npj | digital medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

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[7,8], Ben D. Coleman[8], Janine H. Collins[9], Sara J. Deakyne Davies[10], James A. Feinstein[11], Asiyah Y. Lin[4], Blake Martin[12], Nicolas A. Matentzoglu[13], Daniella Meeker[14], Justin Reese[15], Jessica Sinclair[16], Sanya B. Taneja[17], Katy E. Trinkley[18], Nicole A. Vasilevsky[19], Andrew E. Williams[20], Xingmin A. Zhang[8], Joshua C. Denny[4], Patrick B. Ryan[21], George Hripcsak[2], Tellen D. Bennett[12], Melissa A. Haendel[12], Peter N. Robinson[8], Lawrence E. Hunter[1,22] and Michael G. Kahn[22]

Common data models solve many challenges of standardizing electronic health record (EHR) data but are unable to semantically integrate all of the resources needed for deep phenotyping. Open Biological and Biomedical Ontology (OBO) Foundry ontologies provide computable representations of biological knowledge and enable the integration of heterogeneous data. However, mapping EHR data to OBO ontologies requires significant manual curation and domain expertise. We introduce OMOP2OBO, an algorithm for mapping Observational Medical Outcomes Partnership (OMOP) vocabularies to OBO ontologies. Using OMOP2OBO, we produced mappings for 92,367 conditions, 8611 drug ingredients, and 10,673 measurement results, which covered 68–99% of concepts used in clinical practice when examined across 24 hospitals. When used to phenotype rare disease patients, the mappings helped systematically identify undiagnosed patients who might benefit from genetic testing. By aligning OMOP vocabularies to OBO ontologies our algorithm presents new opportunities to advance EHR-based deep phenotyping.

npj Digital Medicine (2023)6:89 ; https://doi.org/10.1038/s41746-023-00830-x

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]. The cornerstone of medicine, diagnosis or clinical phenotyping, aims to identify empirically observable traits exhibited by patients (i.e., signs and symptoms) known to be characteristic of a specific disease[7]. Computational phenotyping is the process of converting clinical phenotypes into computer-executable algorithms in order to identify relevant patients from large sources of clinical data, usually EHRs[8]. One promise of EHR-based computational phenotyping is the ability to perform population-level investigations of mechanistic drivers of disease in diverse patient populations[9,10]. Despite significant progress, this objective remains largely aspirational[6,11–14].

Traditionally, computational phenotypes have been imprecise due to their exclusive reliance on EHR data, which has been shown to be insufficient at capturing the phenotypic heterogeneity present in most complex diseases[15–18]. Deep phenotyping, or "the precise and comprehensive analysis of phenotypic abnormalities in which the individual components of the phenotype are observed and described"[7], is a fundamental component of precision medicine that requires timely synthesis of multiple types of patient data[19,20]. Deep phenotyping has been successfully applied to rare disease and genetic disorders[21–33], cancer[34–40], and pregnancy[41–43] using a variety of clinical and -omic data. Despite large-scale biobanking efforts and resources like the UK Biobank (https://www.ukbiobank.ac.uk) and the *All of Us* (AoU) Research Program (https://www.researchallofus.org), most EHRs do not systematically integrate nor have the infrastructure to integrate patient-level genomic data or other forms of external knowledge (e.g., scientific literature) with clinical data[44–46].

Within an EHR, most data used for research (i.e., structured data) are stored using clinical terminologies or vocabularies. A clinical vocabulary is a standard representation of preferred terms which may or may not be hierarchical or have formally defined relationships and is designed to facilitate meaningful and unambiguous information exchange within the medical domain[47–49]. Hundreds of clinical vocabularies have been

[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, 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]Departments of Biomedical Informatics and Pediatrics, University of Colorado School of Medicine, Aurora, CO 80045, USA. [13]Semanticly, Athens, Greece. [14]Yale School of Medicine, New Haven, CT 06510, USA. [15]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [16]HealthLinc, Valparaiso, IN 46383, USA. [17]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA. [18]Department of Family Medicine, University of Colorado Anschutz School of Medicine, Aurora, CO 80045, USA. [19]Translational and Integrative Sciences Lab, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA. [20]Tufts Institute for Clinical Research and Health Policy Studies, Tufts University, Boston, MA 02155, USA. [21]Janssen Research and Development, Raritan, NJ 08869, USA. [22]Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO 80045, USA. ✉email: tiffany.callahan@cuanschutz.edu

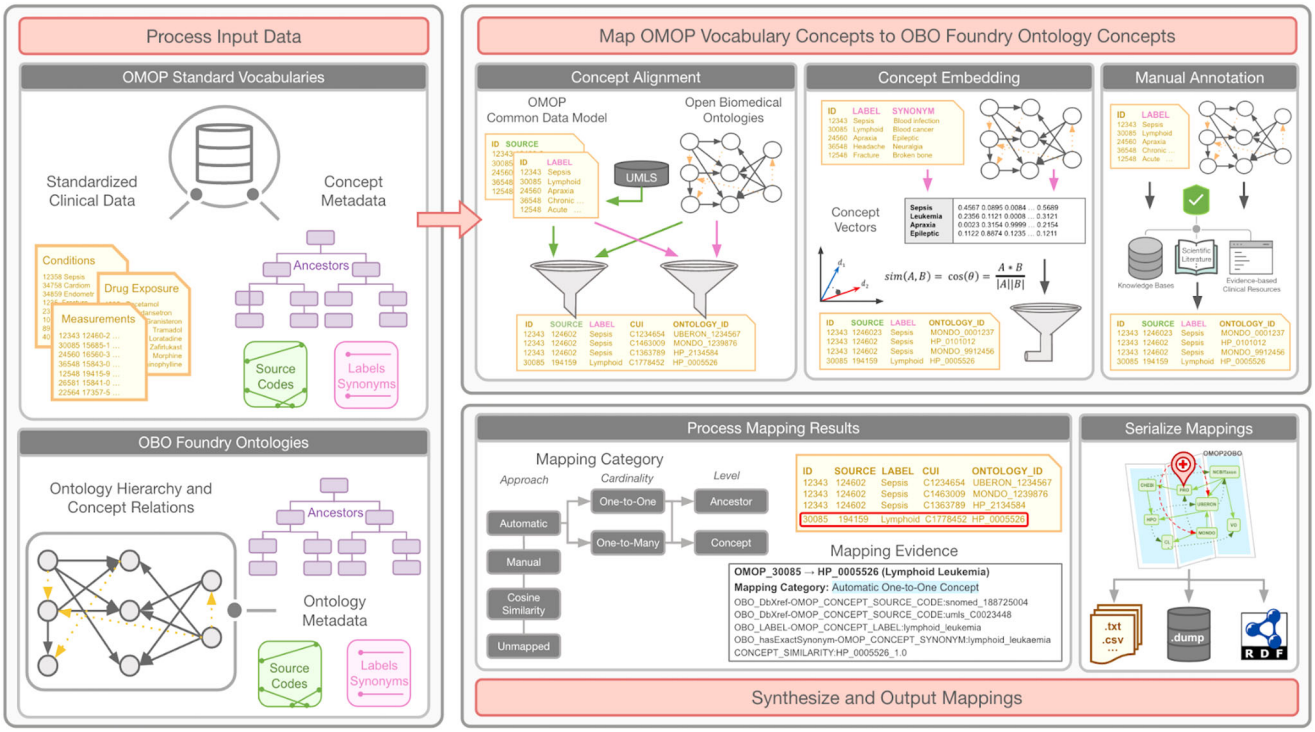

Fig. 1 Overview of the OMOP2OBO algorithm. The OMOP2OBO algorithm consists of three components: (1) Process input data. The algorithm takes as input a table of Observational Medical Outcomes Partnership (OMOP) concepts and a list of one or more OBO (Open Biological and Biomedical Ontology) Foundry ontologies. For both data types, the algorithm expects concept or class identifiers, source codes or database cross-references, labels, synonyms, and ancestor concepts or classes. (2) Map OMOP vocabulary concepts to OBO Foundry Ontology concepts. OMOP concepts are automatically mapped to OBO Foundry ontology concepts. The algorithm includes several different approaches (e.g., concept alignment and concept embedding), prioritizing those that result in high-confidence mappings. (3) Synthesize and output mapping results. The mapping results from the prior component are post-processed to include a mapping category and human-readable evidence. Post-processed mappings are serialized and able output to a variety of file types.

developed and their use differs by hospital and country. Examples include the International Classification of Diseases (ICD)[50], the Logical Observation Identifiers, Names and Codes (LOINC)[51], the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT; https://www.snomed.org), and RxNorm[52]. Most clinical vocabularies were not designed to be integrated or interoperable with other vocabularies, which is one of the long standing barriers preventing the secondary use of EHR data for research[46]. Common data models (CDMs) like the Observational Medical Outcomes Partnership (OMOP)[53] have solved many of the challenges of standardizing, representing, and utilizing clinical EHR data. Unfortunately, most CDMs and associated terminology management systems are not yet able to integrate and interpret genomic data or other sources of external knowledge or publicly available data[54].

