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

Tiffany J. Callahan[1,2*], Adrianne L. Stefanski[1], Jordan M. Wyrwa[3], Chenjie Zeng[4], Anna Ostropolets[2], Juan M. Banda[5], William A. Baumgartner Jr.[1], Richard D. Boyce[6], Elena Casiraghi[7], Ben D. Coleman[8], Janine H. Collins[9], Sara J. Deakyne-Davies[10], James A. Feinstein[11], Melissa A. Haendel[12], Asiyah Y. Lin[4], Blake Martin[13], Nicolas A. Matentzoglu[14], Daniella Meeker[15], Justin Reese[16], Jessica Sinclair[17], Sanya B. Taneja[18], Katy E. Trinkley[19], Nicole A. Vasilevsky[20], Andrew Williams[21], Xingman A. Zhang[22], Peter N. Robinson[8], Patrick Ryan[23], George Hripcsak[2], Tellen D. Bennett[13], Lawrence E. Hunter[1,24], Michael G. Kahn[24]

[1]Computational Bioscience Program, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[2]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA
[3]Department of Physical Medicine and Rehabilitation, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[4]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
[5]Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA
[6]Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA, 15260, USA
[7]Computer Science, Università degli Studi di Milano, Milan, 20122, Italy
[8]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA
[9]Department of Haematology, University of Cambridge, Cambridge, UK
[10]Department of Research Informatics & Data Science, Analytics Resource Center, Children's Hospital Colorado, Aurora, CO, 80045, USA
[11]Adult and Child Center for Health Outcomes Research and Delivery Science (ACCORDS), University of Colorado Anschutz School of Medicine, Aurora, CO, 80045, USA
[12]Center for Health AI, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[13]Departments of Biomedical Informatics and Pediatrics, University of Colorado School of Medicine, Aurora, CO 80045, USA
[14]Semanticly, Athens, Greece
[15]Department of Preventive Medicine, Leonard D. Schaeffer Center for Health Policy and Economics, University of Southern California, Los Angeles, CA, 90033 USA
[16]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[17]HealthLinc, Valparaiso, IN 46383, USA
[18]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, 15260, USA
[19]Department of Clinical Pharmacy and Medicine, University of Colorado Anschutz Skaggs School of Pharmacy and Pharmaceutical Sciences and School of Medicine, Aurora, CO 80045, USA
[20]Translational and Integrative Sciences Lab, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[21]Tufts Institute for Clinical Research and Health Policy Studies, Tufts University, Boston, MA 02155, USA
[22]Sema4, Stamford, CT 06902 USA
[23]Janssen Research and Development, Raritan, NJ 08869, USA
[24]Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO 80045, USA

[*]Corresponding author

ABSTRACT

Common data models solve many challenges of standardizing electronic health record (EHR) data, but are unable to semantically integrate the resources needed for deep phenotyping. Open Biological and Biomedical Ontology (OBO) Foundry ontologies provide semantically computable representations of biological knowledge and enable the integration of a variety of biomedical data. However, mapping EHR data to OBO Foundry ontologies requires significant manual curation and domain expertise. We introduce a framework for mapping Observational Medical Outcomes Partnership (OMOP) standard vocabularies to OBO Foundry ontologies. Using this framework, we produced mappings for 92,367 conditions, 8,615 drug ingredients, and 10,673 measurement results. Mapping accuracy was verified by domain experts and when examined across 24 hospitals, the mappings covered 99% of conditions and drug ingredients and 68% of measurements. Finally, we demonstrate that OMOP2OBO mappings can aid in the systematic identification of undiagnosed rare disease patients who might benefit from genetic testing.

INTRODUCTION

Electronic health record (EHR) adoption, which is nearly universal within the US healthcare system,[1,2] has increased adherence to evidence-based clinical guidelines[3] and facilitated greater patient communication[4] resulting in significant improvements in care.[5] EHRs contain a myriad of systematically collected, longitudinal, patient-level information and are a valuable resource for population-level research.[6] One promise of EHR-based phenotyping is the ability to perform population-level investigations of mechanistic drivers of disease in diverse patient populations.[7,8] Despite significant progress, this objective remains largely aspirational.[6,9–12]

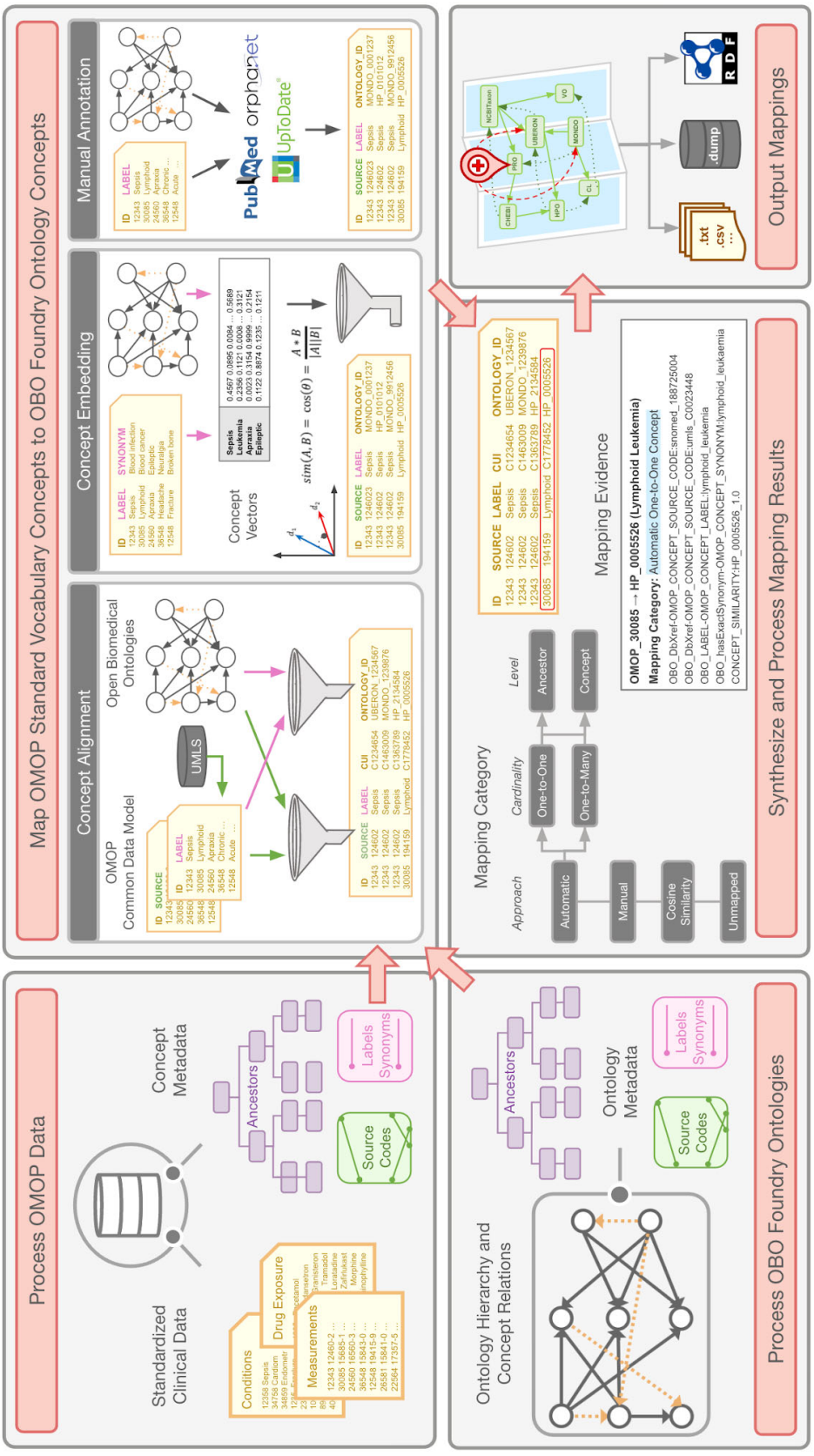
Deep phenotyping, or "the precise and comprehensive analysis of phenotypic abnormalities in which the individual components of the phenotype are observed and described"[13], is a fundamental component of precision medicine that requires the timely synthesis of multiple types of patient data.[14,15] Deep phenotyping has successfully been applied to rare and genetic disorders,[16–28] cancer,[29–35] pregnancy,[36,37] and has been used to identify patients with undiagnosed rare diseases[38]. While common data models (CDMs) like the Observational Medical Outcomes Partnership (OMOP)[39] have solved many of the challenges of standardizing and utilizing clinical EHR data, most do not yet include the resources needed to integrate and interpret molecular data.[40] Ontologies exist for nearly all scales of biological organization and when combined, can provide a semantically rich and biologically accurate representation of molecular entities and mechanisms.[41] Similar to clinical vocabularies, ontologies are classification systems that provide detailed representations of a specific domain of knowledge.[42] Unlike clinical vocabularies, ontologies are semantically computable and interoperable with formally defined relationships, which means they can be logically verified and integrated with data from basic science and clinical research.[42]

Mapping clinical vocabularies to ontologies, like those in the Open Biological and Biomedical Ontology (OBO) Foundry, has been recognized as a fundamental requirement for

use in deep phenotyping.[15,38,42,43] An example of how aligning these resources improves deep phenotyping was demonstrated by Zhang et al., (2019)[44] who mapped Logical Observation Identifiers, Names and Codes (LOINC)[45] to the Human Phenotype Ontology (HPO),[46] which enabled the harmonization of laboratory tests with different codes to common HPO concepts. Aligning clinical vocabularies to OBO Foundry ontologies also has the potential to unlock translational resources, which are otherwise not easily accessible or available in the EHR.[42,43] For example, OpenTargets is a service that aligns genomic data to different disease identifiers and ontologies like the Experimental Factor Ontology[47] and the Mondo Disease Ontology (Mondo),[48,49] in order to improve the systematic identification and prioritization of drug targets for specific diseases and phenotypes.[50] Biomedical ontologies like HPO have also been used in theoretical applications as a way to transform entire EHRs,[51–53] but the majority of existing work has focused on phenotyping particular diseases[54–57] or investigating specific biological[58] or clinical domains.[44,59,60] Due to the time-consuming manual effort required to map clinical vocabularies to OBO Foundry ontologies, no comprehensive mapping across commonly used ontologies currently exist. While automated approaches to create mappings have been developed, they are not yet able to accurately capture the complex clinical semantics underlying the data and knowledge encoded by clinical vocabulary concepts.

To address these limitations and enable large-scale semantically interoperable deep phenotyping, we developed OMOP2OBO, a framework to align clinical vocabularies to OBO Foundry ontologies (Figure 1). Using this method, we created the first healthcare system-scale mappings between clinical vocabularies in the OMOP CDM and eight OBO Foundry ontologies[61] spanning diseases (Mondo[48]), phenotypes (HPO[46]), anatomical entities (Uber Anatomy Ontology [Uberon[62]]; Cell Ontology [CL][63]), organisms (National Center for Biotechnology Information Taxon Ontology [NCBITaxon][64]), chemicals (Chemical Entities of Biological Interest [ChEBI][65]), vaccines (the Vaccine Ontology [VO][66]), and proteins (the Protein Ontology [PRO][67]). We evaluated the mappings in three ways: (1) accuracy, examined by a team of domain

Figure 1: Overview of the OMOP2OBO Framework.



experts; (2) generalizability, examined through comparison to a large set of mapped concepts used at least once in clinical practice from 24 hospital systems; and (3) clinical utility, examined when used to identify patients with an undiagnosed rare disease.

RESULTS

OMOP2OBO is open source (https://github.com/callahantiff/OMOP2OBO) and includes an dashboard (http://tiffanycallahan.com/OMOP2OBO_Dashboard/). Acronyms used in this paper are provided in Supplementary Table 1 and OMOP2OBO resources are described in Supplementary Table 2.

Mapping Data

OMOP Data

OMOP concepts were extracted from a de-identified pediatric dataset normalized to the OMOP CDM.[39,68] Supplementary Table 3 contains the counts of data available for mapping by clinical domain (i.e., conditions, drug ingredients, and measurements) and whether the concepts were used in at least one clinical encounter in a de-identified pediatric dataset (i.e., *Standard Concepts Used in Practice)* or not (i.e., *Standard Concepts Not Used in Practice*). There were 109,709 condition concepts (Systematized Nomenclature of Medicine -- Clinical Terms [SNOMED-CT][69]) and 11,807 drug ingredient concepts (i.e., RxNorm[70]) available to map. For measurements, there were 4,083 concepts (i.e., LOINC[45]), representing 11,269 measurement results available to map. This concept set included 2,477 LOINC2HPO concepts (6,844 measurement results) after excluding 631 overlapping concepts and 11 deprecated concepts. With respect to the *Standard Concepts Used in Practice*, the 29,129 conditions had a median frequency of 25 (max=544,618), the 1,697 drug ingredients had a median frequency of 251 (max=2,267,866), and the 1,606 measurement concepts had a median frequency of 25 (max=56,823,139).

OBO Foundry Ontologies

As shown in Supplementary Figure 1 and Supplementary Table 4, the amount of metadata available for mapping varied across the OBO Foundry ontologies, with NCBITaxon containing the most metadata and Uberon containing the least. A Chi-square test of independence with Yate's correction revealed a significant association between the ontology and the amount of available metadata ($\chi^2(14) = 2{,}664{,}853.82$, $p<0.0001$). Post-hoc tests with Bonferroni adjustment confirmed the ontologies provided significantly different amounts of metadata ($ps<0.0001$).

OMOP2OBO Mappings

Figure 2 includes example mappings and illustrates how the OBO Foundry ontologies were used to map OMOP concepts from each clinical domain. Supplementary Table 5 provides additional details on and examples of the mapping categories. The mapping procedures and resources are described in the **OMOP2OBO Framework** section of the Methods.

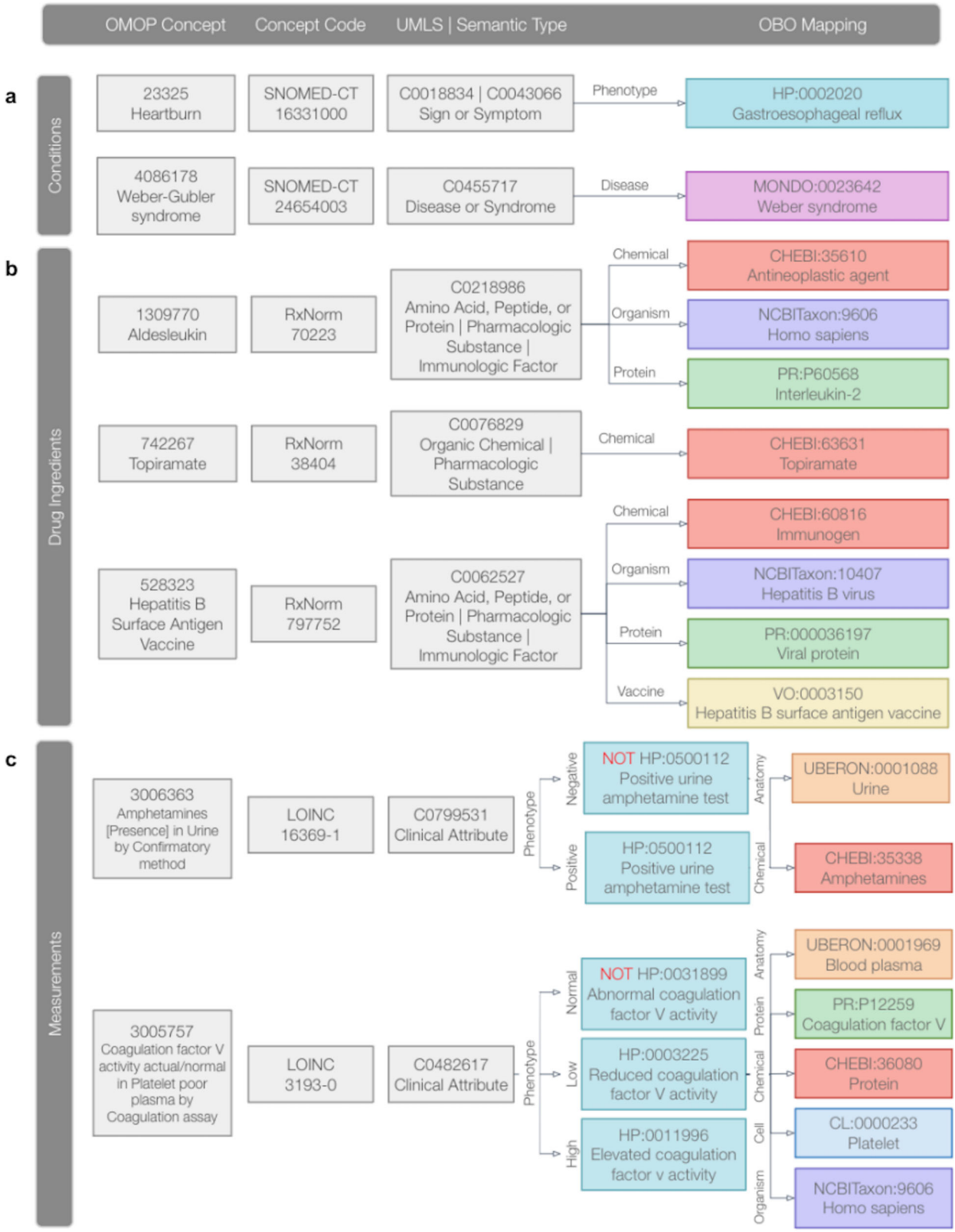
Conditions

Unified Medical Language System (UMLS)[71] concept unique identifiers (CUIs) were found for 96.6% of condition concepts (n=105,976) representing 69 unique Semantic Types.[72] The mapping results for each OBO Foundry ontology are shown in Table 1. Of the 109,709 available concepts, 73,418 mapped to 5,661 unique HPO concepts (83.9% *Standard Concepts Used in Practice*, 60.8% *Standard Concepts Not Used in Practice*) and 63,375 mapped to 9,643 unique Mondo concepts (68.9% *Standard Concepts Used in Practice*, 53.8% *Standard Concepts Not Used in Practice*). Only 50 concepts we attempted to map (excluding purposefully unmapped concepts) were unable to be mapped to at least one OBO Foundry ontology concept.

