoid paralysis) *Purposefully Not Mapped to HP or Mondo* - `OMOP:432499` (Mechanical complication due to coronary bypass graft) → Complication - `OMOP:432498` (Burn of axilla) → Injury - `OMOP:4056963` (Patient on self-medication) → Finding



Supplementary Figure 2: Distribution of Concept Similarity Scores by Clinical Domain and Ontology.

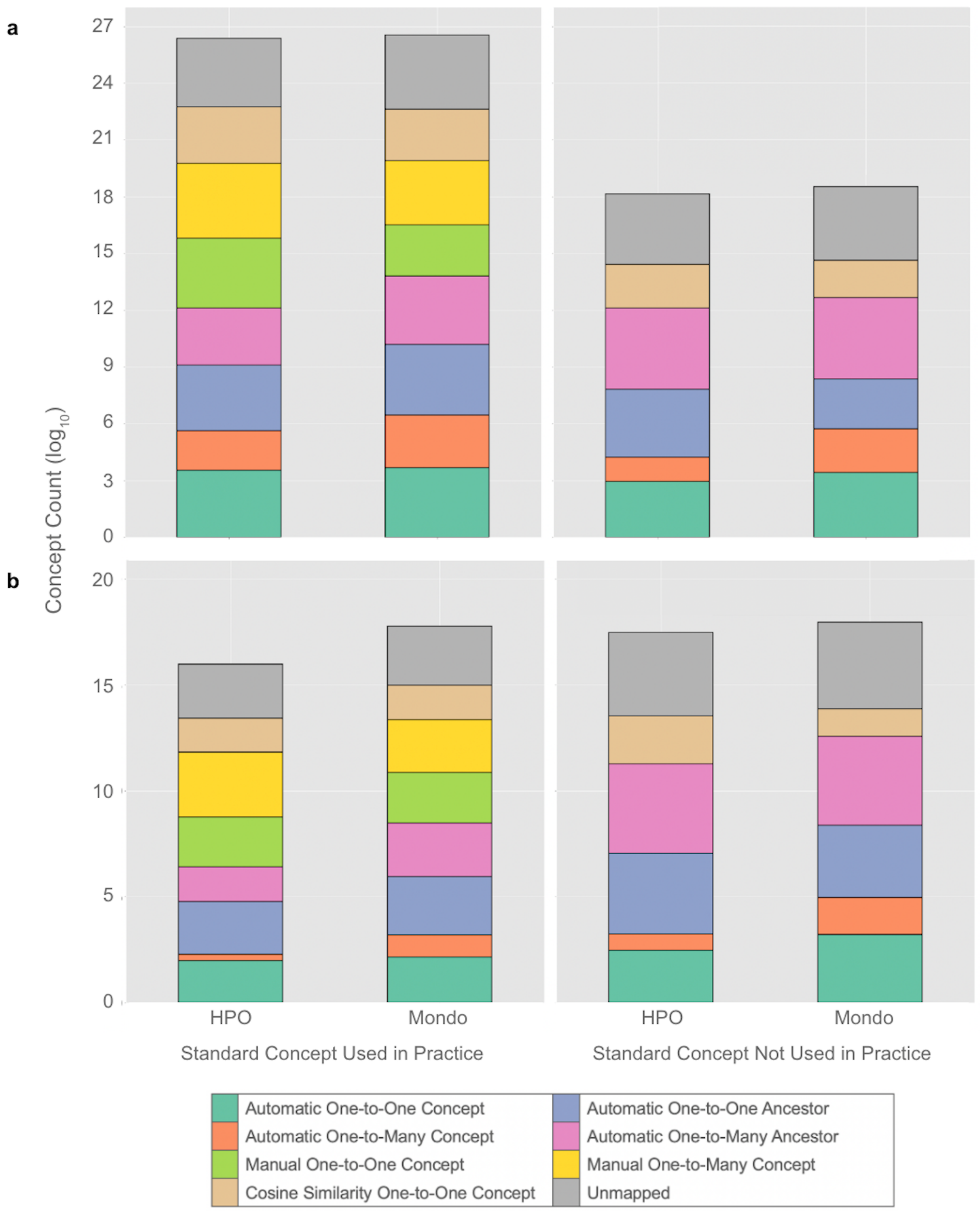
The figure presents the distribution of cosine similarity scores by data wave and OBO Foundry ontology for OMOP (**A**) Conditions, (**B**) Drug Exposure Ingredients, and (**C**) Measurements. Acronyms: OBO (Open Biological and Biomedical Ontology); OMOP (Observational Medical Outcomes Partnership); HPO (Human Phenotype Ontology); Mondo (Monarch Disease Ontology); ChEBI (Chemical Entities of Biological Interest); PRO (Protein Ontology); VO (Vaccine Ontology); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); CL (Cell Ontology); Uberon (Uber-Anatomy Ontology).