Similar to clinical vocabularies, ontologies are classification systems that provide detailed representations of our knowledge of a specific domain[49]. Ontologies, like those in the Open Biological and Biomedical Ontology (OBO) Foundry, 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[55–57]. 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[49]. Mapping clinical vocabularies to ontologies has been recognized as a fundamental requirement for use in deep phenotyping[20,46,49,58]. An example of how aligning these resources improves deep phenotyping was recently demonstrated by Zhang et al. who mapped LOINC to the Human Phenotype Ontology (HPO)[59], which enabled the harmonization of laboratory tests with different clinical codes to common HPO concepts[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 exists. While automated mapping approaches exist, they largely remain unable to correctly capture the complex semantics underlying clinical data and the knowledge encoded by clinical vocabulary concepts[61]. For example, when mapping the concept "Peptic Ulcer without Hemorrhage AND without Perforation but with Obstruction" (SNOMED-CT:54157007) to the HPO, most automated approaches would return a single best mapping, most likely "Peptic Ulcer" (HP:0004398). This HPO concept is much broader in meaning than the clinical concept. A more precise mapping would explicitly capture the presence and absence of all relevant phenotypic features: "Peptic Ulcer" (HP:0004398) and "Gastrointestinal Obstruction" (HP:0004796) and NOT "Gastrointestinal Hemorrhage" (HP:0002239) or "Intestinal Perforation" (HP:0031368). To the best of our knowledge no existing mappings or mapping algorithms are capable of capturing this type of complex semantics.

Building on LOINC2HPO, the goal of this work is to develop OMOP2OBO, an algorithm that enables semantically interoperable mappings between clinical vocabularies in the OMOP CDM to OBO Foundry ontologies (Fig. 1). The resulting mappings will enhance the semantic interoperability of the data represented by the OMOP concepts and have the potential to advance deep EHR-based phenotyping by enabling the identification of relevant patients using existing knowledge of the molecular mechanisms underlying disease rather than billing codes which are prone to error and subject to bias. Using OMOP2OBO, we created healthcare system-scale mappings between clinical vocabularies in the OMOP CDM

Table 1. Clinical data used for input to OMOP2OBO mapping algorithm.

OMOP domain and vocabulary	Data wave[a]	Concept level	Concepts	Labels	Synonyms
Condition SNOMED-CT[b]	Standard Concepts Used in Practice	Concept	29,129	29,129	86,630
		Ancestor	1,421,104	1,389,525	N/A
	Standard Concepts Not Used in Practice	Concept	80,590	80,590	194,264
		Ancestor	3,458,072	3,393,343	N/A
Drug Ingredient RxNorm[c]	Standard Concepts Used in Practice	Concept	1693	1693	1865
		Ancestor	1697	1696	N/A
	Standard Concepts Not Used in Practice	Concept	10,110	10,110	11,235
		Ancestor	10,578	10,578	N/A
Measurement LOINC[d]	Standard Concepts Used in Practice	Concept	1606	1606	41,917
		Ancestor	20,784	21,196	N/A
	Standard Concepts Not Used in Practice	Concept	2477	2477	73,612
		Ancestor	23,457	24,306	N/A

OMOP Observational Medical Outcomes Partnership.
[a]Concepts were mapped in two data waves according to whether or not they had been used at least once in clinical practice (i.e., Concepts Used in Practice) or not (i.e., Concepts Not Used in Practice).
[b]The Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) is the Observational Medical Outcomes Partnership (OMOP) standard vocabulary used for Condition Occurrence concepts.
[c]RxNorm is the OMOP standard vocabulary used for Drug Exposure Ingredient concepts.
[d]The Logical Observation Identifiers, Names and Codes (LOINC) is the OMOP standard vocabulary used for Measurement concepts.

and eight of the most widely used OBO Foundry ontologies[56] spanning diseases, phenotypes, anatomical entities, organisms, chemicals, vaccines, and proteins. The mappings were evaluated on: (1) accuracy, examined by a team of domain 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 through the identification of patients with an undiagnosed rare disease. OMOP2OBO is open source (https://github.com/callahantiff/OMOP2OBO) and includes a custom built interactive dashboard (http://tiffanycallahan.com/OMOP2OBO_Dashboard).

RESULTS

Supplementary Table 1 lists the acronyms and definitions used in the paper. The resources used to build and evaluate the OMOP2OBO algorithm and mappings are described in Supplementary Tables 2 and 3.

OMOP2OBO mapping data: OMOP data

The OMOP2OBO mappings were created using a de-identified pediatric dataset from the Children's Hospital of Colorado (CHCO) normalized to the OMOP CDM (referred throughout the manuscript as "CHCO OMOP Database" and described in detail in Supplementary Table 3)[53,62]. Standardized vocabularies are a fundamental component of the OMOP CDM, which serve as primary vocabularies within each OMOP domain; all other vocabularies within a specific domain are aligned to a standard vocabulary using mappings provided by the CDM[63]. The standard vocabularies used in this work included: SNOMED-CT (the OMOP Condition domain for diseases and clinical findings), RxNorm (the OMOP Drug domain for drug products and vaccines), and LOINC (the OMOP Measurement domain for laboratory tests and assessment scales). Concepts from these three vocabularies, including labels, synonyms, source codes (i.e., standard vocabulary codes), and ancestor concepts obtained from the OMOP CDM, were extracted and used as input to the OMOP2OBO mapping algorithm. Using the CHCO OMOP Database, concepts were organized into two data waves according to whether or not they had been used at least once in clinical practice (i.e., "Concepts Used in Practice") or not (i.e., "Concepts Not Used in Practice"). Only "Concepts Used in Practice" were manually mapped.

The counts of concepts eligible for mapping by OMOP domain and data wave are shown in Table 1. There were 109,719 condition concepts (Concepts Used in Practice: $n = 29{,}129$; Concepts Not Used in Practice: $n = 80{,}590$) and 11,803 drug ingredient concepts (Concepts Used in Practice: $n = 1693$; Concepts Not Used in Practice: $n = 10{,}110$) available to map. For measurements, there were 4083 concepts, representing 11,269 measurement results (Concepts Used in Practice: $n = 1606$ concepts [4425 results]; Concepts Not Used in Practice: $n = 2477$ concepts [6844 results]) available to map. With respect to the Concepts Used in Practice, the 29,129 conditions had a median frequency of 25 (range 1–544,618), the 1693 drug ingredients had a median frequency of 251 (range 1–2,267,866), and the 1606 measurement concepts had a median frequency of 313.5 (range 1–56,823,139).

OMOP2OBO mapping data: OBO Foundry ontologies

Under the guidance of domain experts, eight OBO Foundry ontologies were selected to represent the following domains: diseases (Mondo), phenotypes (HPO), anatomical entities (Uber Anatomy Ontology [Uberon][64]; Cell Ontology [CL][65]), organisms (National Center for Biotechnology Information Taxon Ontology [NCBITaxon][66]), chemicals (Chemical Entities of Biological Interest [ChEBI][67]), vaccines (the Vaccine Ontology [VO][68]), and proteins (the Protein Ontology [PRO][69]). Each set of ontology concepts also included metadata, which was obtained by querying each ontology for labels, definitions, synonyms, and database cross-references (i.e., codes from other vocabularies and ontologies). The amount of metadata available for mapping is shown in Table 2 and varied across the OBO Foundry ontologies, with NCBITaxon containing the most metadata and Uberon containing the least (visualized in Supplementary Fig. 1). A chi-square test of independence with Yate's correction revealed a significant

Table 2. Open Biological and Biomedical Ontology Foundry ontologies used for input to OMOP2OBO mapping algorithm.

Ontology	Classes	Labels	Synonyms	Cross-references
ChEBI	126,169	126,169	269,798	231,247
CL	2238	2238	2124	1376
HPO	15,247	15,247	19,860	19,569
Mondo	22,288	22,288	98,181	159,918
NCBITaxon	2,241,110	2,241,110	263,571	18,426
PRO	215,624	215,624	590,190	195,671
Uberon	13,898	13,898	36,771	51,322
VO	5789	5789	6	0

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, *OMOP* Observational Medical Outcomes Partnership, *PRO* Protein Ontology, *Uberon* Uber-Anatomy Ontology, *VO* Vaccine Ontology.

association between the ontology and the amount of available metadata ($\chi^2(14) = 2{,}664{,}853.8$, $p < 0.0001$). Post hoc tests with Bonferroni adjustment confirmed the ontologies provided significantly different amounts of metadata ($p < 0.0001$ for all significant comparisons).

OMOP2OBO mappings

Figure 2 includes example mappings and illustrates how the OBO Foundry ontologies were used to map concepts from each OMOP domain. As illustrated by this figure, OMOP conditions were mapped to HPO and Mondo, OMOP drug ingredients were mapped to ChEBI, NCBITaxon, PRO, and VO, and OMOP measurements results were mapped to HPO, Uberon, NCBITaxon, PRO, CheBI, and CL. As illustrated in the bottom panel of Fig. 1, each mapping consists of four 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); (3) level (i.e., concept or ancestor); and (4) evidence, which consists of pipe-delimited free-text phrases that explain what fields were used to construct the mapping. Supplementary Table 4 provides additional details on and examples of the OMOP2OBO mapping categories. The mapping procedures and resources are described in the "OMOP2OBO Algorithm" sections of the Methods.