Mapping Categories

The frequency distributions of the *Standard Concepts Used in Practice* by mapping category and ontology are visualized in Figure 3. The majority of automatic mappings were one-to-one at

Figure 2: OMOP2OBO mapping examples by OMOP clinical domain.



the concept-level for *Standard Concepts Used in Practice* (HPO: n=3,601; Mondo: n=4,836) and *Standard Concepts Not Used in Practice* (HPO: n=1,166; Mondo: n=4,261). For the manual approaches, only the *Standard Concepts Used in Practice* were mapped, with the majority being one-to-many (HPO: n=10,425; Mondo: n=2,836). Cosine similarity-scored concept embeddings enabled 5,020 HPO and 755 Mondo mappings (Supplementary Figure 2a). On average, more evidence to support the mappings was found for *Standard Concepts Used in Practice* than *Standard Concepts Not Used in Practice* for Mondo mappings (5.19 vs 2.29) than HPO mappings (3.84 vs 4.28).

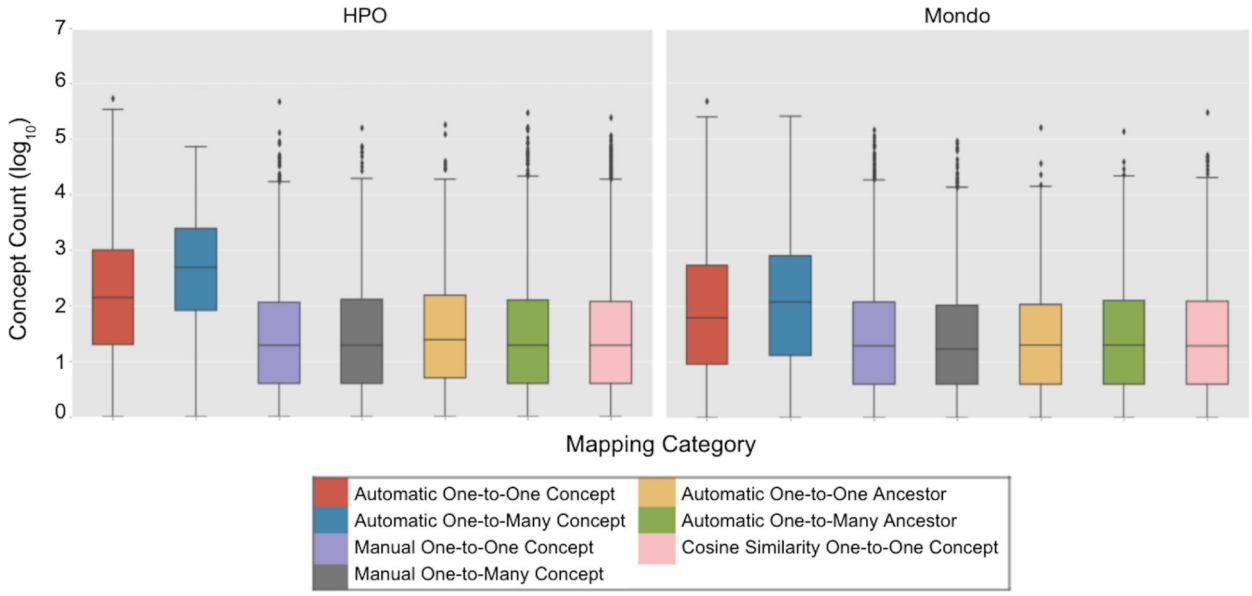
Drug Ingredients

UMLS CUIs were found for 99.2% of drug ingredient concepts (n=11,716) representing 23 unique Semantic Types. The mapping results for each OBO Foundry ontology are shown in Table 2. Of the 11,807 available concepts, 4,719 concepts mapped to 2,739 unique NCBITaxon concepts (27.3% *Standard Concepts Used in Practice*, 42.1% *Standard Concepts Not Used in Practice*), 4,415 mapped to 4,074 unique ChEBI concepts (100% *Standard Concepts Used in Practice*, 26.9% *Standard Concepts Not Used in Practice*), 317 concepts mapped to 145 unique PRO concepts (14.4% *Standard Concepts Used in Practice*, 0.7% *Standard Concepts Not Used in Practice*), and 161 concepts mapped to 134 unique VO concepts (7.4% *Standard Concepts Used in Practice*, 0.4% *Standard Concepts Not Used in Practice*). All of the OMOP concepts were able to be mapped to at least one ChEBI concept.

Mapping Categories

The frequency distributions of the *Standard Concepts Used in Practice* by mapping category and OBO Foundry ontology are visualized in Figure 4. The majority of automated mappings were one-to-one at the concept-level for *Standard Concepts Used in Practice* (ChEBI: n=959; NCBITaxon: n=20; PRO: n=7; VO: n=92) and *Standard Concepts Not Used in Practice* (ChEBI: n=2,192; NCBITaxon: n=135; PR: n=42; VO: n=18). For the manual approaches, only the

Figure 3: Condition Concept Frequency by Mapping Category and Ontology.

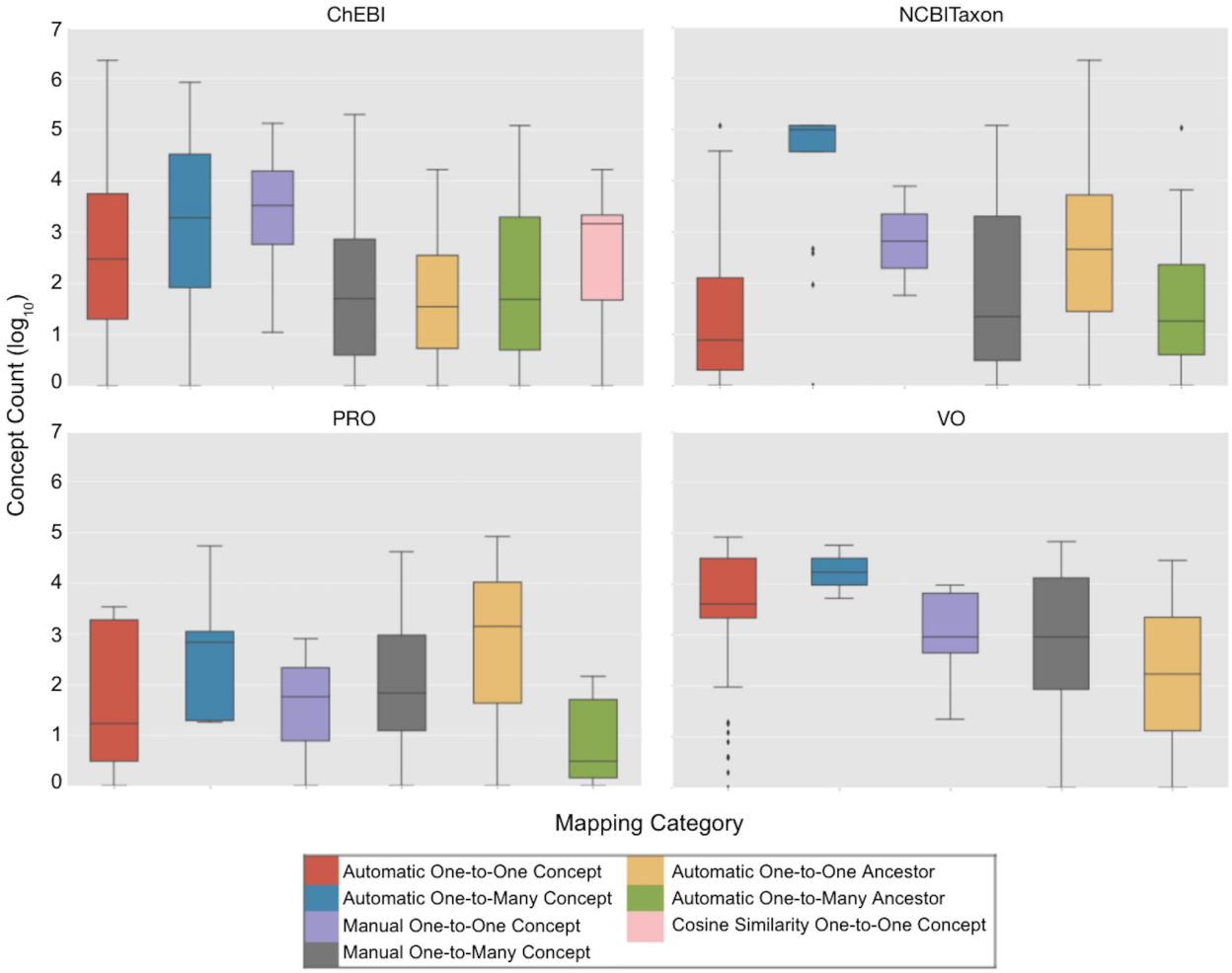


Standard Concepts Used in Practice were mapped with the majority being one-to-one (ChEBI: n=31; NCBITaxon: n=136; PRO: n=9; VO: n=5). Cosine similarity-scored concept embeddings enabled 396 ChEBI, 4,376 NCBITaxon, 224 PRO, 38 VO, and 4,376 NCBITaxon mappings (Supplementary Figure 2b). More evidence to support the mappings was found for Standard *Concepts Used in Practice* than *Standard Concepts Not Used in Practice* in ChEBI, excluding mappings to PRO and NCBITaxon (ChEBI: 46.41 vs 9.01; VO: 6.15 vs 5.00; PR: 1.32 vs. 5.59; NCBITaxon: 1.70 vs 2.19).

Measurements

UMLS CUIs were found for 95.8% of measurement concepts (n=3,868) representing a single Semantic Type. The mapping results for each OBO Foundry ontology are shown in Table 3. Of the 11,807 measurement results, 10,888 results concepts mapped to 1,118 unique HPO concepts (92.4% *Standard Concepts Used in Practice*, 92.4% *Standard Concepts Not Used in Practice*), 10,888 results concepts mapped to 48 unique Uberon concepts (99.3% *Standard Concepts Used in Practice*, 99.4% *Standard Concepts Not Used in Practice*), 9,902 results concepts mapped to 446 unique ChEBI concepts (78.8% *Standard Concepts Used in*

Figure 4: Drug Ingredient Concept Frequency by Mapping Category and Ontology.



Practice, 93.7% *Standard Concepts Not Used in Practice*), 7,460 results concepts mapped to 428 unique NCBITaxon concepts (58.1% *Standard Concepts Used in Practice*, 71.4% *Standard Concepts Not Used in Practice*), 4,855 results concepts mapped to 176 unique PRO concepts (35.5% *Standard Concepts Used in Practice*, 47.9% *Standard Concepts Not Used in Practice*), and 1,045 results concepts mapped to 41 unique CL concepts (13.9% *Standard Concepts Used in Practice*, 6.3% *Standard Concepts Not Used in Practice*). Only five concepts we attempted to map (excluding purposefully unmapped concepts) were unable to be mapped to at least one OBO Foundry ontology concept.

Mapping Categories

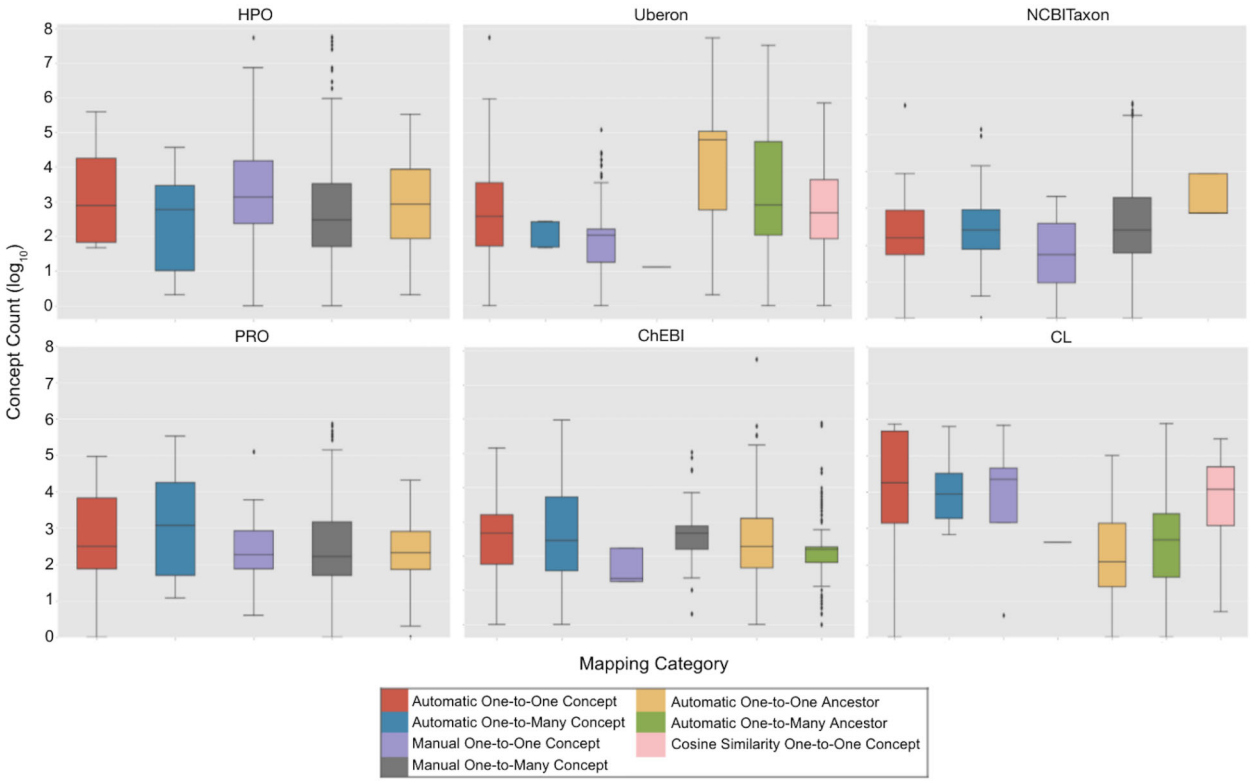
The frequency distributions of the *Standard Concepts Used in Practice* by mapping category and OBO Foundry ontology are visualized in Figure 5. The majority of the automated mappings were one-to-one at the concept-level for *Standard Concepts Used in Practice* (ChEBI: n=263; CL: n=182; HPO: n=17; NCBITaxon: n=320; PRO: n=44; Uberon: n=1793) and *Standard Concepts Not Used in Practice* (ChEBI: n=400; CL: n=186; HPO: n=3; NCBITaxon: n=444; PRO: n=12; Uberon: n=3589). For manual approaches, the majority were one-to-one (ChEBI: n=1,369 and 2,376; CL: n=256 and 178; HPO: 3,902 and 6,671; NCBITaxon: n=2,019 and 3,516; PRO: n=1,261 and 3,047; Uberon: n=406 and 462) for *Standard Concepts Used in Practice* and *Standard Concepts Not Used in Practice*, respectively. Cosine similarity-scored concept embeddings enabled 464 ChEBI, 105 CL, 113 HPO, 158 NCBITaxon, 4,132 PRO, and 142 Uberon mappings (Supplementary Figure 2c). On average, more evidence to support the mappings was found for *Concepts Used in Practice* than *Standard Concepts Not Used in Practice* for all of the OBO Foundry ontologies (HPO: 1.11 vs 1.04; Uberon: 3.03 vs 2.79; NCBITaxon: 1.67 vs 1.47; PRO: 1.11 vs 0.98; ChEBI: 3.51 vs 3.28) except CL (3.35 vs 3.76).