Supplementary Figure 3: OMOP2OBO Coverage of Condition Concepts.

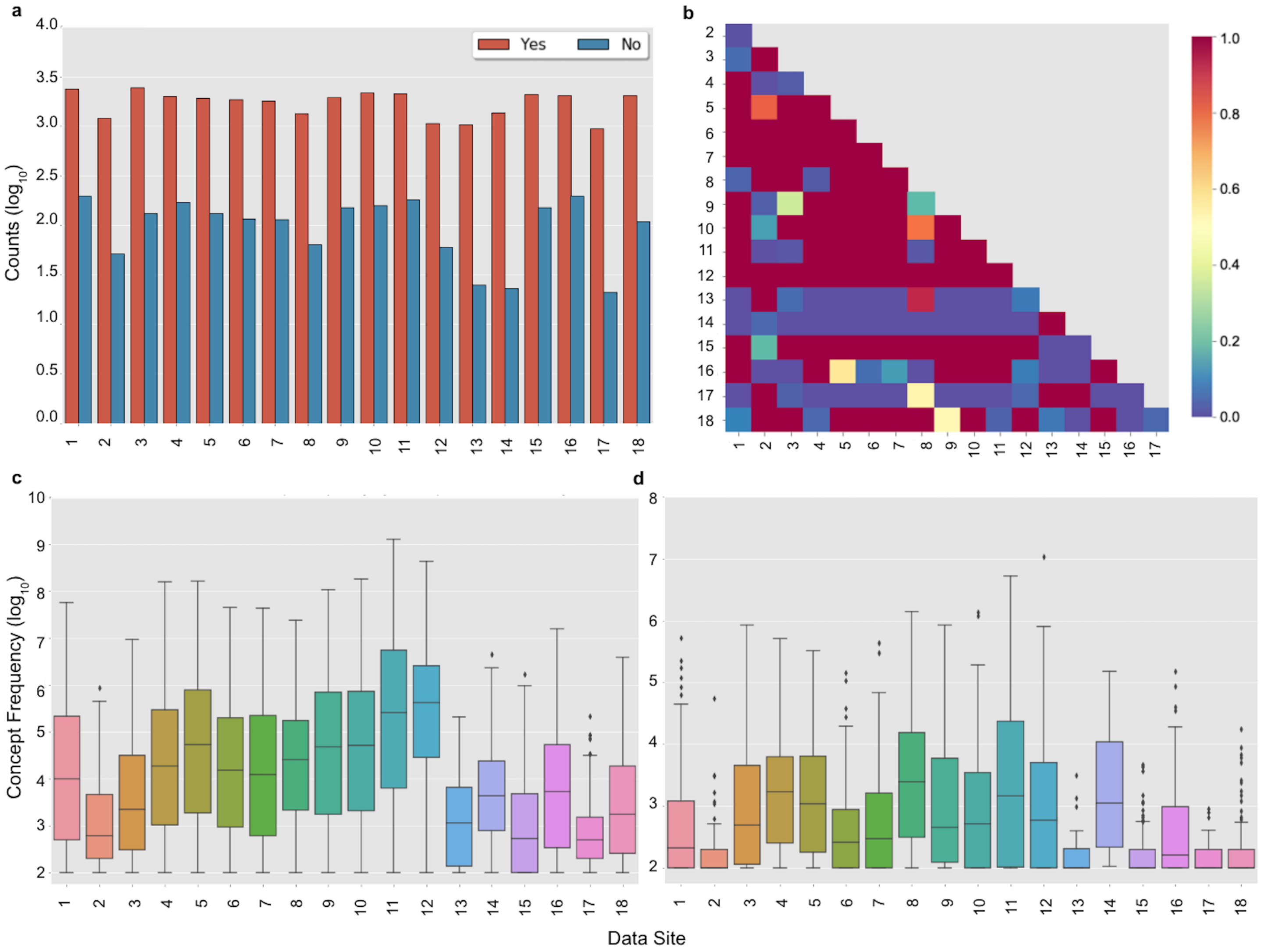
(**A**) This figure visualizes the log count of OMOP2OBO condition concepts covered at each of the Concept Prevalence study sites. (**B**) This figure visualizes the p-values at each Concept Prevalence Study site from the pairwise comparisons of the frequency of concepts from each site that overlapped with the OMOP2OBO mapping set. Post-hoc tests with Bonferroni adjustment to correct for multiple comparisons confirmed that 107 of the 276 pairwise site comparisons had significantly different coverage (ps<0.001). (**C**) This figure visualizes the frequency of the covered OMOP2OBO concepts at each Concept Prevalence site. (**D**) This figure visualizes the frequency of Concept Prevalence site concepts not covered by the OMOP2OBO mappings.