OMOP2OBO mappings: conditions

Unified Medical Language System (UMLS)[70] concept unique identifiers (CUIs) were found for 96.6% of condition concepts ($n = 105{,}976$) representing 69 unique Semantic Types[71]. The mapping results for each OBO Foundry ontology are displayed in Fig. 3 and detailed in Supplementary Table 5. Of the 109,719 available concepts, 66.9% ($n = 73{,}417$) mapped to 5654 unique HPO concepts (Concepts Used in Practice: 83.9%; Concepts Not Used in Practice: 60.8%) and 57.8% ($n = 63{,}374$) mapped to 9637 unique Mondo concepts (Concepts Used in Practice: 68.9%; Concepts Not Used in Practice: 53.8%). Only 50 concepts we attempted to map (excluding purposefully unmapped concepts) were unable to be mapped to at least one OBO Foundry ontology concept.

The frequency distributions of the Concepts Used in Practice by mapping category and ontology are visualized in Fig. 4. The majority of automatic mappings were one-to-one at the concept-level for Concepts Used in Practice (HPO: $n = 3601$; Mondo: $n = 4836$) and Concepts Not Used in Practice (HPO: $n = 1166$; Mondo: $n = 4261$). The majority of the manual mappings were one-to-many (HPO: $n = 10{,}328$; Mondo: $n = 2835$). Cosine similarity-scored concept embeddings enabled 1374 HPO (Concepts Used in Practice: median 0.5, range 0.2–1; Concepts Not Used in Practice: median 0.4, range 0.2–1) and 667 Mondo (Concepts Used in Practice: median 0.8, range 0.2–1; Concepts Not Used in Practice: median 1, range 0.2–1) mappings (Supplementary Fig. 2a). On average, more evidence was found for mappings to Concepts Not Used in Practice than Concepts Used in Practice for HPO (8.9 vs. 3.9) and Mondo (12.4 vs. 10.6).

OMOP2OBO mappings: drug ingredients

UMLS CUIs were found for 99.3% of drug ingredient concepts ($n = 11{,}716$) representing 23 unique Semantic Types. The mapping results for each OBO Foundry ontology are displayed in Fig. 3 and detailed in Supplementary Table 6. Of the 11,803 available concepts, 37.4% ($n = 411$) mapped to 4072 unique ChEBI concepts (Concepts Used in Practice: 100%; Concepts Not Used in Practice: 26.9%), 21.5% ($n = 4661$) mapped to 2535 unique NCBITaxon concepts (Concepts Used in Practice: 23.9%; Concepts Not Used in Practice: 42.1%), 2.1% ($n = 4249$) mapped to 142 unique PRO concepts (Concepts Used in Practice: 10.5%; Concepts Not Used in Practice: 0.7%), and 1.3% ($n = 154$) mapped to 132 unique VO concepts (Concepts Use in Clinical Practice: 6.9%; Concepts Not Used in Practice: 0.4%). All of the OMOP concepts were mapped to at least one ChEBI concept.

The frequency distributions of the Concepts Used in Practice by mapping category and OBO Foundry ontology are visualized in Fig. 5. The majority of automated mappings were one-to-one at the concept-level for Concepts Used in Practice (ChEBI: $n = 959$; NCBITaxon: $n = 20$; PRO: $n = 1$; VO: $n = 90$) and Concepts Not Used in Practice (ChEBI: $n = 2192$; NCBITaxon: $n = 135$; PRO: $n = 42$; VO: $n = 18$). The majority of the manual mappings were one-to-one (ChEBI: $n = 321$; NCBITaxon: $n = 230$; PRO: $n = 157$; VO: $n = 21$). Cosine similarity-scored concept embeddings enabled 109 ChEBI (Concepts Used in Practice: median 1, range 0.3–1; Concepts Not Used in Practice: median 1, range 0.3–1), 4241 NCBITaxon (Concepts Used in Practice: median 0.6, range 0.3–1; Concepts Not Used in Practice: median 0.6, range 0.3–1), 18 PRO (Concepts Used in Practice: median 0.8, range 0.4–1; Concepts Not Used in Practice: median 1, range 0.6–1), and 17 VO (Concepts Used in Practice: median 1, range 0.4–1; Concepts Not Used in Practice: median 0.8, range 0.4–1) mappings (Supplementary Fig. 2b). On average, more evidence was found for mappings to Concepts Not Used in Practice than Concepts Used in Practice for ChEBI and PRO, excluding NCBITaxon and VO (ChEBI: 7.6 vs. 7.6; PRO: 3.9 vs. 1; NCBITaxon: 1.2 vs. 1.1; VO: 3 vs. 4.1).

OMOP2OBO mappings: measurements

UMLS CUIs were found for 94.8% of measurement concepts ($n = 3869$) representing a single Semantic Type. The mapping results for each OBO Foundry ontology are displayed in Fig. 3 and detailed in Supplementary Table 7. Of the 11,269 measurement results, 96.6% ($n = 10{,}888$) mapped to 1115 unique HPO concepts (Concepts Used in Practice: 92.4%; Concepts Not Used in Practice: 99.4%) and 45 unique Uberon concepts (Concepts Used in Practice: 92.4%; Concepts Not Used in Practice: 99.4%), 76.8% ($n = 8657$) mapped to 425 unique NCBITaxon concepts (Concepts Used in Practice: 64.4%; Concepts Not Used in Practice: 84.9%), 42.6% ($n = 4804$) mapped to 172 unique PRO concepts (Concepts Used in Practice: 35.5%; Concepts Not Used in Practice: 47.2%), 87.9% ($n = 9904$) mapped to 443 unique ChEBI concepts (Concepts Used in Practice: 78.9%; Concepts Not Used in Practice: 93.7%), and 9.9% ($n = 1114$) mapped to 38 unique CL concepts (Concepts Used in Practice: 15.3%; Concepts Not Used in Practice: 6.4%). Only 13 concepts we attempted to map (excluding purposefully unmapped concepts) were unable to be mapped to at least one OBO Foundry ontology concept.