Validation

Accuracy

The goal of this task was to verify the accuracy of the OMOP2OBO mappings through domain expert review. Of the 2,000 condition mappings, 73.8% were correct (n=1,477). Of the 523 (26.2%) incorrect mappings, 165 (31.5%) could be improved by creating more specific mappings or replacing multiple concepts with a general ancestor concept. Of the 116 drug ingredient mappings, 70.7% (n=82) were correct. Of the 34 (29.3%) incorrect mappings, 14 (41.2%) could be improved by creating more specific mappings or replacing multiple concepts with a general ancestor concept. Measurement concepts were reviewed at the result-level using a survey and manual domain expert review. On the survey, 92.9% (n=251) of the mappings

Figure 5: Measurement Concept Frequency by Mapping Category and Ontology.



were found to be correct. Of the 1,350 measurement results, 97.3% (n=1,314) were correct. The error rates for each clinical domain are expected to be much lower in the final mapping set as all identified errors were corrected and improvements were made to the algorithm.

Generalization

The goal of this evaluation was to characterize the coverage of standard concepts in the OMOP2OBO mapping set to standard concepts utilized at least once in practice in the Observational Health Data Sciences and Informatics (OHDSI) Concept Prevalence Study.[73]

Conditions

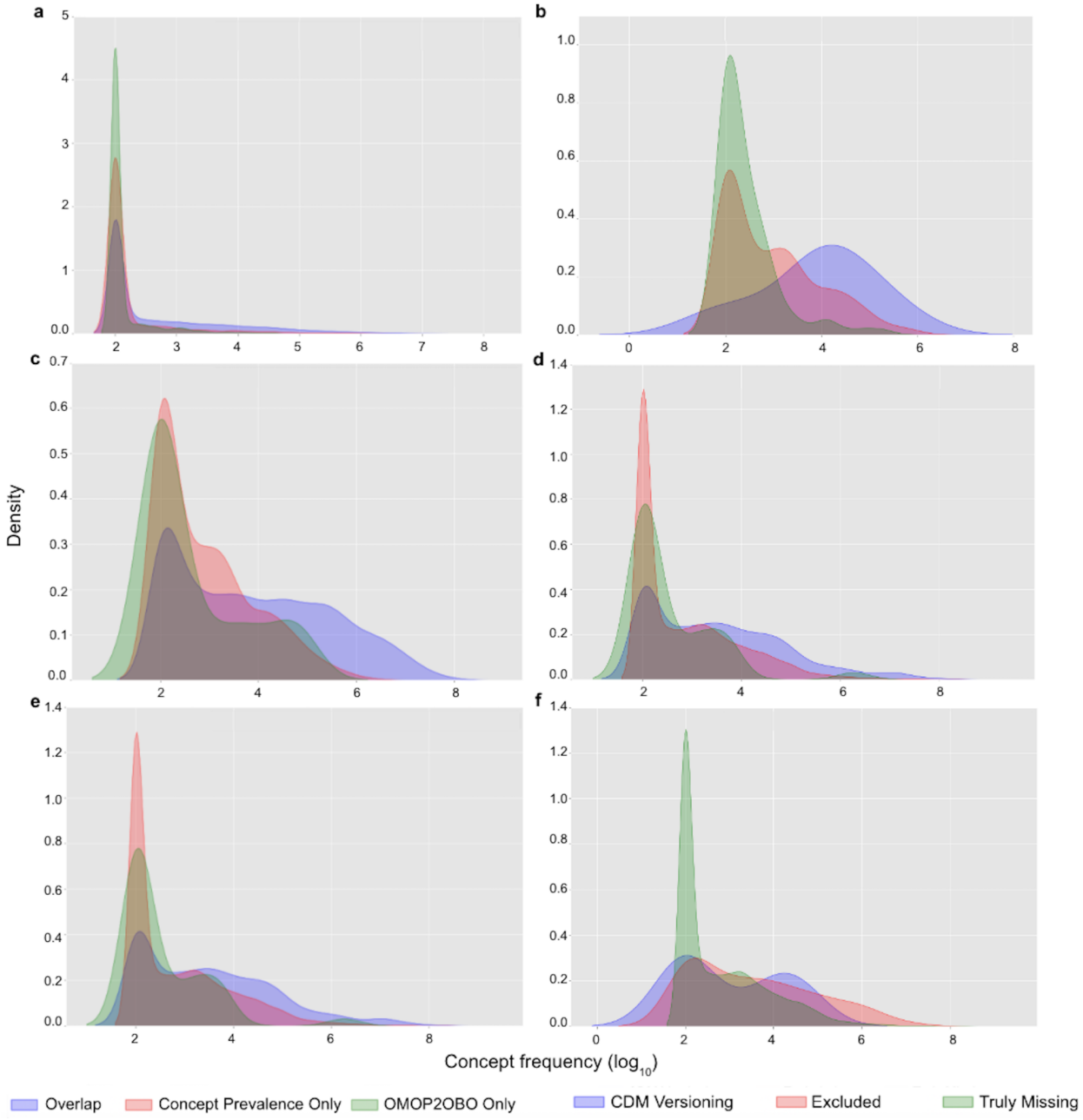
The Concept Prevalence Study contained data for 62,335 distinct concepts from 24 sites. The resulting OMOP2OBO mapping set contained 92,367 eligible concepts, which covered 92.5% (99.5% weighted coverage) of the Concept Prevalence Study concepts (n=57,663 concepts; average frequency: 526.9 [max=87,285,164.4]). Of the remaining concepts, 34,704 were only

found in OMOP2OBO (average frequency: 131.7 [max=39,975]) and 4,672 were only found in the Concept Prevalence Study (average frequency: 173.6 [max=8,254,186.5]). These findings are visualized in Figure 6a. OMOP2OBO concept coverage ranged from 93-99.7% across the 24 Concept Prevalence Study sites (Supplementary Figure 3). A Chi-Square test of independence with Yate's correction revealed a significant association between the sites and OMOP2OBO coverage ($\chi^2(23) = 7{,}559.1$, $p<0.0001$). Post-hoc tests using Bonferroni adjustment confirmed that 38.8% of the pairwise site comparisons had significantly different OMOP2OBO coverage (ps<0.001). The OMOP2OBO concept count by OBO Foundry ontology, data wave, and coverage type are shown in Supplementary Figure 4.

Error Analysis

Results for the 4,672 concepts used in at least one Concept Prevalence Study site but missing from OMOP2OBO are visualized in Figure 6b. Roughly 7.9% (n=367) of concepts were accounted for using a newer version of the OMOP CDM and occurred in an average of 2.6 sites with a mean frequency of 27,412.3 (max=3,539,698.5). 90.6% (n=4,231) of concepts purposefully excluded from the OMOP2OBO mapping set (i.e., no clear pathological or biological origin) occurred in an average of 1.7 sites with a mean frequency of 6,139.3 (max=8,254,186.5). The remaining concepts (1.6%; n=74) were truly missing and occurred in an average of 2.7 sites with a mean frequency of 5,320 (max=100,483). The top-five most frequently occurring missing concepts (reported as the average frequency across the 24 sites and number of sites with that concept) were: (1) *Increased fluid intake* (SNOMED:249480002; n=100,483; one site); (2) *COVID-19* (SNOMED:840539006; n=93,585; one site); (3) *Polycystic ovary syndrome* (SNOMED:237055002; n=62,900.3; 3 sites); (4) *Saddle embolus of pulmonary artery with acute cor pulmonale* (SNOMED:15964701000119109; n=22,324.4; 10 sites); and (5) *Adjustment disorder with mixed anxiety and depressed mood* (SNOMED:782501005; n=18,453; one site). Domain expert review found that these concepts were likely missing due to

Figure 6: OMOP2OBO - Concept Prevalence Coverage.



differences in patient populations and coding practices. Comparable concepts in OMOP2OBO were identified.

Drug Ingredients

The Concept Prevalence Study contained data for 4,588 concepts from 18 sites. The

OMOP2OBO mapping set contained 8,615 eligible concepts, which covered 87.9% (99.9% weighted coverage) of the Concept Prevalence Study concepts (n=4,037 concepts; average frequency: 8,071.6 [max=125,634,570.4]). Of the remaining concepts, 4,578 were only found in OMOP2OBO (average frequency: 468.9 [max=69,311]) and 551 were only found in the Concept Prevalence Study (average frequency: 801.2 [max=1,795,364.8]). These findings are visualized in Figure 6c. OMOP2OBO concept coverage ranged from 91.2-98.4% across the 18 Concept Prevalence Study sites (Supplementary Figure 5). A Chi-Square test of independence with Yate's correction revealed a significant association between the sites and OMOP2OBO coverage ($\chi^2(17) = 195.6$, $p<0.0001$). Post-hoc tests using Bonferroni adjustment confirmed that 34.6% of the pairwise site comparisons had significantly different OMOP2OBO coverage (ps<0.001). The OMOP2OBO concept count by OBO Foundry ontology, data wave, and coverage type are shown in Supplementary Figure 6.

Error Analysis

Results for the 551 concepts missing from OMOP2OBO are visualized in Figure 6d. Roughly 0.9% (n=5) of concepts were accounted for using a newer version of the OMOP CDM and occurred in an average of 8.4 sites with a mean frequency of 51,732 (max=221,229.7). 82.8% (n=456) of concepts purposefully excluded from the OMOP2OBO mapping set (i.e., no clear pathological or biological origin) occurred in an average of 3.9 sites with a mean frequency of 18,847.3 (max=1,077,258.9). The remaining concepts (16.3%; n=90) were truly missing and occurred in an average of 2.7 sites with a mean frequency of 3,361.2 (max=175,551.3). The top-five most frequently occurring missing concepts were (reported as the average frequency across the 18 sites and number of sites with that concept): (1) *hepatitis A virus strain CR 326F antigen, inactivated* (RxNorm:2274413; n=175,551.3; 14 sites); (2) *erenumab* (RxNorm:2045613; n=60,618; 10 sites); (3) *fremanezumab* (RxNorm:2056691; n=15,579.6; five sites); (4) *galcanezumab* (RxNorm:2058846; n=11,594.8; five sites); and (5) *baloxavir marboxil*

(RxNorm:2099995; n=11,366.7; three sites). Domain expert review of these concepts found that they were likely missing as a result of hospital vendor differences or because they were a new high-risk biologic whose safety and efficacy had not yet been tested or confirmed for use in pediatric populations. Comparable concepts in OMOP2OBO were identified.

Measurements

The Concept Prevalence Study contained data for 25,513 concepts from 18 sites. The resulting OMOP2OBO mapping set contained 3,827 eligible concepts (10,673 results), which covered 11.1% (67.7% weighted coverage) of the Concept Prevalence Study concepts (n=2,260 concepts; average frequency: 3,072.3 [max=183,333,482.4]). Of the remaining concepts, 1,207 were only found in OMOP2OBO (average frequency: 346.9 [max=842,485]) and 20,893 were only found in the Concept Prevalence Study (average frequency: 669.6 [max=1,219,846,862]). These findings are visualized in Figure 6e. OMOP2OBO concept coverage ranged from 91.2-98.4% across the 18 Concept Prevalence Study sites (Supplementary Figure 7). A Chi-Square test of independence with Yate's correction revealed a significant association between the sites and OMOP2OBO coverage ($\chi^2(17) = 3{,}872.3$, $p<0.0001$). Post-hoc tests using Bonferroni adjustment confirmed that 60.8% of the pairwise site comparisons had significantly different OMOP2OBO coverage (ps<0.001). The OMOP2OBO concept count by OBO Foundry ontology, data wave, and coverage type are shown in Supplementary Figure 8.

Error Analysis

Results for the 20,893 concepts missing from OMOP2OBO are visualized in Figure 6f. Roughly 0.1% (n=13) of concepts were accounted for using a newer version of the OMOP CDM and occurred in an average of 3.2 sites with a mean frequency of 9,836.3 (max=221,229.7). 0.8% (n=158) of concepts purposefully excluded from the OMOP2OBO mapping set (i.e., no clear pathological or biological origin) occurred in an average of 5.2 sites with a mean frequency of 282,115.3 (max=14,317,951.9). The remaining concepts (99.2%; n=20,722) were truly missing

and occurred in an average of 2.8 sites with a mean frequency of 218,874.1 (max=1,219,846,862). The top-five most frequently occurring missing concepts were (reported as the average frequency across the 18 sites and number of sites with that concept): (1) *Pulse intensity of Unspecified artery palpation* (LOINC:44974-4; n=1,219,846,862; one site); (2) *Penicillin G potassium [Mass] of Dose* (LOINC:4380-2; n=253,609,945; one site); (3) *Sodium [Moles/volume] in Saliva (oral fluid)* (LOINC:56979-8; n=246,641,211; one site); (4) *Cotinine/Creatinine [Mass Ratio] in Urine* (LOINC:44311-9; n=246,063,202; one site); and (5) *Chloride [Moles/volume] in Saliva (oral fluid)* (LOINC:2074-3; n=234,931,483; one site). Domain expert review of these concepts confirmed that they were likely missing due to inconsistencies in hospital use of LOINC, a finding that's been observed in literature.[74] Comparable concepts in OMOP2OBO were identified.

Clinical Utility

The goal of this evaluation was to examine the clinical utility of the OMOP2OBO mappings when used to identify undiagnosed rare disease patients using data from the All of Us Research Program (AoU).[75] OMOP2OBO mappings that aligned HPO concepts to OMOP condition concepts were compared to Phecode mappings that aligned HPO concepts to International Classification of Diseases [ICD] codes.[38] We assessed the 73 American College of Medical Genetics and Genomics (ACMG) secondary finding genes (ACMG-73),[76] which contain pathogenic variants found to be causative for at least 35 genetic diseases and 2,257 HPO concepts. When querying AoU patients, the Phecode mappings (n=7,815 ICD codes) took ~30 minutes to complete and returned 201,423 patients and the OMOP2OBO mappings (n=3,783 OMOP concepts) took ~10 minutes to complete and returned 198,815 patients. 198,391 patients were found in common, 3,032 patients were only identified by the Phecode mappings, and 424 patients were only identified by the OMOP2OBO mappings. Phenotype Risk Scores (PheRS),[77] which identify patients with clinical features similar to ACMG-73-related genetic

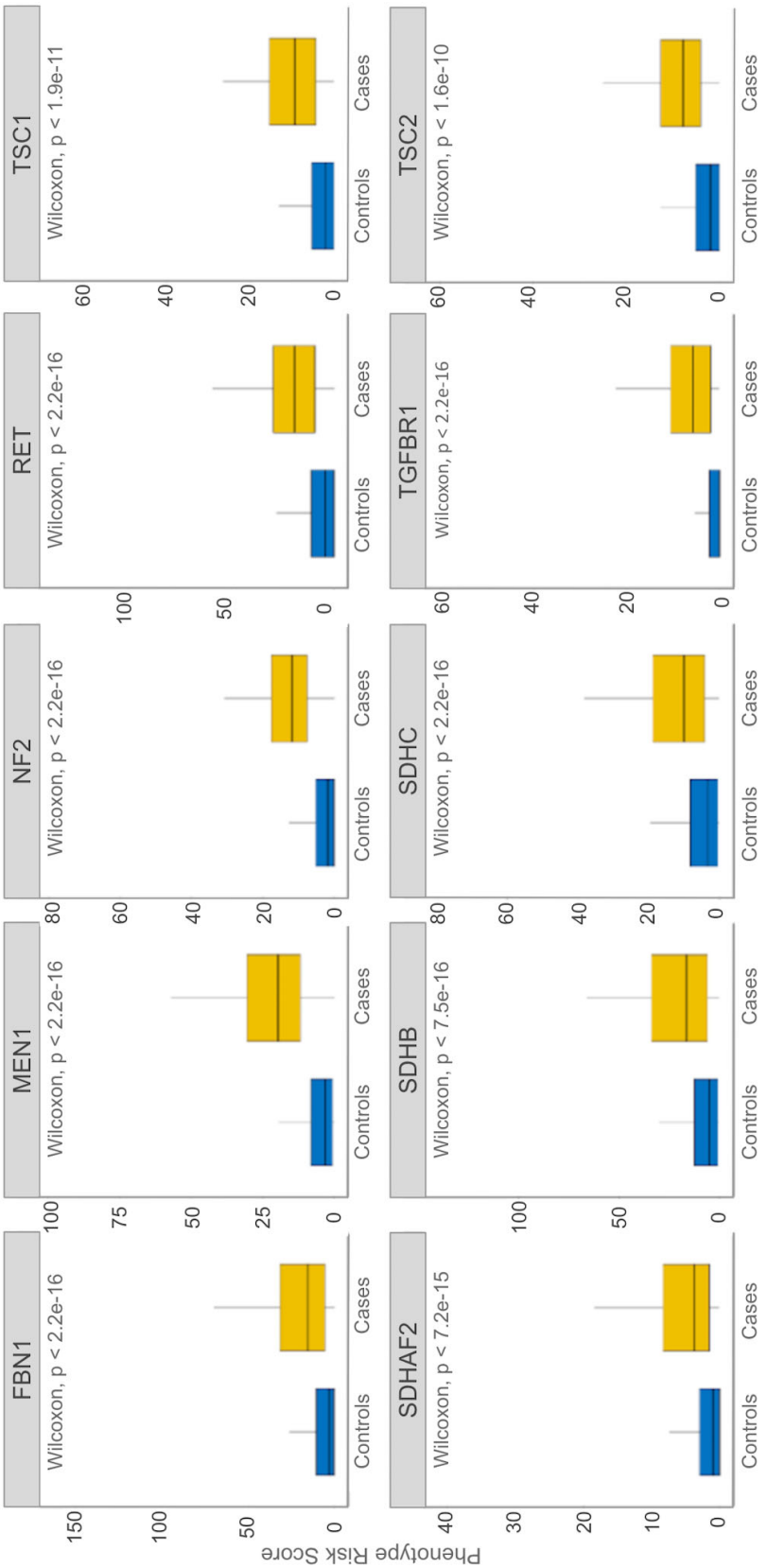
diseases, were highly correlated for both mappings (r^2>0.6 across all diseases). As validation, PheRS of cases (n=504) identified by the OMOP2OBO mappings were compared with controls (n=68,234) for 10 gene-disease pairs, in which the diseases had related diagnosis codes of high positive predictive values. Cases were found to be significantly different than controls for all gene-disease pairs (ps<0.001; Figure 7). These results demonstrate that the OMOP2OBO mappings can aid in the systematic identification of undiagnosed rare disease patients who might benefit from genetic testing.