Database Indices: (1) Ajou University Database; (2) IQVIA US Ambulatory Electronic Medical Record; (3) IQVIA Longitudinal Patient Data Australia; (4) IQVIA Disease Analyzer France; (5) IQVIA Disease Analyzer Germany; (6) The Healthcare Cost and Utilization Project Nationwide Inpatient Sample; (7) IQVIA US Hospital Charge Data Master; (8) IBM MarketScan Commercial Database; (9) IBM MarketScan Multi-State Medicaid Database; (10) IBM MarketScan Medicare Supplemental Database; (11) Japan Medical Data Center database; (12) Medical Information Mart for Intensive Care III; (13) Korea National Health Insurance Service/National Sample Cohort; (14) Optum De-Identified Clinformatics Data-Mart-Database—Date of Death; (15) Optum De-Identified Clinformatics Data-Mart-Database— Socio-Economic Status; (16) Optum De-identified Electronic Health Record Dataset; (17) IQVIA US LRxDx Open Claims; (18) Premier Healthcare Database; (19) University of Southern California PScanner; (20) Stanford Medicine Research Data Repository; (21) Tufts Medical Center Database; (22) University of Colorado Anschutz Medical Campus Health Group; (23) Australian Electronic Practice-based Research Network; (24) Columbia University Medical Center Database.



Supplementary Figure 4: OMOP2OBO Condition Concept Count by Ontology and Data Wave.