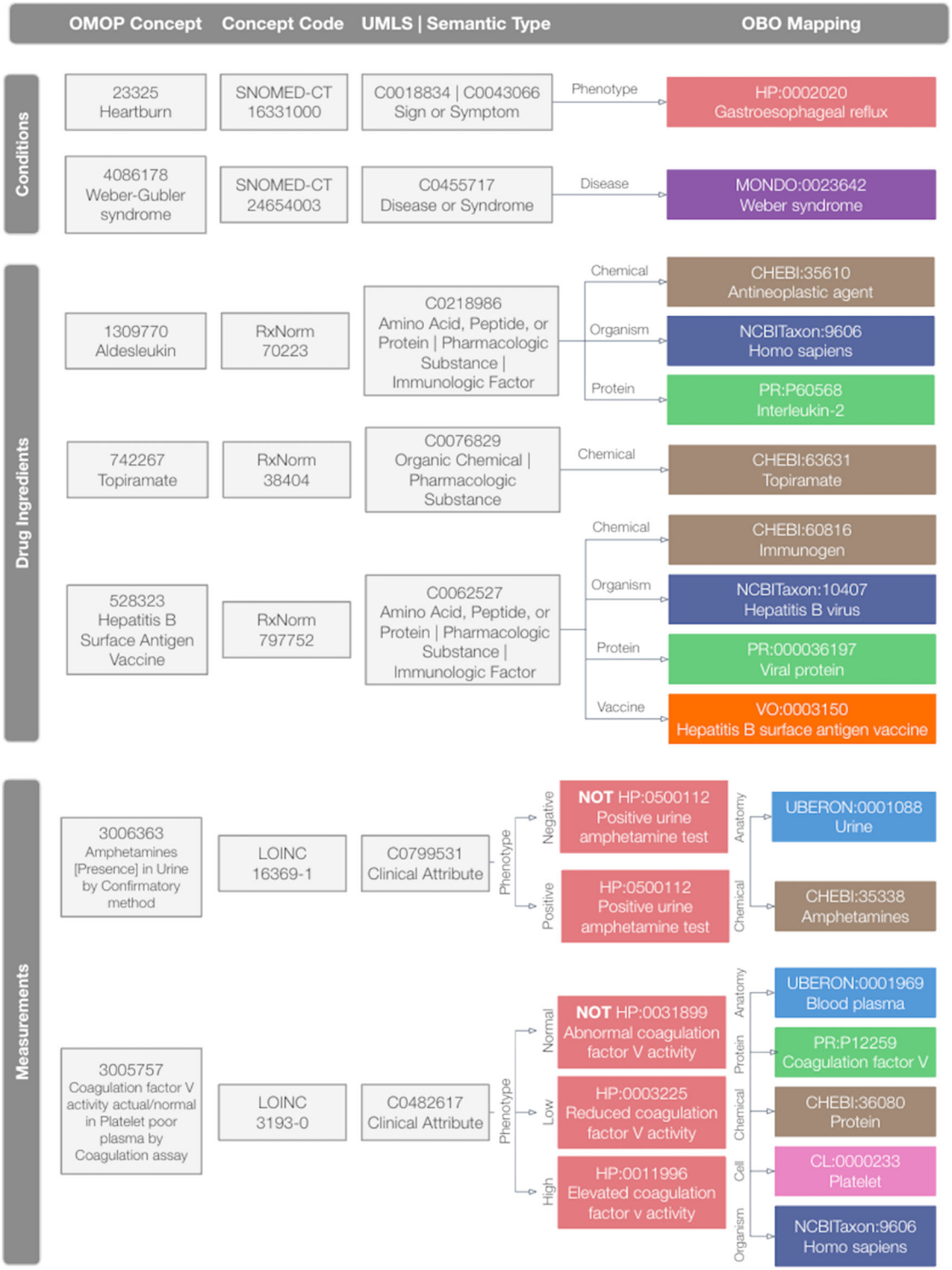

Fig. 2 OMOP2OBO mapping examples by OMOP domain. This figure illustrates which OBO (Open Biological and Biomedical Ontology) Foundry ontologies were used for each OMOP (Observational Medical Outcomes Partnership) domain and provides example mappings. OMOP conditions were mapped to HPO and Mondo. OMOP drug ingredients were mapped to ChEBI, NCBITaxon, PRO, and VO. OMOP measurements were mapped to ChEBI, CL, HPO, NCBITaxon, PRO, and Uberon. UMLS Unified Medical Language System, 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.

The frequency distributions of the Concepts Used in Practice by mapping category and OBO Foundry ontology are visualized in Fig. 6. The majority of the automated mappings were one-to-one at the concept-level for Concepts Used in Practice (HPO: $n = 17$; Uberon: $n = 1793$; NCBITaxon: $n = 444$; PRO: $n = 44$; ChEBI: $n = 264$; CL: $n = 182$) and Concepts Not Used in Practice (HPO: $n = 3$; Uberon: $n = 3589$; NCBITaxon: $n = 444$; PRO: $n = 12$; ChEBI: $n = 515$; CL: $n = 186$). The majority of the manual mappings were

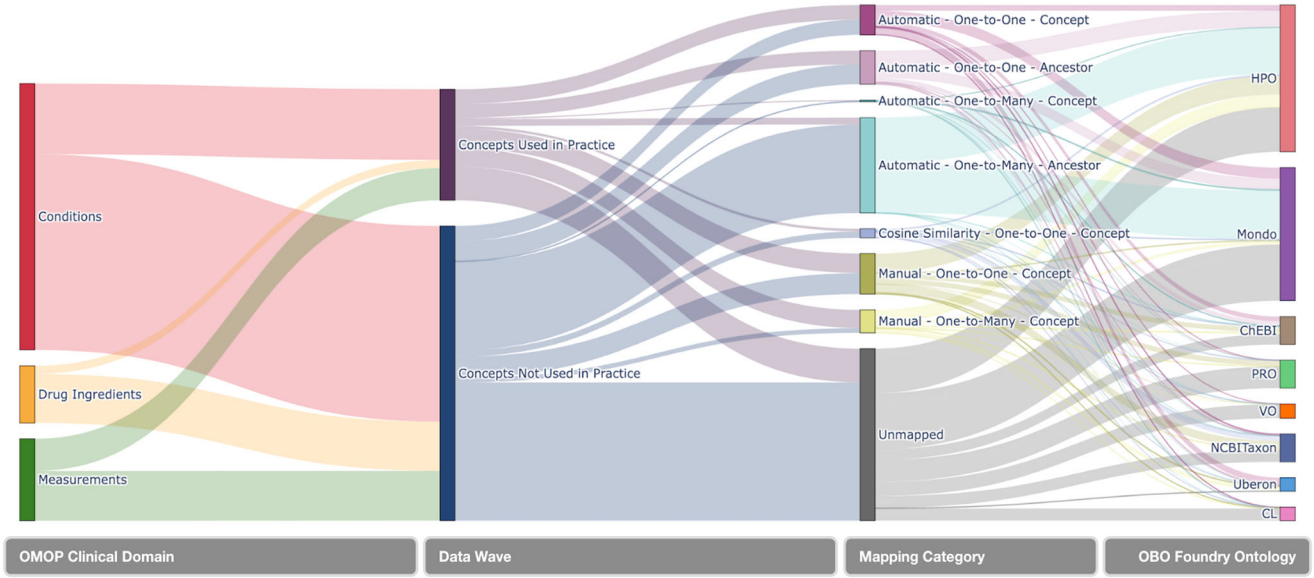

Fig. 3 OMOP concept mapping results by domain, concept type, mapping category, and ontology. This figure features a Sankey Diagram illustrating the mapping flow implemented by the OMOP2OBO algorithm beginning with OMOP (Observational Medical Outcomes Partnership) concepts from the Conditions, Drugs, and Measurements domains, which were grouped by Data Wave (i.e., whether or not the concept has been used at least once in clinical practice), and organized by mapping category. The flow lines in the diagram are weighted by the count of OMOP concepts from the Children's Hospital Colorado pediatric OMOP database. OBO Open Biological and Biomedical Ontology, HPO Human Phenotype Ontology, Mondo Monarch Disease Ontology, ChEBI Chemical Entities of Biological Interest, NCBITaxon National Center for Biotechnology Information Taxon Ontology, PRO Protein Ontology, VO Vaccine Ontology, Uberon Uber-Anatomy Ontology, CL Cell Ontology.

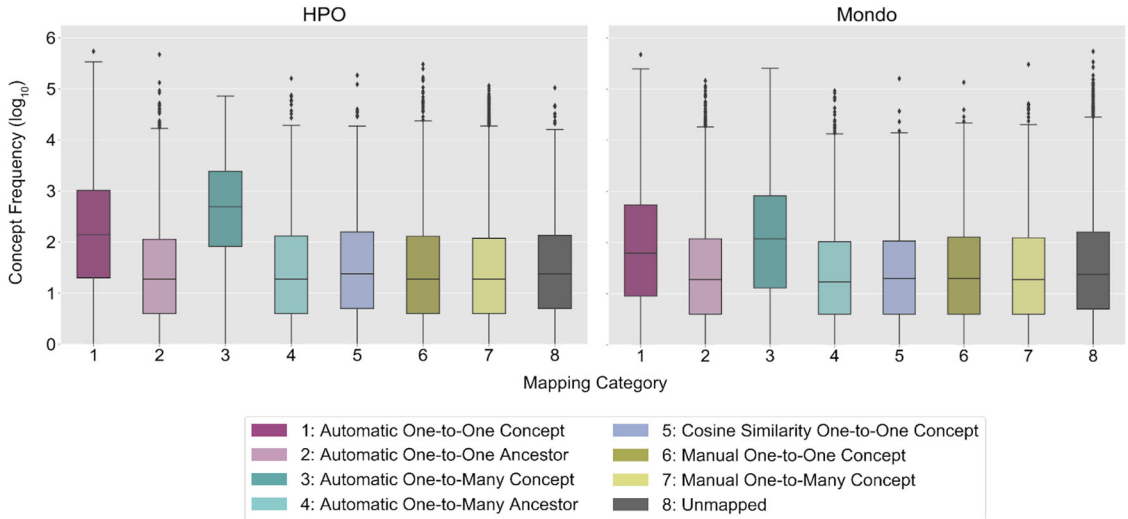

Fig. 4 Condition concept frequency of use in clinical practice by mapping category and ontology. This figure presents the frequency distributions of OMOP (Observational Medical Outcomes Partnership) condition concepts used at least once in clinical practice (log 10 scale) in the Children's Hospital Colorado pediatric OMOP database by mapping category and OBO (Open Biological and Biomedical Ontology) Foundry ontology. In each boxplot, the box extends from the first to third quartile of the data with a center line used to indicate the median. Whiskers extend from each box by 1.5x the interquartile range and outliers that extend past the whiskers shown as dots. The *x*-axis labels are numbers which correspond to the OMOP2OBO mapping categories: (1) Automatic One-to-One Concept; (2) Automatic One-to-One Ancestor; (3) Automatic One-to-Many Concept; (4) Automatic One-to-Many Ancestor; (5) Cosine Similarity One-to-One Concept; (6) Manual One-to-One Concept; (7) Manual One-to-Many Concept; and (8) Unmapped. HPO Human Phenotype Ontology, Mondo Monarch Disease Ontology.