DISCUSSION

Precise phenotyping is viewed as one of the biggest barriers to a deeper understanding of the genetic and mechanistic basis of human disease.[13] EHR-based deep phenotyping has great potential to advance precision medicine.[7,8] Despite significant progress, this promise remains largely aspirational.[6,9–12] We developed OMOP2OBO to map standard vocabularies in the OMOP CDM to OBO Foundry ontologies and created mappings for 92,367 condition, 8,615 drug ingredient, and 10,673 measurement result concepts. To the best of our knowledge, the OMOP2OBO mappings are the largest and most comprehensive set of publicly available mappings between clinical vocabularies and OBO Foundry ontologies.

Our work differs from existing work, which has largely focused on using ontologies to improve phenotyping in specific diseases (e.g., infectious disease,[54] rare diseases,[55,56] and cancer[57]) and for the investigation of specific biological (e.g., glycobiology[58]) and clinical domains (e.g., laboratory test results[44] and medical diagnoses[59,60]). Our work is most similar to LOINC2HPO,[44] which we have included and expanded in our current mapping set. OMOP2OBO complements existing phenotyping efforts like the Electronic Medical Records and Genomics (eMERGE) Network[78] and the AoU Research Program,[75] by providing access to resources not currently available in EHRs and opportunities to improve the semantic interoperability of definitions through alignment to the OBO Foundry ontologies. One potential use of OMOP2OBO

Figure 7: Phenotype Risk Scores by Disease and Mapping Set for Cases and Controls.



is to aid in the alignment of patient data to ontologies in the Global Alliance for Genomics and Health's Phenopacket schema,[79] which was designed to support the global exchange of computable patient-level phenotypic information.

Several recent publications have demonstrated the value of the OMOP2OBO mappings. The OMOP2OBO mappings have been used to characterize differences in definitions of long COVID,[80] generate long COVID phenotypes,[81,82] and to improve the categorization and prediction of psychiatric diseases among patients with long COVID.[83] Additionally, our recent work in pediatric rare disease subphenotyping demonstrated that patient representations constructed from the OMOP2OBO mappings produced more clinically meaningful clusters than representations built using OMOP concepts alone.[84] We further demonstrated the value of the mappings by leveraging them to successfully integrate external gene expression data from an independent sample of pediatric patients resulting in more clinically-meaningful and biologically-actionable phenotypes than those generated using only clinical data.

In this work, we examined whether the OMOP2OBO mappings could identify undiagnosed rare disease patients, a problem which has been frequently identified in the literature.[85–87] The PheRS, when used with Phecodes,[88] has shown great promise when used to identify underdiagnosed rare disease patients using only EHR data.[38,77] Using AoU data and the ACMG-73 gene list, we demonstrated that queries using the OMOP2OBO mappings identified 98.5% of the patients identified by the Phecode mappings using fewer codes. Additionally, the PheRS results demonstrated that the OMOP2OBO mappings were highly correlated to each genetic disease and performed comparably with the Phecode mappings, the current state-of-the-art for this task.[38] Our applied example of using deep phenotyping to identify undiagnosed rare diseases validates the clinical utility of OMOP2OBO and demonstrates the tremendous promise of the mappings. Further application of OMOP2OBO is needed to help elucidate its full potential.

Limitations and Future Work

OMOP2OBO has not been optimized for performance; all possible ancestors are mapped when unable to generate a mapping at the concept-level. A prioritization strategy would significantly improve performance. OMOP2OBO does not take advantage of all of the knowledge available in the UMLS. Leveraging information in the mapping and hierarchy tables could improve the automatically map concepts and would enable use of other UMLS-aligned resources like SemMedDB.[89] We only evaluated the accuracy of a small subset of the manual mappings. It is important to evaluate the remaining manually derived mappings as well as update all of the mappings with citations from the resources that they were derived from. Similarly, it is important to evaluate the automatically derived mappings as their accuracy depends upon the quality of the resources from which they were built and ontologies are subject to a variety of errors.[90–92] The *Accuracy* evaluation revealed limitations of our expert review procedures; some of the experts experienced challenges when trying to use the OBO ontologies, which may have negatively impacted the results. Future evaluations will provide better training and outcomes other than correct/incorrect will be considered. OMOP standard clinical vocabularies are also dependent upon a large set of CDM-specific mappings and may be subject to similar errors as our mappings. In the future, we need to evaluate the logical consistency of the mappings and output them using Semantic Web standards like RDF/XML and the Simple Standard for Sharing Ontological Mappings or SSSOM.[93]

METHODS

OMOP2OBO is open source (https://github.com/callahantiff/OMOP2OBO), available on PyPI (https://pypi.org/project/omop2obo/), and includes an interactive dashboard that summarizes the current mapping set (http://tiffanycallahan.com/OMOP2OBO_Dashboard/). We also created a dedicated Zenodo Community, which provides access to data, mappings, and presentations (https://zenodo.org/communities/omop2obo). A list of the acronyms used in this paper are

provided in Supplementary Table 1 and the resources used by OMOP2OBO are described in Supplementary Table 2.

OMOP2OBO Framework

Mapping Resources

The National Library of Medicine's UMLS[71] MRCONSO and MRSTY tables (2020AA version[94]) were used to annotate each OMOP concept with a UMLS CUI and a Semantic Type.[72]

Overview

The OMOP2OBO framework (Figure 1) consists of five components:

1. **Process OMOP Data.** The framework takes as input a table of OMOP concepts including concept identifiers, codes, labels, synonyms, and concept ancestors.
2. **Process OBO Foundry Ontologies.** Using OWLTools (April 06, 2020 release),[95] one or more OBO Foundry ontologies are downloaded and current classes, dbXRefs, labels, and synonyms are extracted.
3. **Map OMOP Standard Vocabulary Concepts to OBO Foundry Ontology Concepts.** This step consists of two tasks: (1) Concept Alignment: exact-string matches between OMOP and OBO Foundry ontology concept labels, definitions, and synonyms are obtained. Prior to alignment, the label and synonym fields are both made lowercase. This step also obtains exact matches between OMOP standard concepts and source codes to OBO Foundry ontology dbXRefs. To increase the likelihood of finding a match, the OMOP standard concepts and source codes are first merged with terminologies in the UMLS using core functionality from OHDSI Ananke,[96] a program developed to align OMOP concepts to UMLS CUIs. Prior to performing this alignment, the OMOP standard concepts and source codes and the OBO Foundry ontology dbXRefs are normalized using a custom dictionary (source_code_vocab_map.csv[97]). This resource ensures that concepts referenced by the same code using different prefixes or symbols can be aligned (e.g.,

SNOMED:1234567 and snomed_1234567). If a mapping at the concept-level cannot be found, mappings to the concept's ancestor are attempted and (2) Concept Embedding: using scikit-learn,[98] a bag-of-words[99] vector space model with term-frequency inverse-document frequency[100] and L2 normalization is used to learn embeddings from labels and synonyms for all OMOP and OBO Foundry ontology concepts and concept ancestors. Prior to building the model, all text fields are made lowercase, stop words are removed using the wordnet list from Python's NLTK library,[101] white spaces are removed, and word-level tokenization and lemmatization are applied. Next, cosine similarity is used to compute scores between all pairwise combinations of OMOP and OBO Foundry ontology concepts and ancestor concepts. To improve the efficiency of this process, only the top 75% of pairs with scores >=0.25 are output, which was decided after visualizing the score distribution using a histogram. All thresholds and cut-offs are customizable. All OMOP concepts unable to be automatically mapped will require manual curation.

4. **Synthesize and Process Mapping Results.** Each mapping includes a category and human-readable evidence. The mapping category is constructed by combining the following elements: (1) the approach used to create it (i.e., "automatic", "manual", or "cosine similarity"), (2) cardinality (i.e., one-to-one or one-to-many), and (3) level (i.e., concept or ancestor). Mapping evidence consists of pipe-delimited free-text phrases that explain what fields were used to construct the mapping. Supplementary Table 5 provides additional details on and examples of the mapping categories.
5. **Output Mappings.** Mappings can be output in a variety of file types, like flat file, database dump, or RDF/XML file.

OMOP2OBO Mappings

Figure 2 includes example mappings and illustrates how the OBO Foundry ontologies were used to map OMOP concepts from each clinical domain. Supplementary Table 5 provides

additional details on and examples of the mapping categories.

Mapping Data

OMOP Data

OMOP concepts were extracted from a de-identified copy of the Children's Hospital Colorado pediatric OMOP database stored within University of Colorado Anschutz Medical Campus Health Data Compass infrastructure (created in October 2018).[102] The data conformed to the structure defined by the National Pediatric Learning Health System (PEDSnet) OMOP CDM, which is an adaptation of the OMOP CDM version 5.0.[39,68] Use of these data was approved by the Colorado Multiple Institutional Review Board (#15-0445).

Concept lists were derived from standard OMOP vocabularies (i.e., SNOMED-CT[69] [v20180131], RxNorm[70] [v20180507], and LOINC[45] [v2.64]) from the Condition Occurrence, Drug Exposure, and Measurement tables. There is one exception, one frequently used measurement concept was from a local pediatric-specific source vocabulary. Two waves of data were utilized: (1) all concepts associated with at least one patient and visit occurrence in the PEDSnet OMOP database (i.e., -- *Standard Concepts Used in Practice*); and (2) all standard OMOP concepts not used in clinical practice (i.e., *Standard Concepts Not Used in Practice)*. For each concept set, additional metadata were extracted from the OMOP CDM including concept codes (i.e., codes from each standard vocabulary), labels, synonyms, and ancestor concepts (codes, labels, and synonyms were also extracted for each concept ancestor). Additional information for each concept set is available on the project's GitHub Wiki (https://github.com/callahantiff/OMOP2OBO/wiki). The OMOP CDM is built on an extensive set of mappings between source codes, standard concepts, and source codes and standard concepts. While we leverage these mappings when building ours (i.e., leveraging source codes mapped to standard concepts), we do not verify their quality.

Data Preprocessing

No preprocessing was required for concepts from the Condition Occurrence table. Concepts from the Drug Exposure table were extracted at the ingredient-level. For concepts from the Measurement table, a scale and result type were created. The scale (i.e., ordinal, nominal, quantitative, qualitative, narrative, doc, and panel) of each measurement was identified from the OMOP CDM or by parsing the concept synonym field. For all *Standard Concepts Used in Practice,* reference ranges were used to determine the result type; concepts with numeric reference ranges were typed as "Normal/Low/High" and concepts with reference ranges that included "positive" or "negative" were typed as "Positive/Negative". *Standard Concepts Not Used in Practice* with an ordinal scale or with synonyms that contained the words "presence" or "screen" were typed as "Positive/Negative". Concepts with a quantitative scale were typed as "Normal/Low/High". All other scale types were typed as "Unknown Result Type". While it is possible to infer the result type from the scale type (e.g., all concepts with a quantitative scale have result type "Normal/Low/High" and all concepts with an ordinal scale have result type "Positive/Negative"), our approach was developed to maximize the inclusion of concepts from all scale types.

OBO Foundry Ontologies

OBO Foundry ontologies were selected under the advice of several clinicians, molecular biologists, and professional OBO Foundry biocurators to cover the following domains: diseases (Mondo[48] [v2020-09-14]), phenotypes (HPO[46] [v2020-08-11]), anatomical entities (CL[63] [v2020-05-21], Uberon[62] [v2020-06-30]), organisms (NCBITaxon[64] [v2020-04-18]), chemicals (ChEBI[65] [v191]), vaccines (VO[66] [v1.1.102]), and proteins (PRO[67] [v61.0]). Similar to the clinical concepts, each ontology was queried to obtain labels, definitions, synonyms (including synonym type), and dbXRefs. All OBO Foundry ontologies were downloaded in September 2020 using OWLTools (April 06, 2020 release).[95] Similar to the OMOP CDM, the OBO Foundry ontologies

contain automatically and manually-derived mappings. While we leverage these mappings when building ours (i.e., dbXRefs mapped to ontology concepts), we do not perform any verification of their quality.

Mapping Constraints

Some additional constraints were applied to Condition Occurrence and Measurement concepts in order to ensure the mapping process was reproducible and as a means to prioritize concepts requiring manual mapping.

Conditions

For *Standard Concepts used in Practice*, UMLS Semantic Types were used to identify all concepts that had a clear pathological or biological origin. All remaining concepts (e.g., accidents, injuries, external complications, and findings without clear interpretations) were marked as unmapped and the reason for exclusion was provided in the evidence field. The Semantic Types were also used to group OMOP concepts such that those typed as “Findings” or “Signs and Symptoms” were treated as phenotypes and only mapped to HPO and concepts typed as “Disease or Syndrome” were only mapped to Mondo. For *Standard Concepts Not Used in Clinical Practice*, all possible automatic mappings were obtained and concepts which were unable to be mapped automatically were marked as unmapped and “NOT YET MAPPED” was provided as the mapping evidence. This same approach was applied to drug ingredients.

Measurements

Mappings were created for each result type using the procedures defined by LOINC2HPO[44]; results were annotated with respect to their result type: **Concepts with result type “Normal/Low/High”.** For example, *Corticotropin [Mass/volume] in Plasma --4th specimen post XXX challenge* (LOINC:12460-2). Results above the reference range are mapped to *Increased circulating ACTH level* (HP:0003154). Results below the reference range are mapped to *Decreased circulating ACTH level* (HP:0002920). Results within the reference are mapped to

Abnormality of circulating adrenocorticotropin level and logically negated (NOT HP:0011043).

Concepts with result type "Positive/Negative". For example, *Amphetamine [Presence] in Urine by Screen Method* (LOINC:19343-3). Positive results are mapped to *Positive urine amphetamine test* (HP:0500112). Negative results are mapped to *Positive urine amphetamine test* and logically negated (NOT HP:0500112). Also consistent with the procedures adopted by LOINC2HPO, all concepts lacking sufficient detail (i.e., non-specific body substances) were marked as unmapped and "Unspecified Sample" was provided as the mapping evidence.

LOINC2HPO Extensions

The initial set of measurement concepts was supplemented with the latest LOINC2HPO annotations,[44] which was downloaded on 08/02/2020 from the LOINC2HPO annotation Github repository.[103] OMOP2OBO expands the LOINC2HPO mappings by including the measurement substance (i.e., body fluids, tissues, and organs via Uberon), the entity being measured (i.e., chemicals, metabolites, or hormones via ChEBI; cell types via CL; and proteins via PRO), and the species of the measured entity (i.e., organism taxonomy via NCBITaxon). All modifications to the original LOINC2HPO annotations were recorded in the mapping evidence field, enabling users to easily identify when an original LOINC2HPO annotation had been updated.

Mapping Evaluation

The accuracy, generalizability, and clinical utility of mappings were evaluated.

Accuracy

For conditions and drug ingredients, 20% of the manual one-to-many mappings (n=2,000 conditions; n=16 drug ingredients) were manually verified by a practicing resident physician and clinical pharmacist, respectively. Only mappings to *Standard Concepts Used in Practice* were evaluated. Measurement mappings to HPO were evaluated in two ways: (1) Survey. A subset of the mappings (n=270) were independently validated by five domain experts including three practicing pediatric clinicians, a PhD-level molecular biologist, and a master's-level

epidemiologist using a Qualtrics Survey.[104] Any mapping that did not meet agreement by at least one clinician and both the biologist and the epidemiologist were re-evaluated by the most senior clinician. These mappings were also vetted on the LOINC2HPO GitHub tracker[105] by members of the biocuration team. (2) Biocurator Validation. A random subset of 1,350 measurement results were manually verified by an OBO Foundry biocurator. Additional details are provided on the project's GitHub Wiki (https://github.com/callahantiff/OMOP2OBO/wiki/Accuracy).