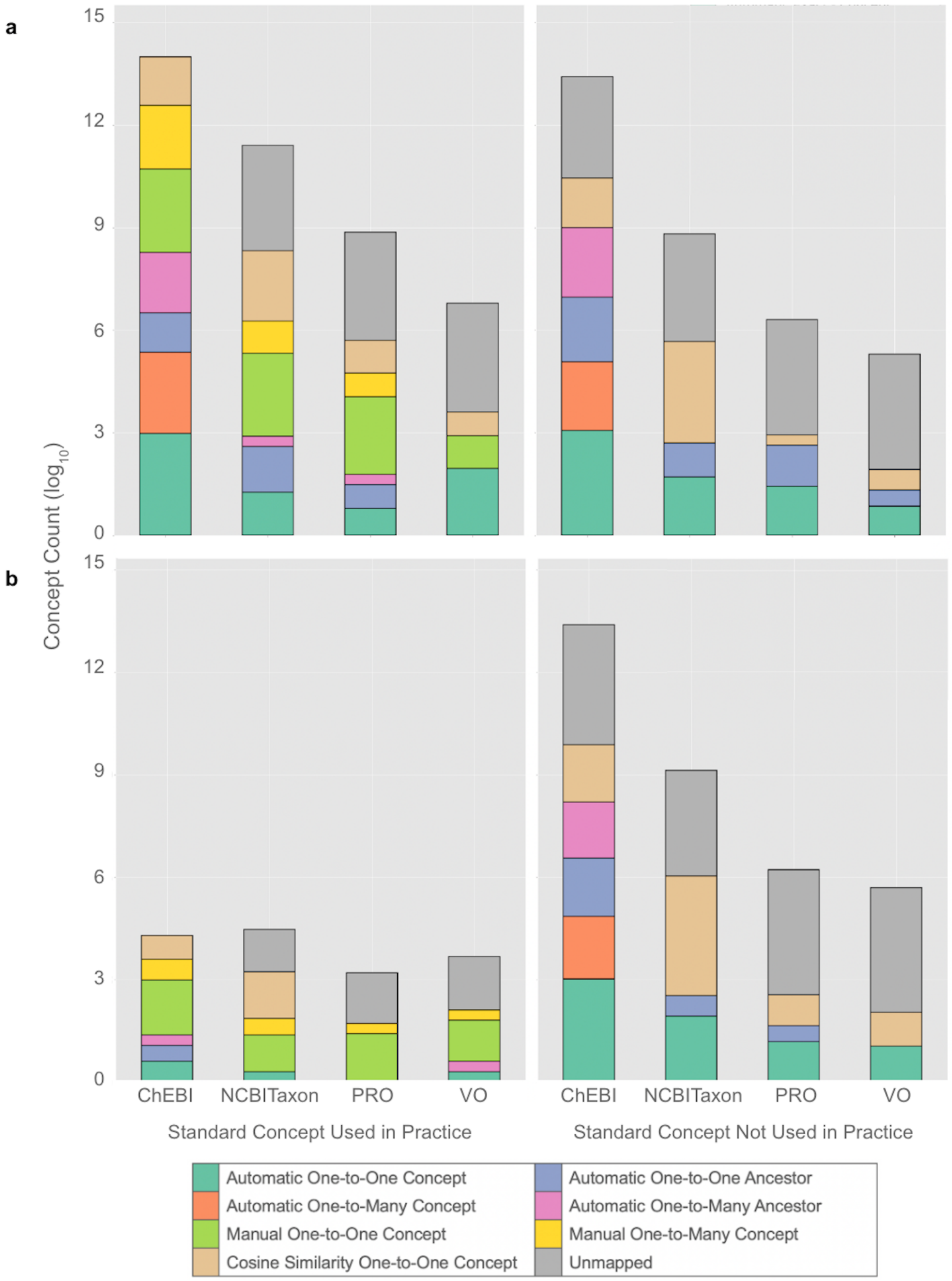
(**A**) This figure visualizes the log count of OMOP2OBO condition concepts that overlapped with concepts in the Concept Prevalence Study by OBO Foundry ontology and data wave. (**B**) This figure visualizes the log count of OMOP2OBO condition concepts that were not found in the Concept Prevalence Study. Acronyms: OBO (Open Biological and Biomedical Ontology); OMOP (Observational Medical Outcomes Partnership); HPO (Human Phenotype Ontology); Mondo (Monarch Disease Ontology).



Supplementary Figure 5: OMOP2OBO Coverage of Drug Ingredient Concepts.