one-to-one (HPO: $n = 3902$; Uberon: $n = 406$; NCBITaxon: $n = 2300$; PRO: $n = 1267$; ChEBI: $n = 1377$; CL: $n = 319$). Cosine similarity-scored concept embeddings enabled 113 HPO (Concepts Used in Practice: median 0.4, range 0.3–0.8; Concepts Not Used in Practice: median 0.4, range 0.3–0.9), 142 Uberon (Concepts Used in Practice: median 0.3, range 0.3–0.8; Concepts Not Used in Practice: median 0.4, range 0.3–0.7), 150 NCBITaxon (Concepts Used in Practice: median 0.4, range 0.3–0.7; Concepts Not Used in Practice: median 0.4, range 0.3–0.7), 132 PRO (Concepts Used in Practice: median 0.4, range 0.3–0.7; Concepts Not Used in Practice: median 0.4, range 0.3–0.6), 476 ChEBI (Concepts Used in Practice: median 0.4, range 0.3–1; Concepts Not Used in Practice: median 0.3, range 0.3–0.6), and 102 CL (Concepts Used in Practice: median 0.4, range 0.3–1; Concepts Not Used in Practice: median 0.4, range 0.3–1) mappings (Supplementary Fig. 2c). On average, more evidence was found for mappings to

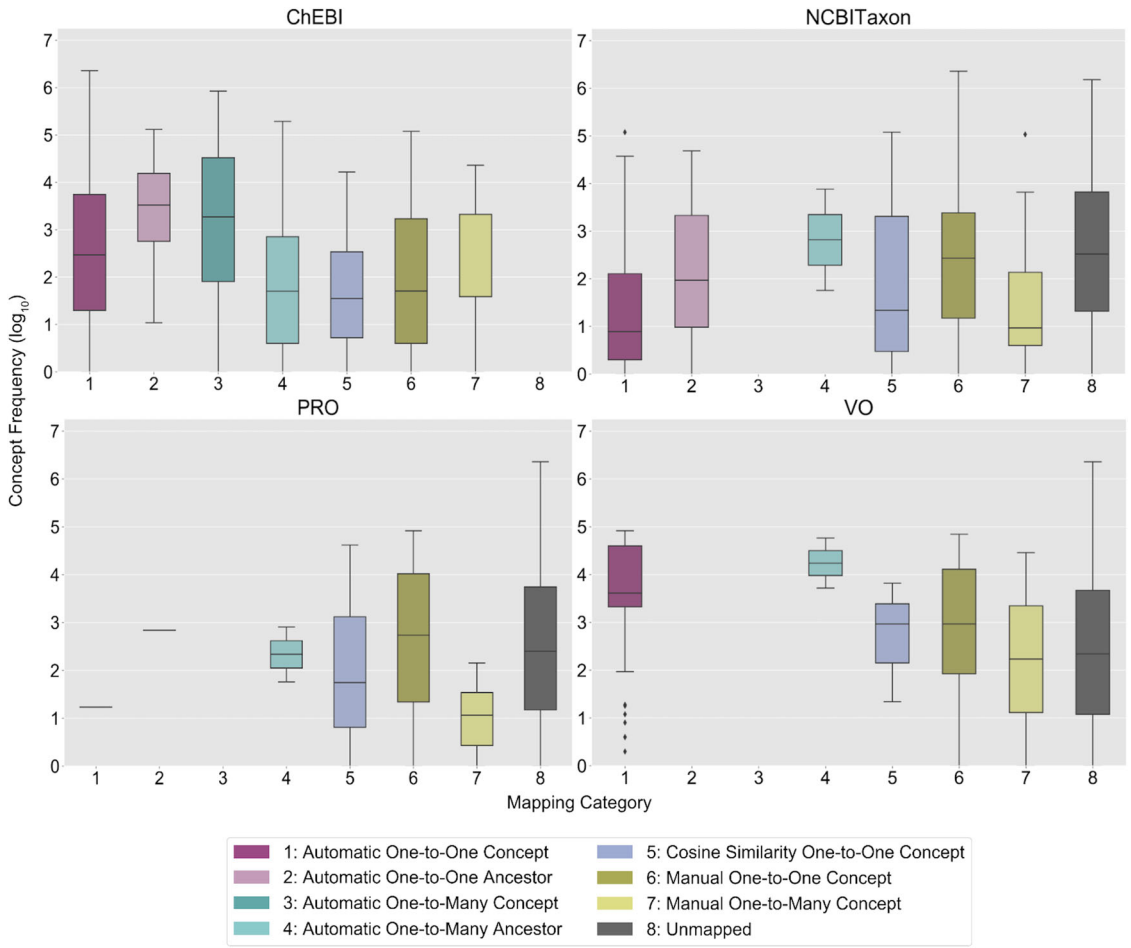

Fig. 5 Drug ingredient concept frequency of use in clinical practice by mapping category and ontology. This figure presents the frequency distributions of OMOP (Observational Medical Outcomes Partnership) drug exposure ingredient concepts used at least once in clinical practice (log 10 scale) in the Children's Hospital Colorado pediatric OMOP database by mapping category and OBO (Open Biological and Biomedical Ontology) Foundry ontology. In each boxplot, the box extends from the first to third quartile of the data with a center line used to indicate the median. Whiskers extend from each box by 1.5x the interquartile range and outliers that extend past the whiskers shown as dots. The *x*-axis labels are numbers which correspond to the OMOP2OBO mapping categories: (1) Automatic One-to-One Concept; (2) Automatic One-to-One Ancestor (3) Automatic One-to-Many Concept; (4) Automatic One-to-Many Ancestor; (5) Cosine Similarity One-to-One Concept; (6) Manual One-to-One Concept; (7) Manual One-to-Many Concept; and (8) Unmapped. ChEBI Chemical Entities of Biological Interest, NCBITaxon National Center for Biotechnology Information Taxon Ontology, PRO Protein Ontology, VO Vaccine Ontology.

Concepts Used in Practice than Concepts Not Used in Practice for HPO, Uberon, and PRO (HPO: 1.03 vs. 1.02; Uberon: 2.3 vs. 1.9; PRO: 1.1 vs. 1; NCBITaxon: 1.3 vs. 1.4; ChEBI: 2.7 vs. 2.9; CL: 2.5 vs. 2.8).

Validation: accuracy

The goal of this task was to verify the accuracy of randomly selected sets of manual one-to-one and one-to-many OMOP2OBO mappings from each OMOP domain through domain expert review. Of the 2000 condition mappings, 73.9% were correct ($n = 1477$). Of the 116 reviewed drug ingredient mappings, 70.7% ($n = 82$) were correct. Upon review, it was found that 165 (31.6%) of the incorrect condition and 14 (41.2%) of the incorrect drug ingredient mappings could be improved by creating more specific mappings through adding new concepts to the OBO Foundry ontologies or by replacing multiple mappings to broad ancestor concepts with a single best representative 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 were found to be correct. Of the 1350 measurement results, 97.3% ($n = 1314$) were correct.

In addition to expert review, each mapping was inspected at least twice by a member of the research team (T.J.C.). If we assume that the automatic one-to-one mappings created using resources provided by the UMLS, OMOP CDM, and OBO Foundry ontologies are correct and exclude mappings that occur at the ancestor level (assuming those are too broad) and unmapped concepts, then the following concepts received at least one form of review: (1) Conditions: 18.4% of Mondo and 9.9% of HPO; (2) Drug Ingredients: 95.3% of NCBITaxon, 90.3% of VO, 85.3% of ChEBI, and 33.3% of PRO; and (3) Measurements: 79.2% of HPO, 50.8% of Uberon, 48.5% of CL, 12.7% of ChEBI, 10.6% of NCBITaxon, and 3.9% of PRO.

Validation: generalization

The goal of this evaluation was to characterize the generalizability or coverage of concepts in the OMOP2OBO mapping set to a set of OMOP standard concepts that are commonly utilized in clinical practice. The Observational Health Data Sciences and Informatics (OHDSI) Concept Prevalence Study contains OMOP standard concepts that are commonly utilized in practice from several independent study sites across the OHDSI network (see