Generalizability

The generalizability of the OMOP2OBO mappings (only mappings to *Standard Concepts Used in Practice*) were examined using data from the OHDSI Concept Prevalence Study.[73] The Concept Prevalence study was designed to provide researchers with additional context regarding the frequency at which different OMOP concepts are used in clinical practice across the OHDSI network. In addition to the Concept Prevalence Study sites (n=22), data were obtained from two independent academic medical centers, bringing the total number of sites to 24. Consistent with the Concept Prevalence Study procedures, all concepts occurring fewer than 10 times were removed and all remaining concepts occurring fewer than 100 times were assigned a count of 100. The OMOP2OBO mappings were filtered to remove all concepts without at least one ontology mapping.

Coverage of all standard OMOP concepts in the OMOP2OBO mapping set was assessed by identifying: (1) concepts that existed in the OMOP2OBO set and in at least one Concept Prevalence Study site (i.e., Overlap); (2) concepts only present in the OMOP2OBO set (i.e., OMOP2OBO Only); and (3) concepts only present in the Concept Prevalence Study set (i.e., Concept Prevalence Only). An error analysis was performed to examine the Concept Prevalence Only concept set. Three scenarios were examined: (i) CDM Versioning**:** concepts that could be recovered using a newer version of the OMOP CDM (v5.3.1; 02/25/2022); (ii) Excluded Concepts**:** concepts without clear pathological or biological origin that were

purposefully excluded from the OMOP2OBO mapping set; and (iii) Truly Missing**:** concepts that could not be accounted for using the prior two scenarios. For all scenarios, concept frequency within the Concept Prevalence Study sites was used as a measure of concept importance. Findings from each scenario were reviewed by a practicing resident physician and a clinical pharmacist. Detailed procedures and timelines are provided on GitHub (https://github.com/callahantiff/OMOP2OBO/wiki/Generalizability).

Clinical Utility

The PheRS[77] can be used to identify patients who are clinically similar to Online Mendelian Inheritance in Man (OMIM)[106] Mendelian profiles but lack formal diagnosis and has demonstrated utility for identifying underdiagnosed rare disease patients using only EHR data.[38,77] We examined whether OMOP2OBO mappings could be used to help identify undiagnosed rare disease patients who lacked relevant diagnosis codes in their clinical records. For this evaluation, ACMG-73 genes (v3.0), which have specific mutations known to cause disorders, have well-defined phenotypes, and are clinically actionable, were used to generate a list of Mendelian diseases.[76] The OMIM[106] database was used to identify the Mendelian diseases associated with each gene, which resulted in 35 genetic diseases. Aligning the genes to phenotypes, using the HPO gene annotation table,[107] produced a list of 2,257 HPO concepts. To calculate the phenotypic burden of each genetic disease, OMOP concepts from the OMOP2OBO HPO mappings (v2.0.0 beta) and ICD concepts from the Phecode HPO to ICD mappings[38] were queried against AoU data[75] (v6; n=230,000 patients). PheRS for each gene were then calculated for patients from each the OMOP2OBO and Phecode mapping sets. As a validation, 10 gene-disease pairs, for which diagnosis codes were available and showed good prediction were examined: (i) NF2 with neurofibromatosis; (ii) SDHAF2, SDHB, and SDHC with paragangliomas; (iii) MEN1 and RET with multiple endocrine neoplasia; (iv) TSC1 and TSC2 with tuberous sclerosis complex; and (v) FBN1 and TGFBR1 with Marfan Syndrome. For each

gene, a one-sided Wilcoxon rank sum test was performed in order to determine if PheRS were significantly higher for cases than controls. Cases were defined as patients with at least two occurrences of a relevant diagnosis code and control patients had no instances of these codes. Cases and controls were matched on age, sex, and length of EHR record. Results were verified by a PhD-level Epidemiologist specializing in genetics (CZ).

Technical Specifications

OMOP2OBO was developed using Python 3.6.2 on a single machine with 8 cores and 16GB of RAM. All code and project information are publicly available and detailed on GitHub (https://github.com/callahantiff/OMOP2OBO). The OMOP2OBO (v1.0) mappings are publicly available from Zenodo.[108–110] The OMOP2OBO Mapping Dashboard was built with R (v4.2.1) using Rmarkdown (v2.14) and flexdashboard (v0.5.2).

Descriptive and inferential statistics were performed to evaluate the data available for mapping and the OMOP2OBO mapping set. Chi-square tests of independence with Yate's correction were used to: (1) assess differences in the proportions of metadata available from each OBO Foundry ontology; and (2) assess differences in the proportions of mapped concepts between OHDSI Concept Prevalence sites. Post-hoc tests using Bonferroni adjustment to correct for multiple comparisons were performed for significant omnibus tests. Analyses were performed in Jupyter Notebooks (v6.1.6) using the scipy (v1.4.1), statsmodels (v0.12.1), statistics (v1.0.3.5), and numpy (v.1.18.1) libraries. Visualizations were created using matplotlib (v.3.3.2). The *Clinical Utility* evaluation was performed in the AoU Researcher Workbench[111] using R (v4.1.2) and Python (v3.7). Analyses were performed on a machine with 16 CPUs and 60GB of memory.

DATA AVAILABILITY

Supplementary Table 2 provides a complete list of the resources used in this project. OMOP concepts are available for download through Athena (https://athena.ohdsi.org/). The MRCONSO and MRSTY tables (2020AA) are available through the UMLS (https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html). The OBO Foundry ontologies are publicly available (https://obofoundry.org/). The OMOP2OBO v1.0 mappings can be downloaded from Zenodo: Condition Occurrence (https://doi.org/10.5281/zenodo.6774363); Drug Exposure Ingredients (https://doi.org/10.5281/zenodo.6774401); and Measurements (https://doi.org/10.5281/zenodo.6774443).

CODE AVAILABILITY

The OMOP2OBO is available through GitHub (https://github.com/callahantiff/OMOP2OBO) and PyPI (https://pypi.org/project/omop2obo/).

ACKNOWLEDGEMENTS

This work was supported by funding from the National Library of Medicine (T15LM009451) to LEH and (T15LM007079) to GH. The authors thank colleagues at the Health Data Compass Warehouse and OMOP2OBO and Machine Learning Working Groups at the National COVID Cohort Collaboration (NCATS U24TR002306) for piloting testing the mappings. The authors would also like to thank Drs. Paul Schofield or his feedback on the mappings and the Denny Lab (National Human Genome Research Institute) for reviewing the mappings. The All of Us Research Program would not be possible without the partnership of its participants.

AUTHOR CONTRIBUTIONS

MGK and LEH served as primary supervisors of this work. TJC, MGK, and ALS conceived and

developed the analyses. TJC and WAB developed the OMOP2OBO framework with feedback from NAV and JMB. ALS and JMW helped develop documentation. RDB, AO, PBR, GH, DM, SJDD, and AW provided data for the evaluation. PNR, XAZ, MAH, NAM, SBT, EC, BDC, BM, JS, AYL, JHC, JR, JMW, ALS, JAF, TDB, NAV, KET, and CZ reviewed, evaluated or aided in pilot testing the mappings and/or assisted with the error analysis. CZ performed the *Clinical Utility* evaluation. TJC drafted the manuscript and all authors reviewed it and provided feedback. All authors read and approved the final version of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1 Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. *Health Aff* 2017; **36**: 1416–1422.
2 Atasoy H, Greenwood BN, McCullough JS. The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization. *Annu Rev Public Health* 2019; **40**: 487–500.
3 Dexter PR, Perkins S, Overhage JM, Maharry K, Kohler RB, McDonald CJ. A computerized reminder system to increase the use of preventive care for hospitalized patients. *N Engl J Med* 2001; **345**: 965–970.
4 King J, Patel V, Jamoom EW, Furukawa MF. Clinical benefits of electronic health record use: national findings. *Health Serv Res* 2014; **49**: 392–404.
5 Evans RS. Electronic Health Records: Then, Now, and in the Future. *Yearb Med Inform* 2016; **Suppl 1**: S48–61.
6 Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S *et al.* From Big Data to Precision Medicine. Frontiers in Medicine. 2019; **6**. doi:10.3389/fmed.2019.00034.
7 Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet* 2016; **17**: 353–373.
8 Rossi RL, Grifantini RM. Big Data: Challenge and Opportunity for Translational and Industrial Research in Healthcare. *Frontiers in Digital Humanities* 2018; **5**: 13.
9 Jha S, Topol EJ. Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. *JAMA* 2016; **316**: 2353–2354.
10 Butte AJ. Big data opens a window onto wellness. Nat. Biotechnol. 2017; **35**: 720–721.
11 Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA* 2018; **319**: 1317–1318.
12 Hinton G. Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA* 2018; **320**: 1101–1102.
13 Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat* 2012; **33**: 777–780.
14 Delude CM. Deep phenotyping: The details of disease. *Nature* 2015; **527**: S14–5.
15 Weng C, Shah NH, Hripcsak G. Deep phenotyping: Embracing complexity and temporality-Towards scalability, portability, and interoperability. *J Biomed Inform* 2020; **105**: 103433.
16 Dorsey ER, Omberg L, Waddell E, Adams JL, Adams R, Ali MR *et al.* Deep Phenotyping of Parkinson's Disease. *J Parkinsons Dis* 2020; **10**: 855–873.
17 Georgiou M, Robson AG, Singh N, Pontikos N, Kane T, Hirji N *et al.* Deep Phenotyping of PDE6C-Associated Achromatopsia. *Invest Ophthalmol Vis Sci* 2019; **60**: 5112–5123.
18 Fassihi H, Sethi M, Fawcett H, Wing J, Chandler N, Mohammed S *et al.* Deep phenotyping of 89 xeroderma pigmentosum patients reveals unexpected heterogeneity dependent on the precise molecular defect. *Proc Natl Acad Sci U S A* 2016; **113**: E1236–45.
19 Russo RS, Gambello MJ, Murphy MM, Aberizk K, Black E, Lindsey Burrell T *et al.* Deep phenotyping in 3q29 deletion syndrome: recommendations for clinical care. Genetics in Medicine. 2021; **23**: 872–880.
20 Daich Varela M, Jani P, Zein WM, D'Souza P, Wolfe L, Chisholm J *et al.* The peroxisomal disorder spectrum and Heimler syndrome: Deep phenotyping and review of the literature. *Am J Med Genet C Semin Med Genet* 2020; **184**: 618–630.
21 Mei C, Fedorenko E, Amor DJ, Boys A, Hoeflin C, Carew P *et al.* Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. *Eur J Hum Genet* 2018; **26**: 676–686.
22 Droogmans G, Swillen A, Van Buggenhout G. Deep Phenotyping of Development, Communication and Behaviour in Phelan-McDermid Syndrome. *Mol Syndromol* 2020; **10**:

294–305.
23 Fernandes SA, Cooper GE, Gibson RA, Kishnani PS. Benign or not benign? Deep phenotyping of liver Glycogen Storage Disease IX. *Mol Genet Metab* 2020; **131**: 299–305.
24 Mak E, Bickerton A, Padilla C, Walpert MJ, Annus T, Wilson LR *et al.* Longitudinal trajectories of amyloid deposition, cortical thickness, and tau in Down syndrome: A deep-phenotyping case report. *Alzheimers Dement* 2019; **11**: 654–658.
25 Mishra R, Jain V, Gupta D, Saxena R, Kulshreshtha S, Ramprasad VL *et al.* Robinow Syndrome and Brachydactyly: An Interplay of High-Throughput Sequencing and Deep Phenotyping in a Kindred. *Mol Syndromol* 2020; **11**: 43–49.
26 Welsink-Karssies MM, Ferdinandusse S, Geurtsen GJ, Hollak CEM, Huidekoper HH, Janssen MCH *et al.* Deep phenotyping classical galactosemia: clinical outcomes and biochemical markers. *Brain Commun* 2020; **2**: fcaa006.
27 Shim Y, Go YJ, Kim SY, Kim H, Hwang H, Choi J *et al.* Deep phenotyping in 1p36 deletion syndrome. *Ann Child Neurol* 2020; **28**: 131–137.
28 Spedicati B, Cocca M, Palmisano R, Faletra F, Barbieri C, Francescatto M *et al.* Natural human knockouts and Mendelian disorders: deep phenotyping in Italian isolates. *Eur J Hum Genet* 2021; **29**: 1272–1281.
29 Yurkovich JT, Tian Q, Price ND, Hood L. A systems approach to clinical oncology uses deep phenotyping to deliver personalized care. *Nat Rev Clin Oncol* 2020; **17**: 183–194.
30 Papadimitriou K, Tsakirakis N, Malandrakis P, Vitsos P, Metousis A, Orologas-Stavrou N *et al.* Deep Phenotyping Reveals Distinct Immune Signatures Correlating with Prognostication, Treatment Responses, and MRD Status in Multiple Myeloma. *Cancers* 2020; **12**. doi:10.3390/cancers12113245.
31 Christopoulos P, Bozorgmehr F, Brückner L, Chung I, Krisam J, Schneider MA *et al.* Brigatinib versus other second-generation ALK inhibitors as initial treatment of anaplastic lymphoma kinase positive non-small cell lung cancer with deep phenotyping: study protocol of the ABP trial. *BMC Cancer* 2021; **21**: 743.
32 Sirinukunwattana K, Aberdeen A, Theissen H, Sousos N, Psaila B, Mead AJ *et al.* Improving the diagnosis and classification of Ph-negative myeloproliferative neoplasms through deep phenotyping. bioRxiv. 2019; : 762013.
33 Nagaoka K, Shirai M, Taniguchi K, Hosoi A, Sun C, Kobayashi Y *et al.* Deep immunophenotyping at the single-cell level identifies a combination of anti-IL-17 and checkpoint blockade as an effective treatment in a preclinical model of data-guided personalized immunotherapy. *J Immunother Cancer* 2020; **8**. doi:10.1136/jitc-2020-001358.
34 Song TH, Cao M, Min J, Im H, Lee H, Lee K. Deep Learning-Based Phenotyping of Breast Cancer Cells Using Lens-free Digital In-line Holography. bioRxiv. 2021; : 2021.05.29.446284.
35 Kuai R, Ochyl LJ, Bahjat KS, Schwendeman A, Moon JJ. Designer vaccine nanodiscs for personalized cancer immunotherapy. *Nat Mater* 2017; **16**: 489–496.
36 Paquette AG, Hood L, Price ND, Sadovsky Y. Deep phenotyping during pregnancy for predictive and preventive medicine. *Sci Transl Med* 2020; **12**. doi:10.1126/scitranslmed.aay1059.
37 Davidson L, Boland MR. Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. *Brief Bioinform* 2021; **22**. doi:10.1093/bib/bbaa369.
38 Bastarache L, Hughey JJ, Goldstein JA, Bastraache JA, Das S, Zaki NC *et al.* Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J Am Med Inform Assoc* 2019; **26**: 1437–1447.
39 Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012; **19**: 54–60.
40 Kho AN, Rasmussen LV, Connolly JJ, Peissig PL, Starren J, Hakonarson H *et al.* Practical

challenges in integrating genomic data into the electronic health record. *Genet Med* 2013; **15**: 772–778.
41 Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. *Brief Bioinform* 2015; **16**: 1069–1080.
42 Haendel MA, Chute CG, Robinson PN. Classification, Ontology, and Precision Medicine. *N Engl J Med* 2018; **379**: 1452–1462.
43 Haendel MA, McMurry JA, Relevo R, Mungall CJ, Robinson PN, Chute CG. A Census of Disease Ontologies. *Annu Rev Biomed Data Sci* 2018; **1**: 305–331.
44 Zhang XA, Yates A, Vasilevsky N, Gourdine JP, Callahan TJ, Carmody LC *et al.* Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *npj Digital Medicine* 2019; **2**: 1–9.
45 McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R *et al.* LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem* 2003; **49**: 624–633.
46 Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res* 2019; **47**: D1018–D1027.
47 Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 2010; **26**: 1112–1118.
48 Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M *et al.* The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* 2017; **45**: D712–D722.
49 Vasilevsky NA, Matentzoglu NA, Toro S, Flack JE IV, Hegde H, Unni DR *et al.* Mondo: Unifying diseases for the world, by the world. bioRxiv. 2022. doi:10.1101/2022.04.13.22273750.
50 Open Targets Team. Open Targets. 2017.https://www.opentargets.org/.
51 Sun H, Depraetere K, De Roo J, Mels G, De Vloed B, Twagirumukiza M *et al.* Semantic processing of EHR data for clinical research. *J Biomed Inform* 2015; **58**: 247–259.
52 Legaz-García M del C, Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT. A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. *Knowledge-Based Systems* 2016; **105**: 175–189.
53 Santo JM do E, do Espírito Santo JM, de Paula EV, Medeiros CB. Exploring Semantics in Clinical Data Interoperability. Lecture Notes in Computer Science. 2019; : 201–210.
54 Kafkas Ş, Abdelhakim M, Hashish Y, Kulmanov M, Abdellatif M, Schofield PN *et al.* PathoPhenoDB, linking human pathogens to their phenotypes in support of infectious disease research. *Sci Data* 2019; **6**: 79.
55 Thompson R, Papakonstantinou Ntalis A, Beltran S, Töpf A, de Paula Estephan E, Polavarapu K *et al.* Increasing phenotypic annotation improves the diagnostic rate of exome sequencing in a rare neuromuscular disorder. *Hum Mutat* 2019; **40**: 1797–1812.
56 Tang X, Chen W, Zeng Z, Ding K, Zhou Z. An ontology-based classification of Ebstein's anomaly and its implications in clinical adverse outcomes. *Int J Cardiol* 2020. doi:10.1016/j.ijcard.2020.04.073.
57 Edgren H, Mano B, Laaksonen M. Abstract 2276: Efficient curation and ontology mapping of clinical and phenotypic data. *Cancer Res* 2018; **78**: 2276–2276.
58 Gourdine J-PF, Brush MH, Vasilevsky NA, Shefchek K, Köhler S, Matentzoglu N *et al.* Representing glycophenotypes: semantic unification of glycobiology resources for disease discovery. *Database* 2019; **2019**. doi:10.1093/database/baz114.
59 Dhombres F, Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies—Investigating partial mappings between HPO and SNOMED CT. *J Biomed Semantics* 2016; **7**: 3.