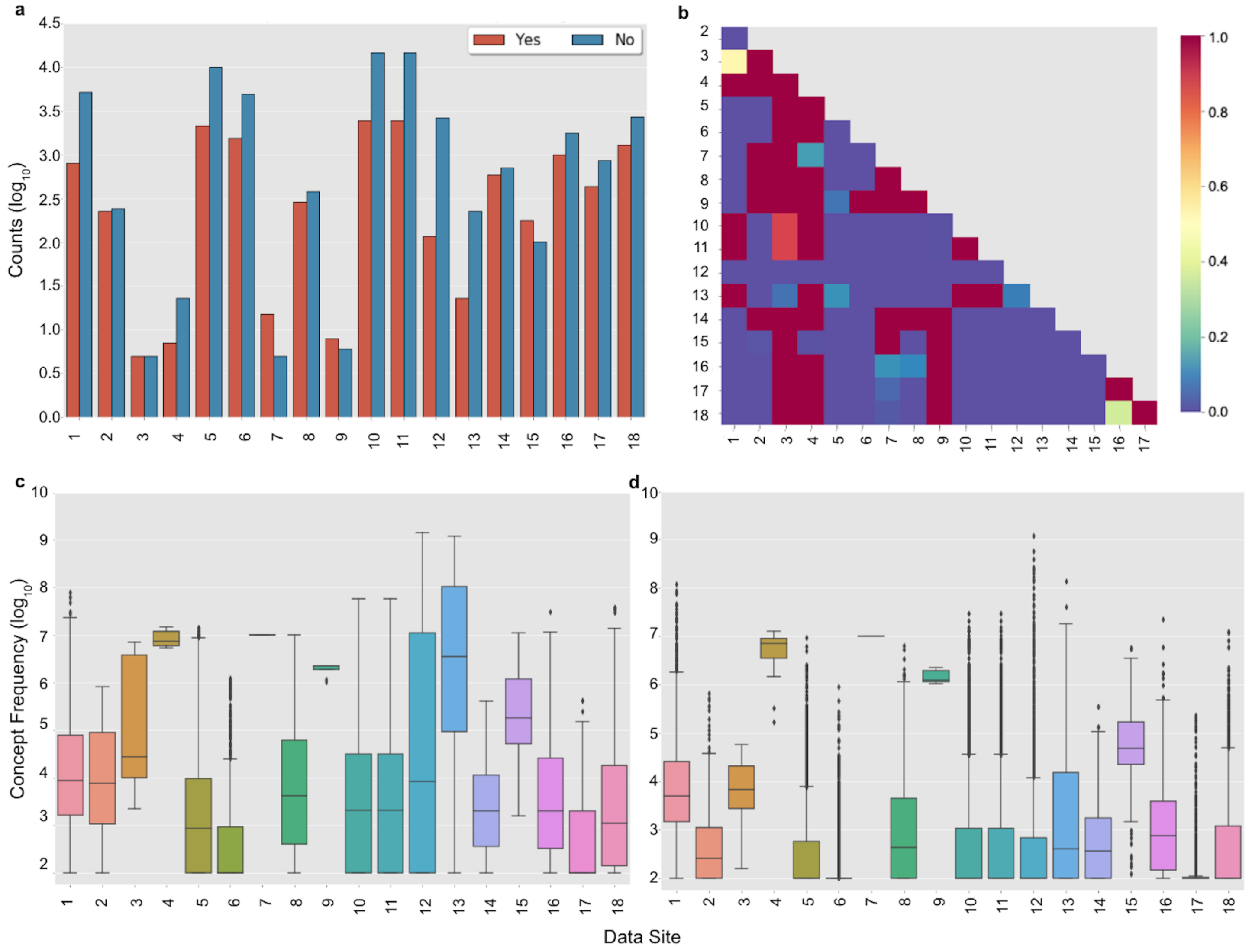
(**A**) This figure visualizes the log count of OMOP2OBO drug ingredient concepts covered at each of the Concept Prevalence study sites. (**B**) This figure visualizes the p-values at each Concept Prevalence Study site from the pairwise comparisons of the frequency of concepts from each site that overlapped with the OMOP2OBO mapping set. Post-hoc tests with Bonferroni adjustment to correct for multiple comparisons confirmed that 53 of the 153 pairwise site comparisons had significantly different coverage (ps<0.001). (**C**) This figure visualizes the frequency of the covered OMOP2OBO concepts at each Concept Prevalence site. (**D**) This figure visualizes the frequency of Concept Prevalence site concepts not covered by the OMOP2OBO mappings.