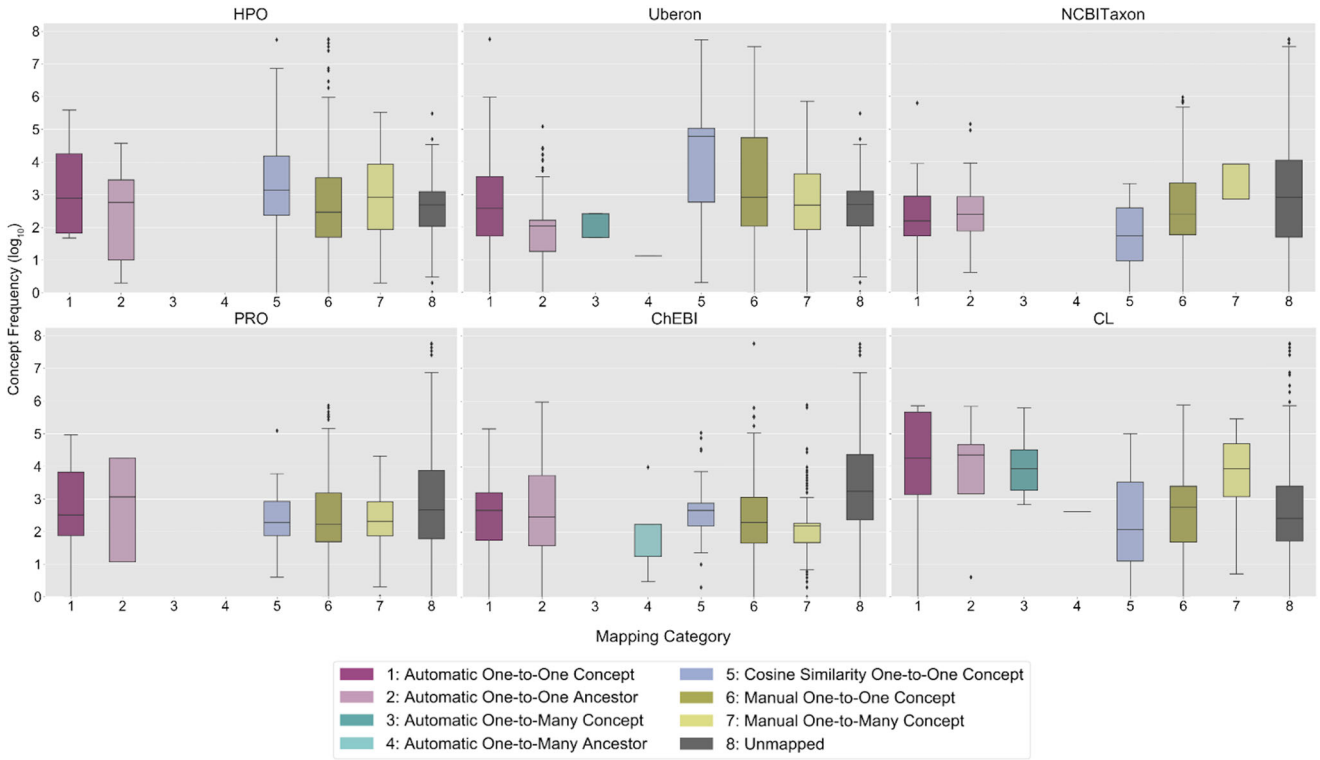

Fig. 6 Measurement concept frequency of use in clinical practice by mapping category and ontology. This figure presents the frequency distributions of OMOP (Observational Medical Outcomes Partnership) measurement concepts used at least once in clinical practice (log 10 scale) in the Children's Hospital Colorado pediatric OMOP database by mapping category and OBO (Open Biological and Biomedical Ontology) Foundry ontology. In each boxplot, the box extends from the first to third quartile of the data with a center line used to indicate the median. Whiskers extend from each box by 1.5x the interquartile range and outliers that extend past the whiskers shown as dots. The *x*-axis labels are numbers which correspond to the OMOP2OBO mapping categories: (1) Automatic One-to-One Concept; (2) Automatic One-to-One Ancestor; (3) Automatic One-to-Many Concept; (4) Automatic One-to-Many Ancestor; (5) Cosine Similarity One-to-One Concept; (6) Manual One-to-One Concept; (7) Manual One-to-Many Concept; and (8) Unmapped. 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.

Supplementary Table 3 for more information)[72–75]. For this evaluation, we leveraged data (referred throughout the remainder of the manuscript as the "OHDSI Concept Prevalence Data") from 24 independent study sites, which included hospitals, academic medical centers, and claims databases. For this analysis, the OMOP2OBO mappings were filtered to identify all concepts with at least one valid mapping (i.e., excluding unmapped and not yet mapped concepts) across all of the ontologies mapped within each OMOP domain.

The OHDSI Concept Prevalence Data contained 62,335 condition concepts from 24 independent sites. The filtered OMOP2OBO mapping set contained 92,367 eligible condition concepts, which covered 92.5% (99.5% weighted coverage) of the OHDSI Concept Prevalence Data condition concepts ($n = 57{,}663$ concepts; median 689, range 100–874,824,195). Of the remaining condition concepts, 34,704 were only found in OMOP2OBO (median 100, range 100–39,975) and 4672 were only found in the OHDSI Concept Prevalence Data (median 100, range 100–52,739,431). These findings are visualized in Fig. 7. OMOP2OBO concept coverage ranged from 93–99.7% across the 24 OHDSI Concept Prevalence Data sites. Supplementary Fig. 3a presents the counts of OMOP condition concepts in the OHDSI Concept Prevalence Data by site. A chi-square test of independence with Yate's correction revealed a significant association between the OHDSI Concept Prevalence Data sites and OMOP2OBO coverage ($\chi^2(23) = 7559.1$, $p < 0.0001$). Post hoc tests using Bonferroni adjustment confirmed that 32% of the pairwise OHDSI Concept Prevalence Data site comparisons had significantly different OMOP2OBO coverage ($p < 0.001$ for all significant comparisons). The results of this analysis are visualized as a heatmap in Supplementary Fig. 3b. The OMOP2OBO concept count by OBO Foundry ontology, data wave, and coverage type are shown in Supplementary Fig. 4.

Results for the 4672 (7.5%) OHDSI Concept Prevalence Data condition concepts missing from OMOP2OBO are visualized in Fig. 8. Roughly 7.9% ($n = 367$) of condition 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 (range 100–3,539,698.5). In total, 90.6% ($n = 4231$) of condition concepts purposefully excluded from the OMOP2OBO mapping set (i.e., no clear pathological or biological origin, not yet mapped, or were unable to be mapped) occurred in an average of 1.7 sites with a mean frequency of 6139.3 (range 100–8,254,186.5). The remaining condition concepts (1.6%; $n = 74$) were truly missing and occurred in an average of 2.7 sites with a mean frequency of 5320.1 (range 100–100,483). The frequency of distributions of the covered condition concepts from OMOP2OBO and Concept Prevalence condition concepts missing from OMOP2OBO in the OHDSI Concept Prevalence Data by site are visualized in Supplementary Fig. 3c, d. The five most frequently occurring missing condition concepts are shown in Table 3. Domain expert review determined these condition concepts were likely missing due to differences in patient populations and coding practices. The domain experts identified comparable condition concepts in the OMOP2OBO mapping set.

The OHDSI Concept Prevalence Data contained 4588 drug ingredient concepts from 18 independent sites. The OMOP2OBO

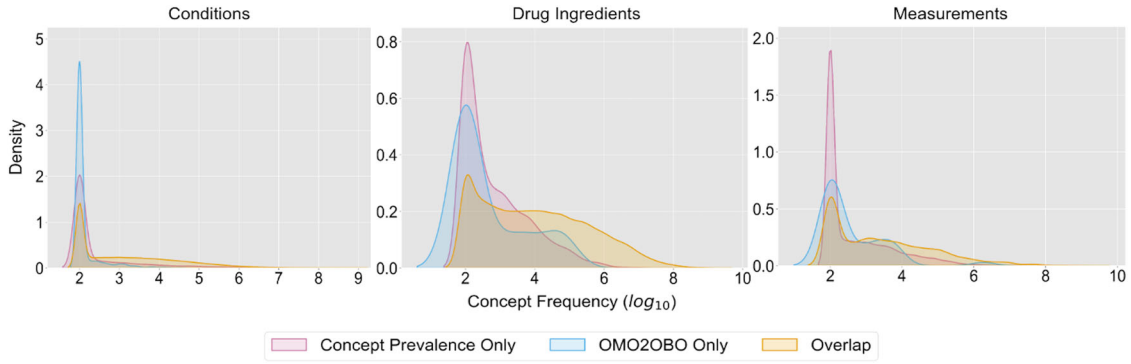

Fig. 7 OMOP2OBO—concept prevalence coverage. This figure visualizes the coverage distributions of Observational Medical Outcomes Partnership (OMOP) concepts over their frequency of use in clinical practice (log 10 scale) within the Concept Prevalence Study data by domain (Conditions [left]; Drugs [middle]; and Measurements [right]). The three modeled distributions include: concepts only found in the Concept Prevalence Study data (magenta), concepts only found in the OMOP2OBO mapping set (blue), and concepts found in both the Concept Prevalence Study data and the OMOP2OBO mapping set (yellow).

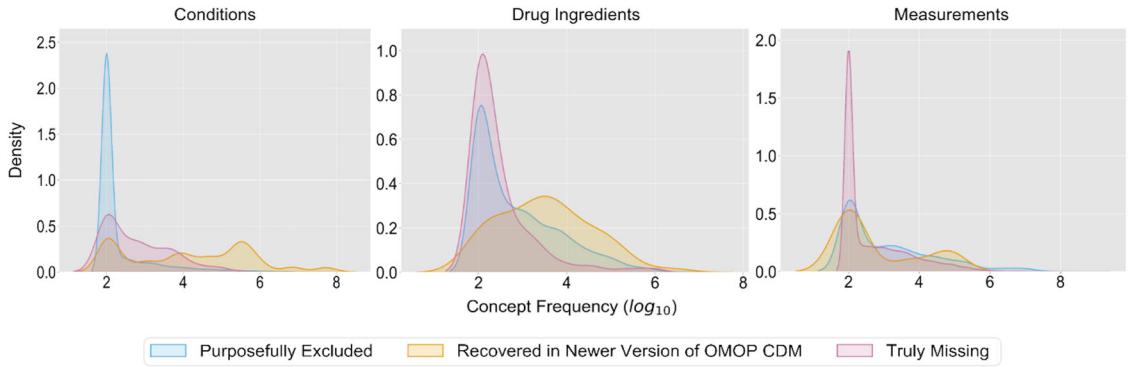

Fig. 8 OMOP2OBO—concept prevalence coverage error analysis. This figure visualizes the distributions of Observational Medical Outcomes Partnership (OMOP) concepts missing from the OMOP2OBO mapping set over their frequency of use in clinical practice (log 10 scale) within the Concept Prevalence Study data by domain (Conditions [left]; Drugs [middle]; and Measurements [right]). The three modeled error analysis distributions include: concepts recovered in a newer version of the OMOP common data model (CDM; magenta), concepts that were purposefully excluded, not yet mapped, or unable to be mapped by OMOP2OBO (blue), and concepts that were truly missing from the OMOP2OBO mapping set (yellow).