60 Raje S, Bodenreider O. Interoperability of Disease Concepts in Clinical and Research Ontologies: Contrasting Coverage and Structure in the Disease Ontology and SNOMED CT. *Stud Health Technol Inform* 2017; **245**: 925–929.
61 Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 2007; **25**: 1251–1255.
62 Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* 2012; **13**: R5.
63 Bard J, Rhee SY, Ashburner M. An ontology for cell types. *Genome Biol* 2005; **6**: R21.
64 Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2008; **36**: D13–21.
65 Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016; **44**: D1214–9.
66 Xiang Z, Todd T, Ku KP, Kovacic BL, Larson CB, Chen F *et al.* VIOLIN: vaccine investigation and online information network. *Nucleic Acids Res* 2008; **36**: D923–8.
67 Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M *et al.* The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res* 2011; **39**: D539–45.
68 Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA *et al.* PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc* 2014; **21**: 602–606.
69 SNOMED International. SNOMED. https://www.snomed.org/.
70 Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc* 2011; **18**: 441–448.
71 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; **32**: D267–70.
72 McCray AT. Representing biomedical knowledge in the UMLS semantic network. In: *High performance medical libraries: Advances in information management for the virtual era*. Meckler Corporation: USA, 1993, pp 45–55.
73 Ostropolets A, Ryan PB, Hripcsak G. OHDSI Network Study: Concept Prevalence. 2019.https://forums.ohdsi.org/t/network-study-concept-prevalence/6562.
74 Lin MC, Vreeman DJ, McDonald CJ, Huff SM. Auditing consistency and usefulness of LOINC use among three large institutions - using version spaces for grouping LOINC codes. *J Biomed Inform* 2012; **45**: 658–666.
75 U.S. Department of Health Services. Join the All of Us Research Program. https://go.joinallofus.org/.
76 Miller DT, Lee K, Chung WK, Gordon AS, Herman GE, Klein TE *et al.* ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med* 2021; **23**: 1381–1390.
77 Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* 2018; **359**. doi:10.1126/science.aal4043.
78 eMERGE Network. 2014.https://emerge-network.org/.
79 Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ *et al.* The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat Biotechnol* 2022; **40**: 817–820.
80 Rando HM, Bennett TD, Byrd JB, Bramante C, Callahan TJ, Chute CG *et al.* Challenges in defining Long COVID: Striking differences across literature, Electronic Health Records, and

patient-reported information. *medRxiv* 2021. doi:10.1101/2021.03.20.21253896.
81 Reese J, Blau H, Bergquist T, Loomba JJ, Callahan T, Laraway B *et al.* Generalizable long COVID subtypes: Findings from the NIH N3C and RECOVER programs. bioRxiv. 2022. doi:10.1101/2022.05.24.22275398.
82 Deer RR, Rock MA, Vasilevsky N, Carmody L, Rando H, Anzalone AJ *et al.* Characterizing Long COVID: Deep Phenotype of a Complex Condition. *EBioMedicine* 2021; **74**: 103722.
83 Coleman B, Casiraghi E, Callahan TJ, Blau H, Chan L, Laraway B *et al.* Manifestations associated with Post Acute Sequelae of SARS-CoV2 infection (PASC) predict diagnosis of new-onset psychiatric disease: Findings from the NIH N3C and RECOVER studies. bioRxiv. 2022. doi:10.1101/2022.07.08.22277388.
84 Callahan TJ, Hunter LE, Kahn MG. *Leveraging a Neural-Symbolic Representation of Biomedical Knowledge to Improve Pediatric Subphenotyping*. 2021 doi:10.5281/zenodo.5746187.
85 Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun* 2017; **8**: 15824.
86 Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheatle-Jarvela AM *et al.* Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N Engl J Med* 2018; **379**: 2131–2139.
87 Groopman EE, Marasa M, Cameron-Christie S, Petrovski S, Aggarwal VS, Milo-Rasouly H *et al.* Diagnostic Utility of Exome Sequencing for Kidney Disease. *N Engl J Med* 2019; **380**: 142–151.
88 Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017; **12**: e0175508.
89 Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* 2012; **28**: 3158–3160.
90 Amith M, He Z, Bian J, Lossio-Ventura JA, Tao C. Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *J Biomed Inform* 2018; **80**: 1–13.
91 Vrandečić D. Ontology Evaluation. In: Staab S, Studer R (eds). *Handbook on Ontologies*. Springer Berlin Heidelberg: Berlin, Heidelberg, 2009, pp 293–313.
92 Gómez-Pérez A. Ontology Evaluation. In: Staab S, Studer R (eds). *Handbook on Ontologies*. Springer Berlin Heidelberg: Berlin, Heidelberg, 2004, pp 251–273.
93 Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ *et al.* A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database* 2022; **2022**. doi:10.1093/database/baac035.
94 National Library of Medicine. UMLS release file archives: 2020AA. 2020.https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html.
95 OWL Collaboration. *OWLTools*. Github, 2020. https://github.com/owlcollab/owltools.
96 Banda JM. *OHDSIananke: OHDSI Ananke - A Tool for Mapping Between OHDSI Concept Identifiers to Unified Medical Language System (UMLS) identifiers*. Githubhttps://github.com/thepanacealab/OHDSIananke.
97 Callahan TJ. OMOP2OBO Code Normalization Dictionary. 2020.https://github.com/callahantiff/OMOP2OBO/blob/master/resources/mappings/source_code_vocab_map.csv.
98 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 2011; **12**: 2825–2830.
99 Harris ZS. Distributional Structure. *Word World* 1954; **10**: 146–162.
100 Rajaraman A, Ullman JD. Data Mining. In: *Mining of Massive Datasets*. Cambridge University Press, 2011, pp 1–17.

101 Bird S, Klein E, Loper E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. ‘O’Reilly Media, Inc.’, 2009https://play.google.com/store/books/details?id=KGIbfiiP1i4C.
102 Health Data Compass. https://www.healthdatacompass.org/.
103 Aaron ZX, Yates A, Vasilevsky NA, Gourdine JP, Callahan TJ, Carmody LC *et al.* LOINC2HPO Annotations. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. 2020.https://github.com/monarch-initiative/loinc2hpo/annotations.tsv.
104 Callahan TJ. Survey to Evaluate OMOP2OBO Measurement Mappings. Qualtrics. 2018.https://survey.az1.qualtrics.com/jfe/form/SV_cAZvVBV7LU0YVa5?Q_CHL=qr.
105 Monarch Initiative. *loinc2hpoAnnotation Issue Tracker*. Github. https://github.com/TheJacksonLaboratory/loinc2hpoAnnotation/issues.
106 Johns Hopkins University. Online Mendelian Inheritance in Man. An Online Catalog of Human Genes and Genetic Disorders. https://www.omim.org/.
107 The Human Phenotype Ontology. genes_to_phenotype.txt. 2022.http://purl.obolibrary.org/obo/hp/hpoa/genes_to_phenotype.txt.
108 Callahan TJ, Wyrwa JM, Vasilevsky NA, Bennett TD, Martin B, Feinstein JA *et al.* *OMOP2OBO Condition Occurrence Mappings*. 2020 doi:10.5281/zenodo.6949688.
109 Callahan TJ, Baumgartner WA, Hunter LD, Kahn MG. *OMOP2OBO Drug Exposure Ingredient Mappings*. 2020 doi:10.5281/zenodo.6949696.
110 Callahan TJ, Vasilevsky NA, Bennett TD, Martin B, Feinstein JA, Baumgartner WA *et al.* *OMOP2OBO Measurement Mappings*. 2020 doi:10.5281/zenodo.6949858.
111 U.S. Department of Health and Human. *The AllofUs Researcher’s Workbench*. 2022https://www.researchallofus.org/data-tools/workbench/.
112 Callahan TJ. *Overview of OMOP2OBO*. 2022 doi:10.5281/zenodo.6998826.

FIGURE LEGENDS

Figure 1: Overview of the OMOP2OBO Framework.
The OMOP2OBO framework[112] consists of five steps: (1) **Query OMOP CDM.** This step processes a table of OMOP concepts including identifiers, source codes, labels, synonyms, and concept ancestors. (2) **Process OBO Foundry Ontologies.** During this step, one or more OBO Foundry ontologies are downloaded and current classes, database cross-references, labels, and synonyms are extracted. (3) **Map OMOP Standard Vocabulary Concepts to OBO Foundry Ontology Concepts.** This step obtains three types of mappings and relies on publicly available resources like the UMLS Metathesaurus[94]. First, exact-string matches between OMOP and OBO Foundry ontology concept labels, definitions, and synonyms are obtained. Exact matches between OMOP standard concepts and source codes and OBO Foundry ontology database cross-references are also obtained. Then, a scoring metric is applied to embeddings learned from concept labels and synonyms. Manual mappings are accepted for all concepts unable to be automatically mapped. (4) **Synthesize and Process Mapping Results.** This steps generates a category (i.e., string constructed by combining: (i) the approach used to create it (i.e., “automatic”, “manual”, “cosine similarity”, or “unmapped”), (ii) cardinality (i.e., one-to-one or one-to-many), and (iii) level (i.e., concept or ancestor) and evidence (i.e., pipe-delimited free-text phrases that explain what fields were used to construct the mapping) for each mapping. (5) **Output Mappings.** Mappings are output as a flat-file, SQL database dump, or an RDF/XML file. Acronyms: OBO (Open Biological and Biomedical Ontology); OHDSI (Observational Health Data Sciences and Informatics); OMOP (Observational Medical Outcomes Partnership); UMLS (Unified Medical Language System).

Figure 2: OMOP2OBO mapping examples by OMOP clinical domain.
This figure illustrates which OBO Foundry ontologies were used for each OMOP clinical domain and provides example mappings. (**A**) OMOP conditions are mapped to HPO and Mondo. (**B**) OMOP drug ingredients are mapped to ChEBI, NCBITaxon, PRO, and VO. (**C**) OMOP measurements are mapped to ChEBI, CL, HPO, NCBITaxon, PRO, and Uberon. Acronyms: OMOP (Observational Medical Outcomes Partnership); UMLS (Unified Medical Language System); OBO (Open Biological and Biomedical Ontology); HP (Human Phenotype Ontology); MONDO (Monarch Disease Ontology); CHEBI (Chemical Entities of Biological Interest); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); PR (Protein Ontology); VO (Vaccine Ontology); UBERON (Uber-Anatomy Ontology); CL (Cell Ontology).

Figure 3: Condition Concept Frequency by Mapping Category and Ontology.
This figure presents the distributions of the frequency of OMOP standard condition concepts used at least once in clinical practice by mapping category and OBO Foundry ontology. Acronyms: HPO (Human Phenotype Ontology); Mondo (Monarch Disease Ontology); OBO (Open Biological and Biomedical Ontology); OMOP (Observational Medical Outcomes Partnership).

Figure 4: Drug Ingredient Concept Frequency by Mapping Category and Ontology.
This figure presents the distributions of the frequency of OMOP drug exposure ingredient concepts used at least once in clinical practice by mapping category and OBO Foundry ontology. Acronyms: ChEBI (Chemical Entities of Biological Interest); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); PRO (Protein Ontology); VO (Vaccine Ontology); OBO (Open Biological and Biomedical Ontology); OMOP (Observational Medical Outcomes Partnership).

Figure 5: Measurement Concept Frequency by Mapping Category and Ontology.
This figure presents the distributions of the frequency of OMOP measurement concepts used at least once in clinical practice by mapping category and OBO Foundry ontology. Acronyms: HPO (Human Phenotype Ontology); Uberon (Uber-Anatomy Ontology); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); PRO (Protein Ontology); ChEBI (Chemical Entities of Biological Interest); CL (Cell Ontology); OBO (Open Biological and Biomedical Ontology); OMOP (Observational Medical Outcomes Partnership).

Figure 6: OMOP2OBO - Concept Prevalence Coverage.
This figure presents the coverage of the OMOP2OBO mappings using Concept Prevalence Study data. Please note that the Y axes vary in scale. These kernel density estimation plots were created for condition concepts (**A** and **B**), drug ingredients (**C** and **D**), and measurement (**E** and **F**) results, where the distribution of the Overlap (OMOP concepts that exist in OMOP2OBO only sets and one or more Concept Prevalence sites), Concept Prevalence only and OMOP2OBO sets are shown on the left, for, respectively from top to bottom, (**A**) condition concepts, (**C**) drug ingredients, and (**E**) measurement results. On the right, the Error Analysis Concepts (concepts that can be accounted for in a newer OMOP CDM version), Excluded Set (purposefully or not yet mapped concepts), and Truly Missing (the concept's missingness cannot easily be accounted for). These distributions were created for (**B**) condition concepts, (**D**) drug ingredients, and (**F**) measurement results.

Figure 7: Phenotype Risk Scores by Disease and Mapping Set for Cases and Controls.
Boxplots of Phenotype Risk Score cases and controls for 10 gene-disease pairs in which the diseases had related diagnosis codes of high positive predictive values in the AllofUs (AoU) data (ps <0.001). The 10 gene-disease pairs were: (i) NF2 with neurofibromatosis; (ii) succinate dehydrogenase genes (i.e., SDHAF2, SDHB, and SDHC) with paragangliomas; (iii) MEN1 and RET with multiple endocrine neoplasia; (iv) TSC1 and TSC2 with tuberous sclerosis complex; and (v) FBN1 and TGFBR1 with Marfan Syndrome.

Table 1: OMOP Condition Concept Mapping Results by OMOP2OBO mapping method.

Ontology			HPO		Mondo	
Used in ≥ 1 Clinical Encounter			Yes	No	Yes	No
OMOP Concepts			24459	48959	20055	43320
Mapped Ontology Concepts			41384	183667	57848	396234
Mapping Category						
Automatic	One-to-One	Concept	3601	1166	4836	4261
		Ancestor	3155	10440	5961	2949
	One-to-Many	Concept	125	25	632	253
		Ancestor	1138	36947	4482	35743
Cosine Similarity	One-to-One	Concept	995	380	553	114
Manual	One-to-One	Concept	5020	0	755	0
	One-to-Many		10425	0	2836	0
Unmapped						
[a]None			50	20771	84	5118
Injury			3323	10730	3323	10726
Carrier Status			23	0	22	103
Complication			906	128	906	103
Finding			368	3	4739	21323
Mapping Evidence						
Database Cross-References			38473	279236	52430	339210
Synonyms			10169	42191	67381	85132
Labels			19343	97920	75795	113565
Similarity			11975	15825	12789	114

[a]The unmapped “None” category for data used in one or more clinical encounters includes concepts that have not yet been mapped. When applied to data not used in a clinical encounter, “None” indicates concepts that were unable to be mapped to an Open Biological and Biomedical Ontology Foundry ontology concept.

Acronyms: HPO (Human Phenotype); Mondo (Mondo Disease Ontology).

Table 2: OMOP Drug Ingredient Concept Mapping Results by OMOP2OBO mapping method.