Database Indices: (1) IQVIA US Ambulatory Electronic Medical Record; (2) IQVIA Longitudinal Patient Data Australia; (3) IQVIA Disease Analyzer Germany; (4) IQVIA US Hospital Charge Data Master; (5) IBM MarketScan Commercial Database; (6) IBM MarketScan Multi-State Medicaid Database; (7) IBM MarketScan Medicare Supplemental Database; (8) Japan Medical Data Center database; (9) Optum De-Identified Clinformatics Data-Mart-Database— Socio-Economic Status; (10) Optum De-identified Electronic Health Record Dataset; (11) Optum De-identified Electronic Health Record Dataset; (12) Premier Healthcare Database; (13) University of Southern California PScanner; (14) Stanford Medicine Research Data Repository; (15) Tufts Medical Center Database; (16) University of Colorado Anschutz Medical Campus Health Group; (17) Australian Electronic Practice-based Research Network; (18) Columbia University Medical Center Database.



Supplementary Figure 6: OMOP2OBO Drug Ingredient Concept Count by Ontology and Data Wave.

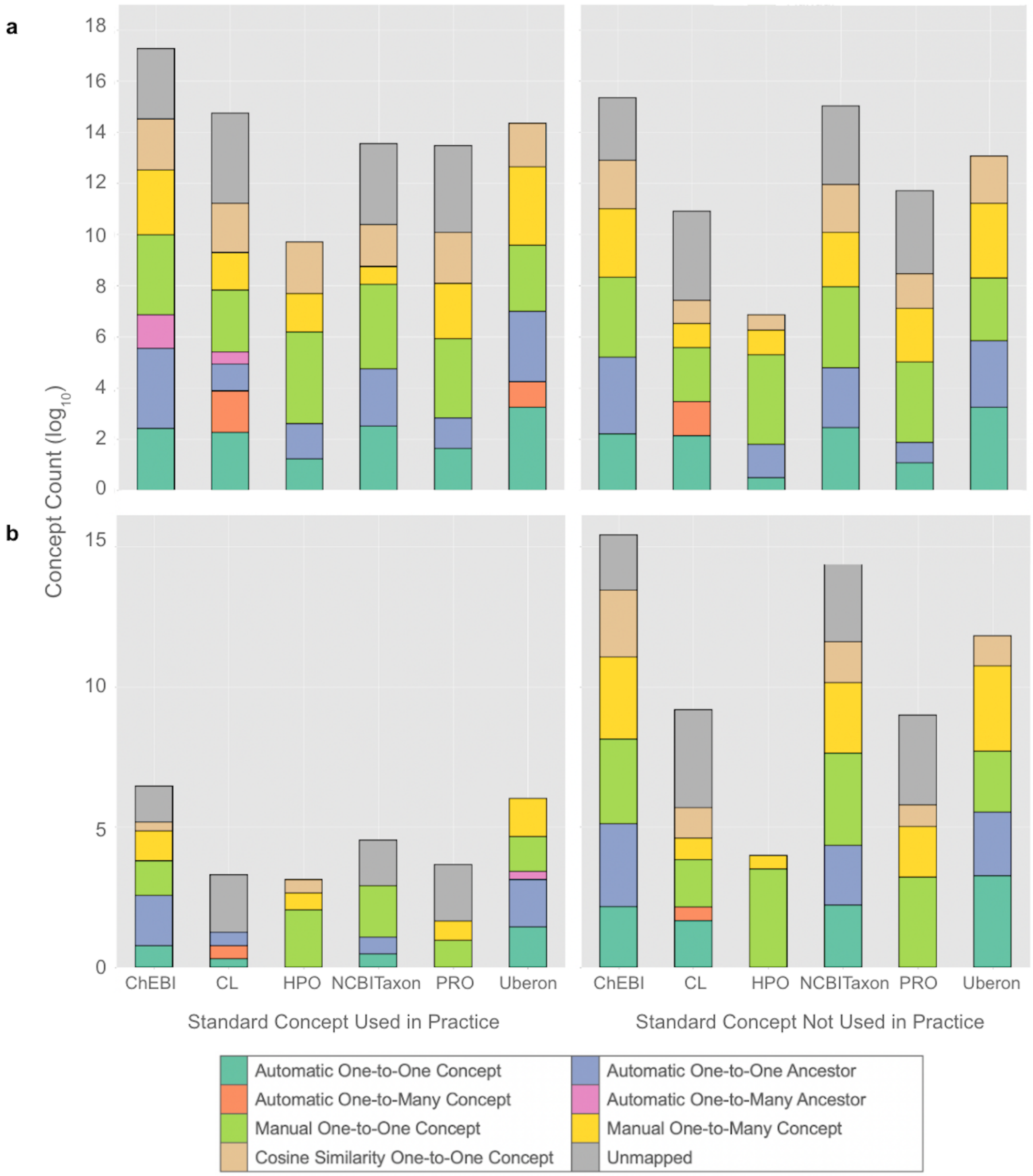
(**A**) This figure visualizes the log count of OMOP2OBO drug ingredient concepts that overlapped with concepts in the Concept Prevalence Study by OBO Foundry ontology and data wave. (**B**) This figure visualizes the log count of OMOP2OBO drug ingredient concepts that were not found in the Concept Prevalence Study. Acronyms: OBO (Open Biological and Biomedical Ontology); OMOP (Observational Medical Outcomes Partnership); ChEBI (Chemical Entities of Biological Interest); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); PRO (Protein Ontology); VO (Vaccine Ontology).