Table 3. Concept prevalence concepts missing from the OMOP2OBO mapping set.

OMOP domain	Concept	Concept label[a]	Average concept frequency[b]	Study sites
Condition	4,091,502	Increased fluid intake	100,483.0	1
	37,311,061	COVID-19	93,585.0	1
	40,443,308	Polycystic ovary syndrome	62,900.3	3
	35,615,055	Saddle embolus of pulmonary artery with acute cor pulmonale	22,324.4	10
	36,684,319	Adjustment disorder with mixed anxiety and depressed mood	18,453.0	1
Drug Ingredient	37,498,625	Hepatitis A virus strain CR 326F antigen, inactivated	175,551.3	14
	1,510,467	Erenumab	60,618.0	10
	35,200,577	Fremanezumab	15,579.6	5
	35,200,800	Galcanezumab	11,594.8	5
	35,201,105	Baloxavir marboxil	11,366.7	3
Measurement	3,045,980	Pulse intensity of Unspecified artery palpation	1,219,846,862.0	1
	3,021,716	Penicillin G potassium [Mass] of Dose	253,609,945.0	1
	40,760,098	Sodium [Moles/volume] in Saliva (oral fluid)	246,641,311.0	1
	3,045,820	Cotinine/Creatinine [Mass Ratio] in Urine	246,063,202.0	1
	3,008,500	Chloride [Moles/volume] in Saliva (oral fluid)	234,931,483.0	1

OMOP Observational Medical Outcomes Partnership.
[a]Concept labels were obtained from the Athena web application (https://athena.ohdsi.org/) on 12/29/2022.
[b]The average concept frequency was calculated as the frequency of each concept divided by the number of Concept Prevalence study sites with that concept by each clinical domain.

mapping set contained 8611 eligible drug ingredient concepts, which covered 87.9% (99.9% weighted coverage) of the OHDSI Concept Prevalence Data concepts ($n = 4037$ concepts; median 7299, range 100–1,308,580,305). Of the remaining drug ingredient concepts, 4574 were only found in OMOP2OBO (median 100, range 100–69,311) and 551 were only found in the OHDSI Concept Prevalence Data (median 300, range 100–10,748,492). These findings are visualized in Fig. 7. OMOP2OBO drug ingredient concept coverage ranged from 91.2–98.4% across the 18 Concept Prevalence Study sites. Supplementary Fig. 5a presents the counts of OMOP drug ingredient concepts in the OHDSI Concept Prevalence Data by site. A chi-square test of independence with Yate's correction revealed a significant association between the OHDSI Concept Prevalence Data sites and OMOP2OBO coverage ($\chi^2(17) = 195.6$, $p < 0.0001$). Post hoc tests using Bonferroni adjustment confirmed that 22% of the pairwise OHDSI Concept Prevalence Data site comparisons had significantly different OMOP2OBO coverage ($p < 0.001$ for all significant comparisons). The results of this analysis are visualized as a heatmap in Supplementary Fig. 5b. The OMOP2OBO drug ingredient concept count by OBO Foundry ontology, data wave, and coverage type are shown in Supplementary Fig. 6.

Results for the 551 (12%) OHDSI Concept Prevalence Data drug ingredient concepts missing from OMOP2OBO are visualized in Fig. 8. Roughly 0.9% ($n = 5$) of drug ingredient 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 (range 100–221,229.7). In total, 82.8% ($n = 456$) of drug ingredient concepts purposefully excluded from the OMOP2OBO mapping set (i.e., not yet mapped) occurred in an average of 3.9 sites with a mean frequency of 18,847.3 (range 100–1,077,258.9). The remaining drug ingredient concepts (16.3%; $n = 90$) were truly missing and occurred in an average of 2.7 sites with a mean frequency of 3361.2 (range 100–175,551.3). The frequency of distributions of the drug ingredient concepts covered by OMOP2OBO and Concept Prevalence drug ingredient concepts missing from OMOP2OBO in the OHDSI Concept Prevalence Data by site are visualized in Supplementary Fig. 5c, d. The five most frequently occurring missing drug ingredient concepts are shown in Table 3. Domain expert review of these drug ingredient 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. The domain experts identified comparable drug ingredient concepts in the OMOP2OBO mapping set.

The OHDSI Concept Prevalence Data contained 25,513 measurement concepts from 18 independent sites. The resulting OMOP2OBO mapping set contained 3828 eligible measurement concepts ($n = 10{,}676$ results), which covered 11.1% (67.7% weighted coverage) of the OHDSI Concept Prevalence Data measurement concepts ($n = 2260$ concepts; median 1355, range 100–1,465,815,430). Of the remaining measurement concepts, 1208 were only found in OMOP2OBO (median 100, range 100–1,842,485) and 20,893 were only found in the OHDSI Concept Prevalence Data (median 109, range 100–1,219,846,862). These findings are visualized in Fig. 7. OMOP2OBO measurement concept coverage ranged from 4.2–75% across the 18 OHDSI Concept Prevalence Data sites. Supplementary Fig. 7a presents the counts of OMOP measurement concepts in the OHDSI Concept Prevalence Data by site. A chi-square test of independence with Yate's correction revealed a significant association between the OHDSI Concept Prevalence Data sites and OMOP2OBO coverage ($\chi^2(17) = 3872.3$, $p < 0.0001$). Post hoc tests using Bonferroni adjustment confirmed that 56% of the pairwise OHDSI Concept Prevalence Data site comparisons had significantly different OMOP2OBO coverage ($p < 0.001$ for all significant comparisons). The results of this analysis are visualized as a heatmap in Supplementary Fig. 7b. The OMOP2OBO measurement concept count by OBO Foundry ontology, data wave, and coverage type are shown in Supplementary Fig. 8.

Results for the 20,893 (81.9%) OHDSI Concept Prevalence Data measurement concepts missing from OMOP2OBO are visualized in Fig. 8. Roughly 0.1% ($n = 13$) of measurement 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 9836.3 (range 100–29,098.2). In total, 0.8% ($n = 158$) of measurement concepts purposefully excluded from the OMOP2OBO mapping set (i.e., not mapped test type, unspecified sample, or were unable to be mapped) occurred in an average of 5.2 sites with a mean frequency of 282,115.3 (range 100–14,317,951.9). The remaining measurement 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 (range 100–1,219,846,862). The frequency of distributions of the measurement concepts covered by OMOP2OBO and Concept Prevalence measurement concepts missing from OMOP2OBO in the OHDSI Concept Prevalence Data by site are visualized in Supplementary Fig. 7c, d. The five most frequently occurring missing measurement concepts (reported as the average frequency across the 18 sites and number of sites with that concept) are shown in Table 3. Domain expert review of these measurement concepts confirmed that they were likely missing due to inconsistencies in hospital use of LOINC, a finding that's been observed in literature[76]. The domain experts identified comparable measurement concepts in the OMOP2OBO mapping set.

Validation: clinical utility

Many patients with a genetic disease never receive a specific diagnosis, even after genetic sequencing[77–80]. Longitudinal EHR data has been used to identify patients with genetic disorders[81–84]. Inspired by the fact that most genetic diseases manifest as a recurring pattern of multiple symptoms or phenotypes affecting multiple organ systems[82], the phenotype risk score (PheRS), which measures the similarity between an individual's diagnosis codes and phenotypic features of known genetic disorders, was developed[81]. While the PheRS has shown great promise for identifying patients with undiagnosed Mendelian disease from EHR data[58], it requires mappings that link ICD codes to HPO concepts, which most EHRs do not contain. The existing mappings[58] developed to support PheRS were manually constructed, which may limit scalability when applied to new data.