Ontology			ChEBI		PRO		VO		NCBITaxon	
Used in ≥ 1 Clinical Encounter			**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
OMOP Concepts			1697	2718	245	72	125	36	463	4256
Mapped Ontology Concepts			2555	3307	270	73	131	36	485	4258
Mapping Category										
Automatic	One-to-One	Concept	959	2192	7	42	92	18	20	135
		Ancestor	17	130	5	19	0	0	22	14
	One-to-Many	Concept	235	169	0	1	0	0	0	1
		Ancestor	61	149	3	0	2	4	2	1
Cosine Similarity	One-to-One	Concept	318	78	214	10	24	14	271	4105
Manual	One-to-One	Concept	31	0	9	0	5	0	136	0
	Constructor		76	0	7	0	2	0	12	0
Unmapped										
[a]None			0	7392	1425	10038	1572	10074	1234	5584
Mapping Evidence										
Database Cross-References			957	759	0	0	0	0	0	0
Synonyms			4567	7732	26	94	92	18	59	199
Labels			5578	9676	25	132	282	97	71	391
Similarity			1352	2562	16	54	100	32	160	4241

[a]The unmapped “None” category for data used in one or more clinical encounters includes concepts that have not yet been mapped. When applied to data not used in a clinical encounter, “None” indicates concepts that were unable to be mapped to an Open Biological and Biomedical Ontology Foundry ontology concept.

Acronyms: ChEBI (Chemical Entities of Biological Interest); PRO (Protein Ontology); and VO (Vaccine Ontology); NCBITaxon (National Center for Biotechnology Information Taxon Ontology).

Table 3: OMOP Measurement Concept Mapping Results by OMOP2OBO mapping method.

Ontology			HPO		Uberon		NCBITaxon		PRO		ChEBI		CL	
Used in ≥ 1 Clinical Encounter			Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
OMOP Concepts			1437	2462	1437	2462	910	1788	567	1237	1226	2327	225	152
Test Results			4087	6801	4087	6801	2572	4888	1572	3283	3487	6415	616	429
Mapped Ontology Concepts			4136	6813	5336	8869	2601	6362	1762	3493	3914	8301	707	474
Mapping Category														
Automatic	One-to-One	Concept	17	3	1793	3589	320	444	44	12	263	400	182	186
		Ancestor	23	20	592	593	184	360	15	6	1382	1912	14	0
	One-to-Many	Concept	0	0	10	0	0	0	0	0	0	0	46	24
		Ancestor	0	0	2	0	0	0	0	0	21	0	3	0
Cosine Similarity	One-to-One	Concept	108	5	50	92	44	114	4103	29	102	362	85	20
Manual	One-to-One		3902	6761	406	462	2019	3516	1261	3047	1369	2376	256	178
	One-to-Many		37	12	1234	2065	5	454	149	189	350	1365	30	21
Unmapped														
[a]None			95	16	95	16	1610	1914	2610	3519	695	387	3566	6734
Other			74	14	74	14	74	14	74	14	74	14	74	14
Mapping Evidence														
Database Cross-References			7	0	6	26	0	0	0	0	409	960	261	145
Synonyms			12	4	5232	8308	465	1627	73	24	2824	6348	486	413
Labels			28	24	1637	1242	310	467	35	14	3035	5797	296	226
Cosine Similarity			234	128	699	553	487	844	165	61	1485	2092	296	232

[a]The unmapped “None” category for data used in one or more clinical encounters includes concepts that have not yet been mapped. When applied to data not used in a clinical encounter, “None” indicates concepts that were unable to be mapped to an Open Biological and Biomedical Ontology Foundry ontology concept.

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

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

Term	Definition
TGFBR1	Transforming Growth Factor Beta Receptor 1
TSC1	Tuberous Sclerosis Complex Subunit 1
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.

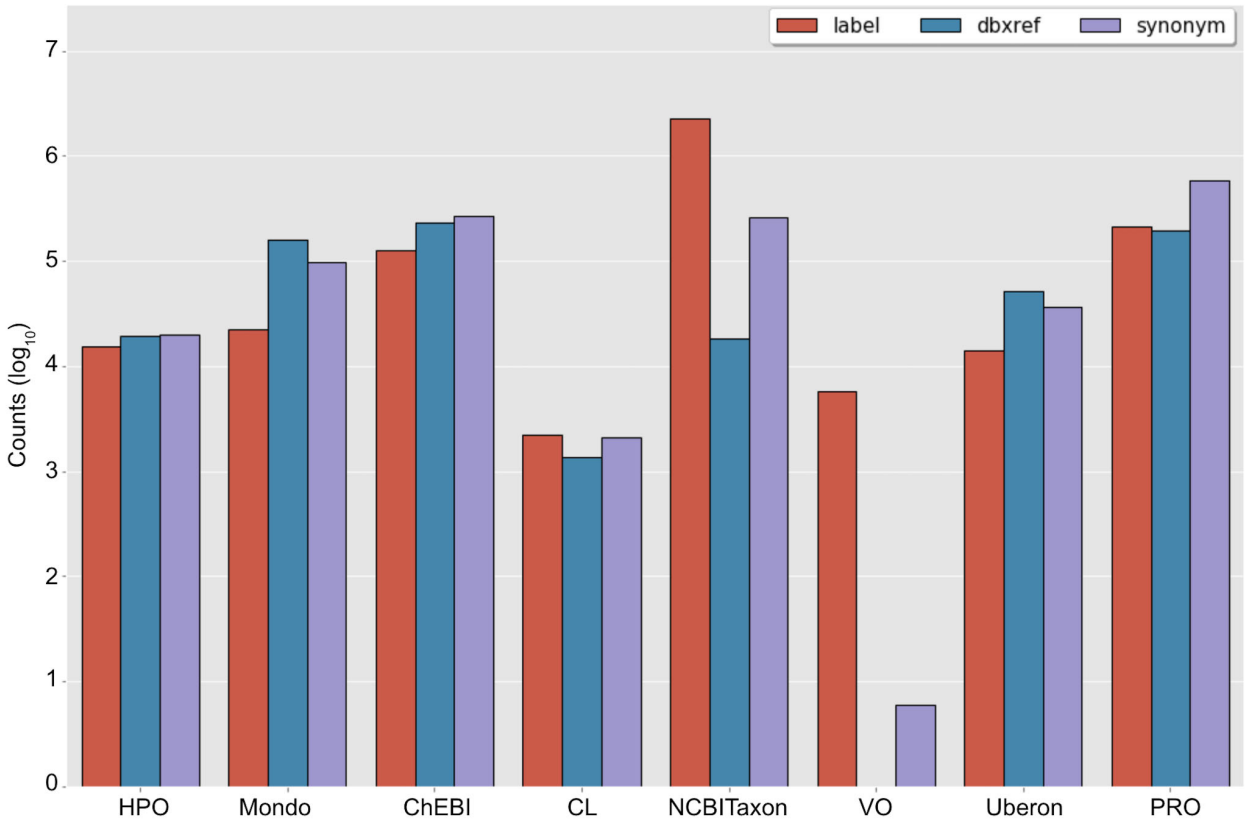
Resource	URL
OMOP2OBO Resources	
PyPI Package	https://pypi.org/project/omop2obo/
GitHub Repository	https://github.com/callahantiff/OMOP2OBO
Project Wiki	https://github.com/callahantiff/OMOP2OBO/wiki
Mapping Dashboard	http://tiffanycallahan.com/OMOP2OBO_Dashboard/
Zenodo Community	https://zenodo.org/communities/omop2obo
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

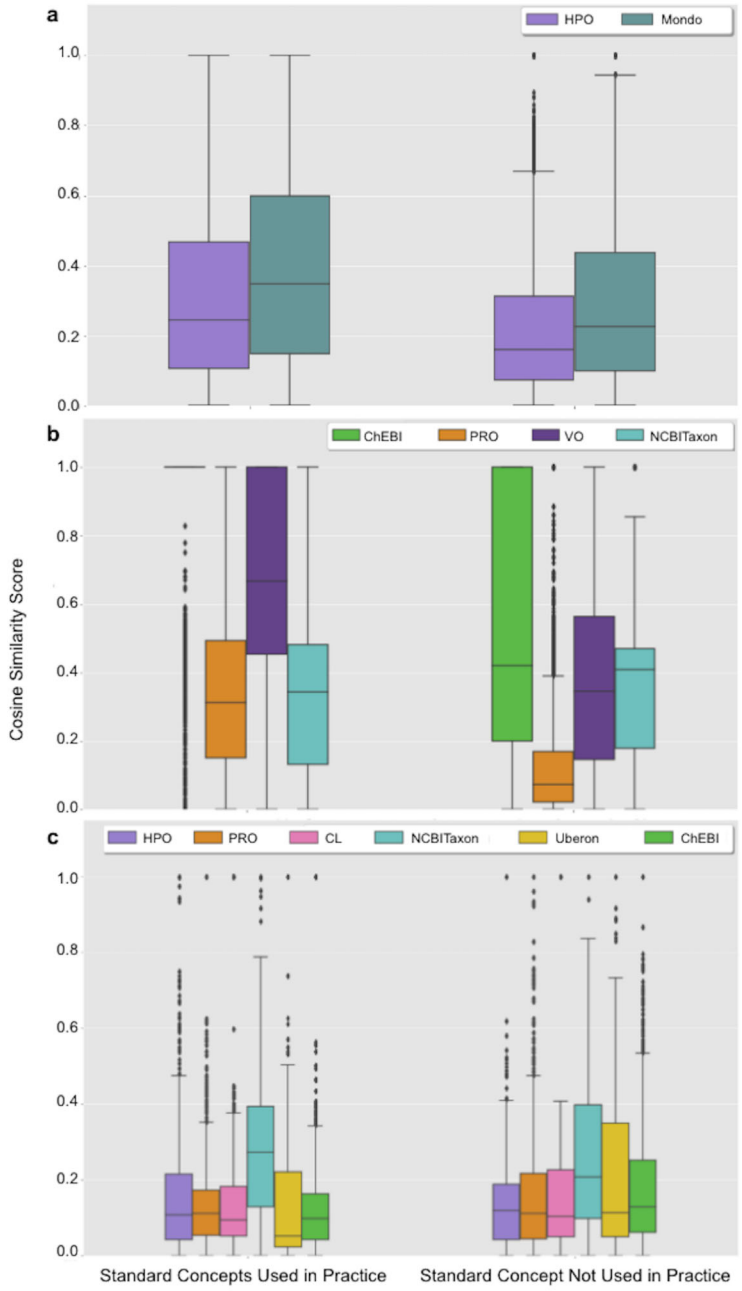
Acronyms: ChEBI (Chemical Entities of Biological Interest); CL (Cell Ontology); HPO (Human Phenotype Ontology); Mondo (Mondo Disease Ontology); NCBITaxon (National Center for Biotechnology Information Taxon Ontology); PRO (Protein Ontology); Uberon (Uber-Anatomy Ontology); VO (Vaccine Ontology).

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.

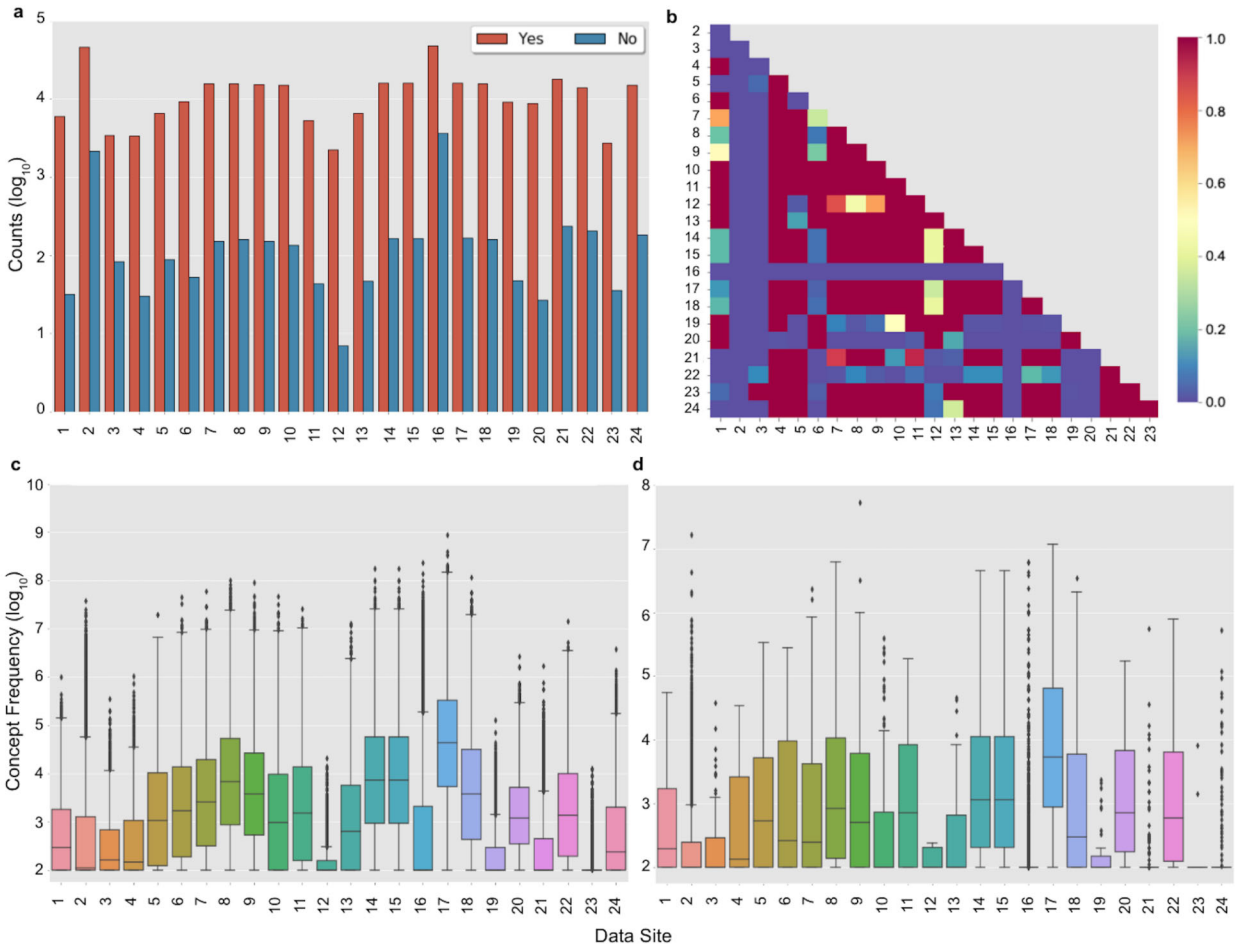
Mapping Category	Definition
Automatic One-to-Many Ancestor	**Definition:** A one-to-many mapping that is automatically generated for a concept's ancestor through exact string mappings to labels or synonyms or exact mappings between codes. For release 1.0, one-to-many mappings indicate that one OMOP concept 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`).
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.

Mapping Category	Definition
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.
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.

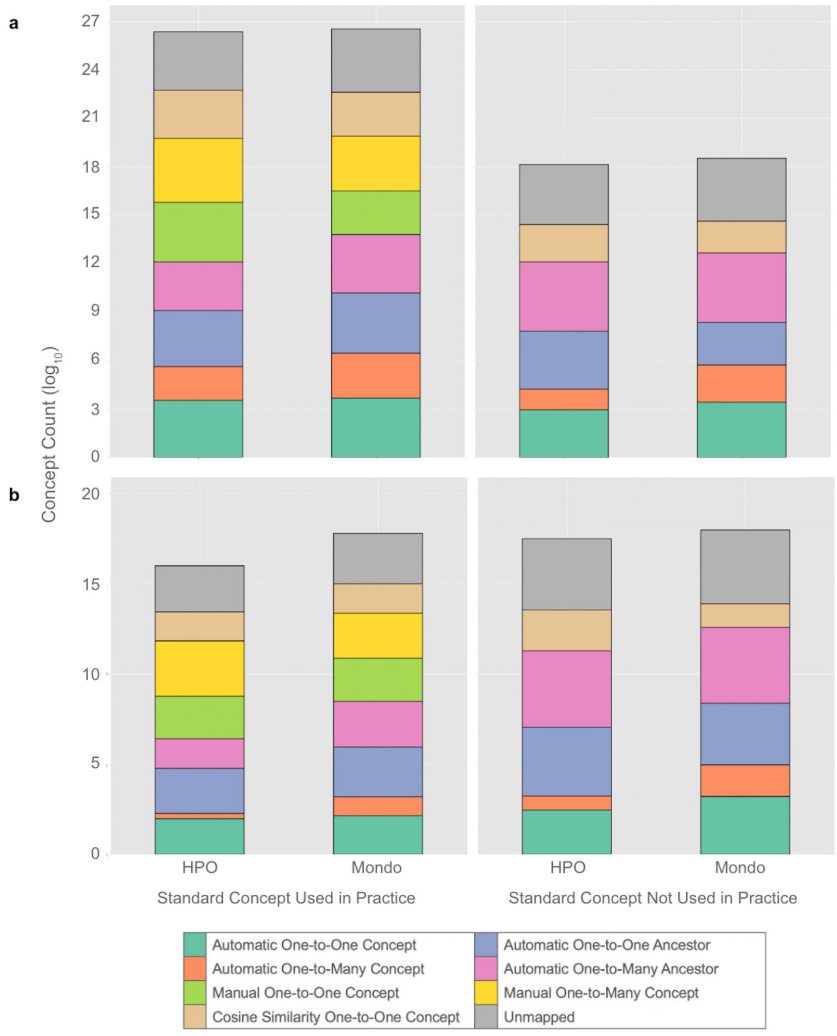
The figure presents the distribution of cosine similarity scores by data wave and OBO Foundry ontology for OMOP (**A**) Conditions, (**B**) Drug Exposure Ingredients, and (**C**) Measurements. Acronyms: OBO (Open Biological and Biomedical Ontology); OMOP (Observational Medical Outcomes Partnership); 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.