Supplementary Figure 7: OMOP2OBO Coverage of Measurement Concepts.

(**A**) This figure visualizes the log count of OMOP2OBO measurement concepts covered at each of the Concept Prevalence study sites. (**B**) This figure visualizes the p-values at each Concept Prevalence Study site from the pairwise comparisons of the frequency of concepts from each site that overlapped with the OMOP2OBO mapping set. Post-hoc tests with Bonferroni adjustment to correct for multiple comparisons confirmed that 93 of the 153 pairwise site comparisons had significantly different coverage (ps<0.001). (**C**) This figure visualizes the frequency of the covered OMOP2OBO concepts at each Concept Prevalence site. (**D**) This figure visualizes the frequency of Concept Prevalence site concepts not covered by the OMOP2OBO mappings.

Database Indices: (1) IQVIA US Ambulatory Electronic Medical Record; (2) IQVIA Longitudinal Patient Data Australia; (3) IQVIA Disease Analyzer France; (4) IQVIA Disease Analyzer Germany; (5) IBM MarketScan Commercial Database; (6) IBM MarketScan Medicare Supplemental Database; (7) Japan Medical Data Center database; (8) Medical Information Mart for Intensive Care III; (9) Korea National Health Insurance Service/National Sample Cohort; (10) Optum De-Identified Clinformatics Data-Mart-Database—Date of Death; (11) Optum De-Identified Clinformatics Data-Mart-Database—Socio-Economic Status; (12) Optum De-identified Electronic Health Record Dataset; (13) Premier Healthcare Database; (14) University of Southern California PScanner; (15) Stanford Medicine Research Data Repository; (16) University of Colorado Anschutz Medical Campus Health Group; (17) Australian Electronic Practice-based Research Network; (18) Columbia University Medical Center Database.



Supplementary Figure 8: OMOP2OBO Measurement Concept Count by Ontology and Data Wave.

(**A**) This figure visualizes the log count of OMOP2OBO measurement concepts that overlapped with concepts in the Concept Prevalence Study by OBO Foundry ontology and data wave. (**B**) This figure visualizes the log count of OMOP2OBO measurement concepts that were not found in the Concept Prevalence Study. Acronyms: OBO (Open Biological and Biomedical Ontology); OMOP (Observational Medical Outcomes Partnership); ChEBI (Chemical Entities of Biological Interest); CL (Cell Ontology); HPO (Human Phenotype Ontology); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); PRO (Protein Ontology); Uberon (Uber-Anatomy Ontology).