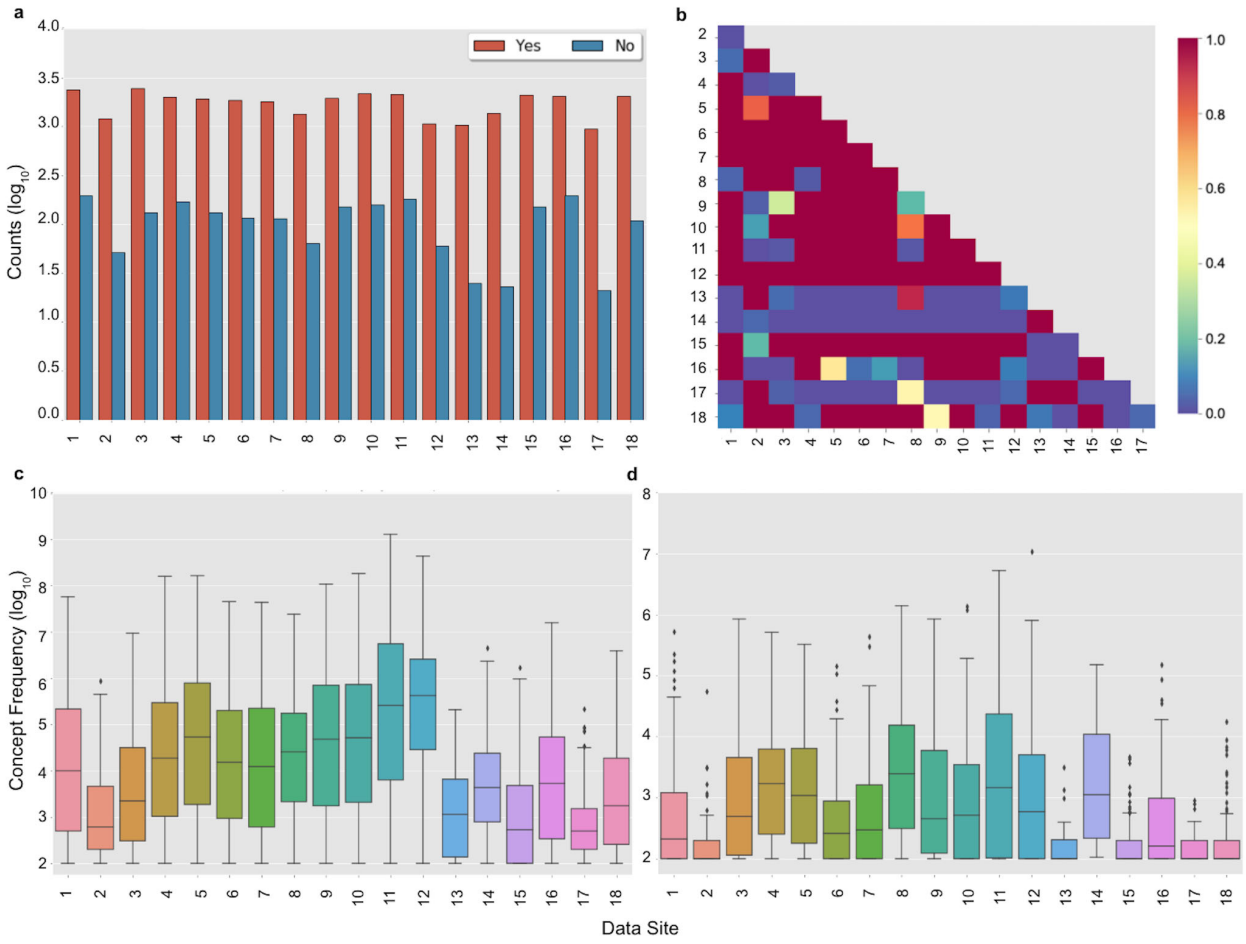
(**A**) This figure visualizes the log count of OMOP2OBO condition concepts covered at each of the Concept Prevalence study sites. (**B**) This figure visualizes the p-values at each Concept Prevalence Study site from the pairwise comparisons of the frequency of concepts from each site that overlapped 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.

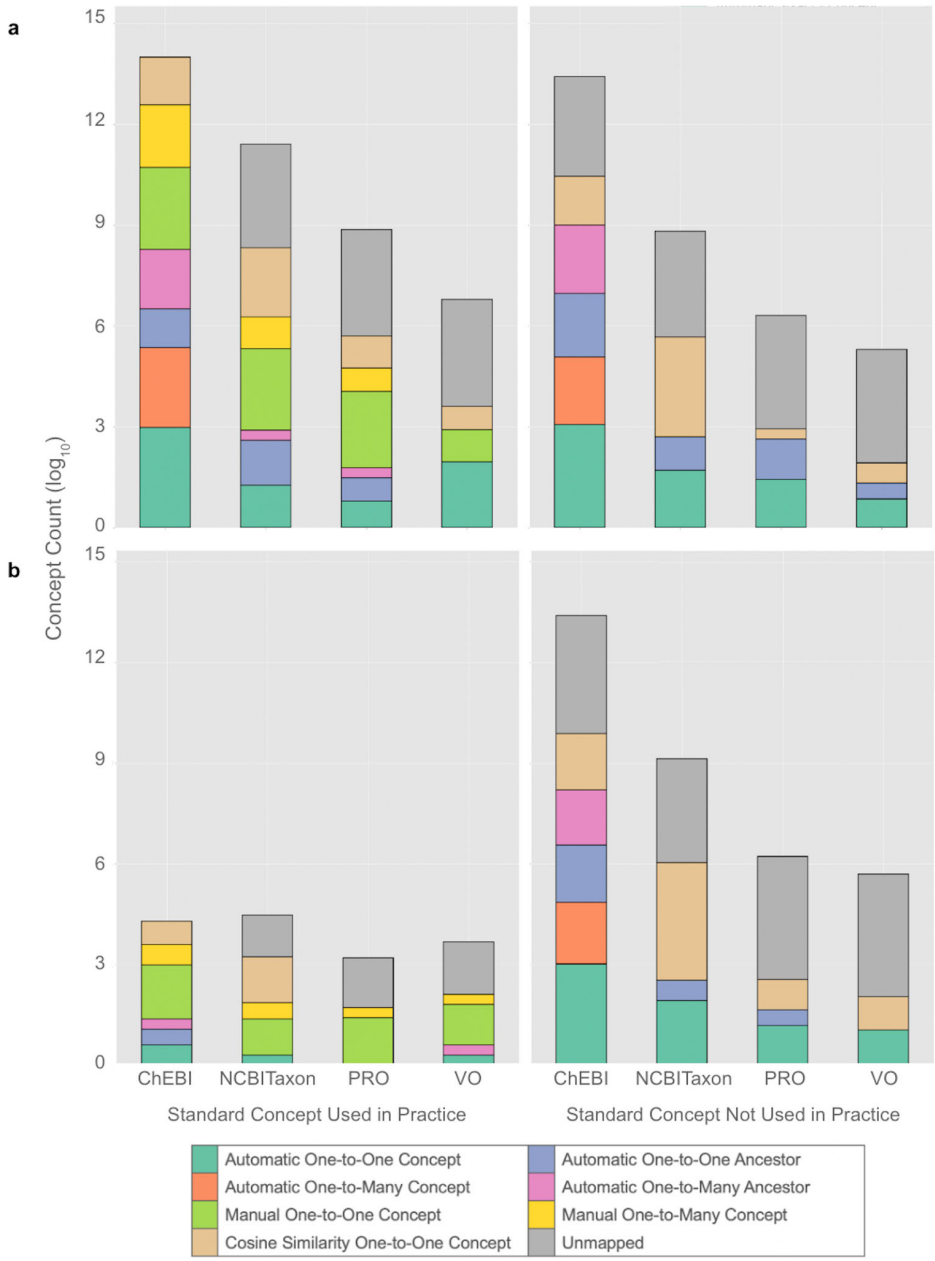
(**A**) This figure visualizes the log count of OMOP2OBO condition concepts that overlapped with concepts in the Concept Prevalence Study by OBO Foundry ontology and data wave. (**B**) This figure visualizes the log count of OMOP2OBO condition concepts that were not found in the Concept Prevalence Study. 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.

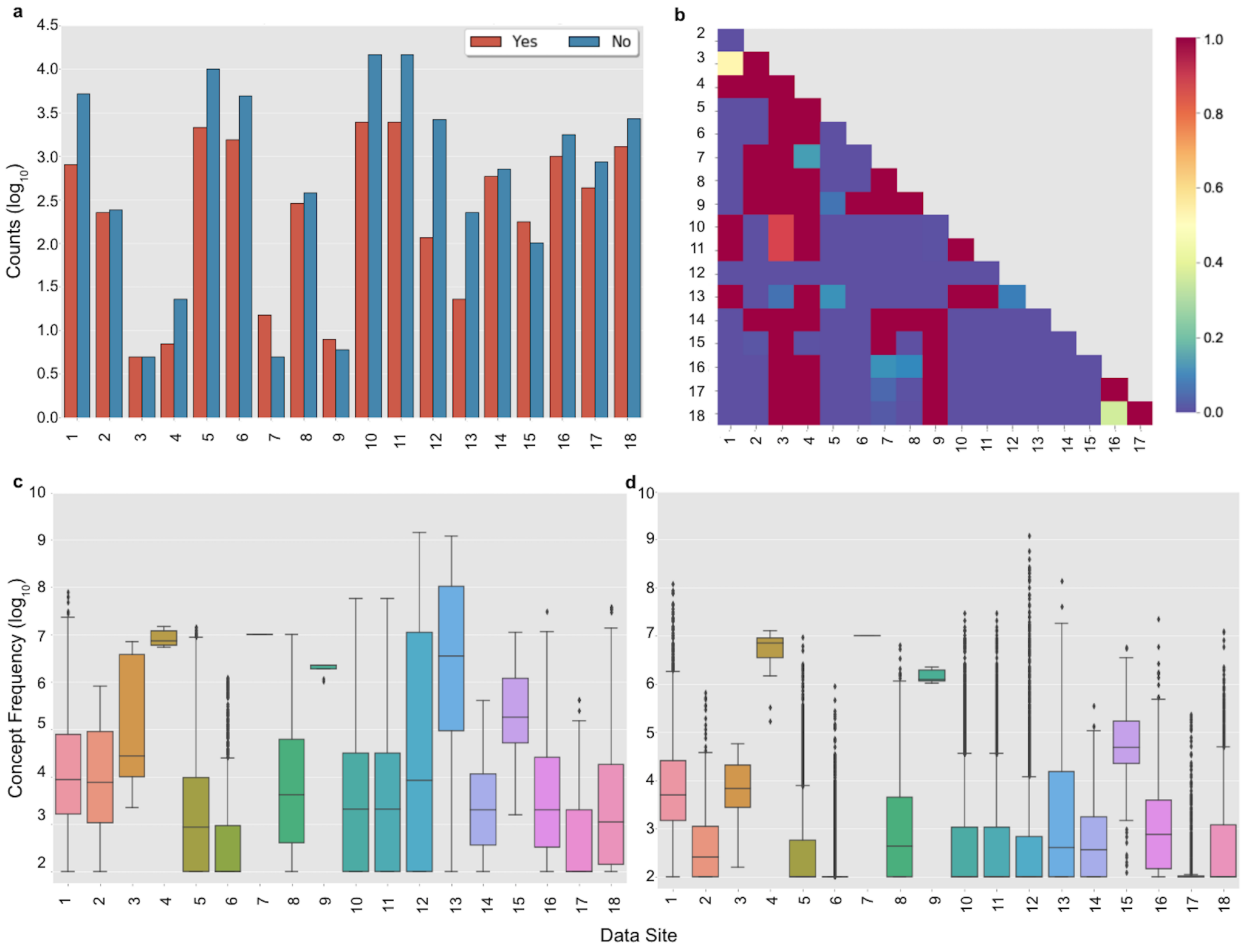
(**A**) This figure visualizes the log count of OMOP2OBO drug ingredient concepts covered at each of the Concept Prevalence study sites. (**B**) This figure visualizes the p-values at each Concept Prevalence Study site from the pairwise comparisons of the frequency of concepts from each site that overlapped 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.

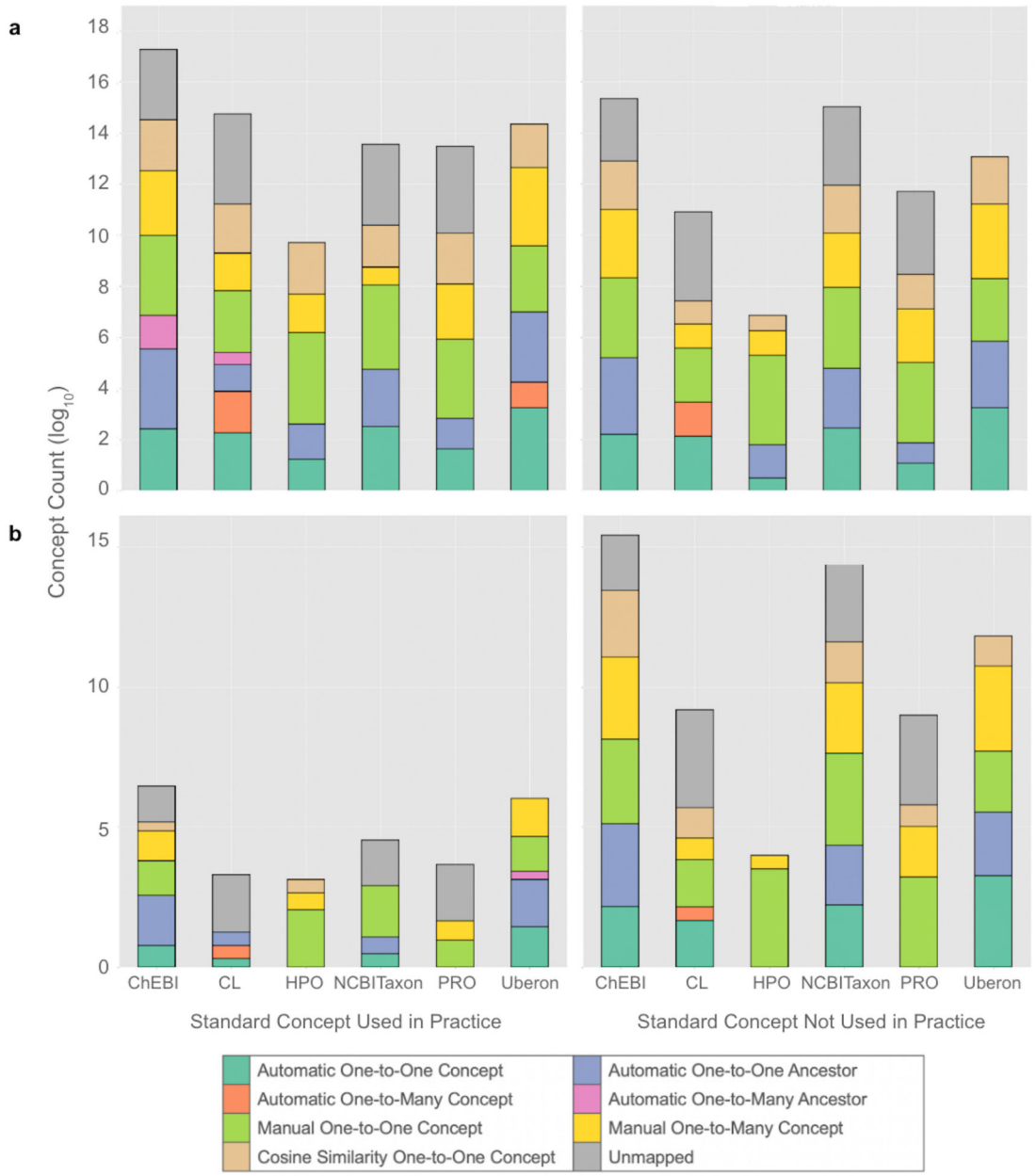
(**A**) This figure visualizes the log count of OMOP2OBO drug ingredient concepts that overlapped with concepts in the Concept Prevalence Study by OBO Foundry ontology and data wave. (**B**) This figure visualizes the log count of OMOP2OBO drug ingredient concepts that were not found in the Concept Prevalence 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).