The goal of this evaluation was to determine if the OMOP2OBO mappings could be used to facilitate the application of the PheRS to EHR data and to compare their performance to an existing set of validated manual mappings. For this analysis, the OMOP2OBO HPO mappings were compared to the ICD-HPO mappings[58] using data from the AoU Research Program. The AoU Data were selected for this task because it provides access to a large sample of EHR data and genetic testing results (see Supplementary Table 3 for additional details on this data source). Five genetic diseases (and their associated genes) for which diagnosis codes have been found to be of high positive predictive value in EHRs[58], were examined: Marfan syndrome (*FBN1* and *TGFBR1*), multiple endocrine neoplasia (*MEN1* and *RET*), neurofibromatosis (*NF2*), paragangliomas (*SDHAF2*, *SDHB*, *SDHC*), and tuberous sclerosis (*TSC1*, *TSC2*). These diseases were associated with 2257 unique phenotypic features (HPO codes). When querying AoU data to identify patients who had at least one of these phenotypic features, the ICD-HPO mappings ($n = 7815$ ICD codes) took ~30 min to complete and returned 210,718 patients and the OMOP2OBO mappings ($n = 3783$ OMOP concepts) took ~10 min to complete and returned 209,342 patients. Of the 208,831 patients found in common, 1887 were only identified by the ICD-HPO mappings, and 601 patients were only identified by the

OMOP2OBO mappings. When the PheRS was applied to patients from both mappings they were found to be highly correlated ($r^2 > 0.6$ across all diseases). This suggests that the patients returned by both mappings were similar.

For additional validation, case-control studies using only the OMOP2OBO mappings were performed: Marfan syndrome (131 cases and 63,086 controls), multiple endocrine neoplasia (86 cases and 72,150 controls), neurofibromatosis (255 cases and 65,256 controls), paraganglioma (105 cases and 65,256 controls), and tuberous sclerosis (38 cases and 58,555 controls). The results of these studies are shown in Supplementary Table 8 and the distributions of PheRS scores for cases and controls for each of the five diseases are visualized in Supplementary Fig. 9. As shown in this figure, PheRS were higher for cases than controls across all examined diseases. These results are further supported by one-sided Wilcoxon rank sum tests, which indicated that the PheRS were significantly higher for cases than controls ($p < 0.001$ for all diseases). Collectively, these results support the use of OMOP2OBO mappings as a scalable alternative to an existing set of validated manual mappings for use with PheRS to aid in the systematic identification of patients who might benefit from genetic testing.

DISCUSSION

In this paper we present OMOP2OBO, an algorithm that semantically aligns conditions, drug ingredients, and measurement results from standard vocabularies in the OMOP CDM to OBO Foundry ontologies. Using OMOP2OBO, we built mappings for 92,367 condition, 8615 drug ingredient, and 10,673 measurement result concepts to ontology concepts representing 9636 diseases, 6309 phenotypes, 83 anatomical entities, 2704 organisms, 4261 chemicals, 132 vaccines, and 272 proteins. The mappings were evaluated on accuracy, generalizability, and clinical utility. For the first task, a panel of 10 domain experts reviewed subsets of the manually-derived mappings from each of the OMOP domains and found that 73.9% of the condition, 70.7% of the drug ingredient, and 92.9% of the measurement result mappings were correct. For the second task, we examined the generalizability of the concepts and found that 99.5% of conditions, 99.9% of drug ingredients, and 68% of measurement results overlapped with concepts used in clinical practice from 24 independent hospitals and claims databases. For the final task, we compared OMOP2OBO HPO mappings to an existing set of validated manual mappings when used to identify patients with five rare genetic diseases using data from the AoU Research Program. Queries using the OMOP2OBO mappings identified 99.3% of the patients returned by the validated manual mappings using fewer codes and one-third of the time. To the best of our knowledge, the OMOP2OBO mappings are the largest set of publicly available mappings between clinical vocabularies and OBO Foundry ontologies. The OMOP2OBO algorithm can easily be incorporated into existing clinical workflows and presents new opportunities to advance EHR-based deep phenotyping (recently published examples are described below).

Existing work to develop mapping sets and mapping algorithms has largely focused on using ontologies to improve the phenotyping of specific diseases (e.g., infectious disease[85], rare diseases[86,87], and cancer[88]) and for the investigation of specific biological (e.g., glycobiology[89]) and clinical domains (e.g., laboratory test results[60] and medical diagnoses[61,90]). Our work is most similar to LOINC2HPO[60], which we have expanded in our current mapping set (with annotations to five additional OBO Foundry ontologies). OMOP2OBO complements existing phenotyping efforts like the Electronic Medical Records and Genomics or eMERGE Network (https://emerge-network.org) and the AoU Research Program, 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.

A portion of the mappings that are automatically derived by OMOP2OBO overlap with existing mappings provided by the OMOP CDM, the UMLS, and the OBO Foundry ontologies. The UMLS and OMOP CDM each align more than 200 vocabularies. At the time of our analysis (and determined using the same data), only the UMLS provided mappings to an OBO Foundry ontology, which included: the Gene Ontology[91] (67,807 CUIs covering 69 vocabularies and an average of 166.9 codes), HPO (16,154 CUIs covering 91 vocabularies and an average of 1668.7 codes), and NCBITaxon (1,776,212 CUIs covering 55 vocabularies and an average of 3236.1 codes). Of these mappings, only the HPO and NCBITaxon are relevant to our work. Of the 1,776,212 CUIs aligned to NCBITaxon, 1128 were mapped to LOINC and 138 were mapped to RxNorm covering 0% of the measurement and 1.1% of the drug ingredient concepts in the CHCO OMOP Database, respectively. Of the 16,154 CUIs aligned to HPO, 993 were mapped to LOINC and 18,212 were mapped to SNOMED-CT covering 0% of the measurement and 4.2% of the condition concepts in the CHCO OMOP Database, respectively. Similar to the OMOP CDM and the UMLS, some of the OBO Foundry ontologies provide mappings to vocabularies in these resources. Collectively, the eight OBO Foundry ontologies used in this work provided 489,794 unique database cross-references from 179 unique data sources. Of these, only the HPO (11,616 ontology concepts to 19,569 codes from 16 data sources), Mondo (22,110 ontology concepts to 159,918 codes from 45 data sources), CL (949 ontology concepts to 1376 codes from 29 data sources), and Uberon (10,865 ontology concepts to 51,322 codes from 91 data sources) mappings were relevant to our work. Of the 19,569 HPO and 159,918 Mondo database cross-references only 3.6% and 15.6% mapped to a condition concept in the CHCO OMOP Database, respectively. These findings highlight that while there are some existing mappings between the resources that OMOP2OBO aligns, at best, they covered only ~15% of the OMOP concepts that we aimed to map supporting the need for its development. Further, it should be noted that the vast majority of the mappings provided by the OMOP CDM, UMLS, and OBO Foundry ontologies are simple one-to-one mappings. While OMOP2OBO contributes one-to-one mappings, it also provides more complex one-to-many mappings.

The OMOP2OBO mappings have been used to characterize differences in definitions of long COVID[92], generate long COVID phenotypes[93,94], and improve the categorization and prediction of psychiatric diseases among patients with long COVID[95]. 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[96]. 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. One potential use of OMOP2OBO is to aid in the alignment of patient data to ontologies in the Global Alliance for Genomics and Health's Phenopacket schema[97], which was designed to support the global exchange of computable patient-level phenotypic information.

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 mapped concepts and would enable use of other UMLS-aligned resources like the SemMedDB[98]. 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 to provide citations from the resources from which they were derived. 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. Providing better training and offering outcomes other than correct/incorrect should be considered. Finally, 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.

There are two primary challenges that remain given the initial development of the OMOP2OBO algorithm and mapping set. The first challenge is to establish procedures and build infrastructure to enable community sharing, monitoring, and updating of the mappings. While the GitHub repository for the OMOP2OBO currently contains policies for contributing to the mapping algorithm, we have yet to establish an infrastructure or policies for the mappings. Future opportunities include the adoption of a system like the one utilized by the Bioregistry (https://bioregistry.io)[99]. The Bioregistry provides extensive governance policies and templates, which make it easy to incorporate new and modify existing identifiers. They also developed a robust, semi-automated infrastructure that facilitates review by the maintainers and triggers rebuilds of the registry anytime changes are made. To improve the shareability of the mappings, we would also like to extend the mapping output formats to include Semantic Web standards like RDF/XML and the Simple Standard for Sharing Ontological Mappings or SSSOM[100]. In addition to creating a system like the Bioregistry, future work may include adoption and adaptation of OBO Foundry protocols for ontology development and maintenance[57,101].

The second challenge is to improve and expand the evaluation of the algorithm and the mapping set. The UMLS, OMOP CDM, and the OBO Foundry ontologies provide mappings between clinical vocabularies and ontologies, which are automatically- or manually-derived (e.g., mappings between source and standard vocabulary concepts, mappings between clinical vocabularies and ontologies, and/or database cross-references mapped to ontology concepts). While the OMOP2OBO algorithm leverages these mappings (i.e., leveraging source codes mapped to standard concepts), verifying the quality of existing mappings was not within the scope of the current work. Currently, no modules in the OMOP2OBO algorithm verify the quality of existing mappings used by OMOP2OBO or mappings generated by it. This should include resources to validate automatic 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[102–104]. To do this, we might leverage pretrained language models and/or develop new machine learning models using trusted resources (e.g., the scientific literature) to verify the database cross-references provided by the OBO Foundry ontologies, UMLS, and OMOP CDM database prior to running OMOP2OBO.

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 the OMOP2OBO algorithm and mappings are described in Supplementary Table 2.

OMOP2OBO algorithm: resources

Although it is possible to apply the OMOP2OBO algorithm to any clinical vocabulary, the OMOP CDM was selected because of its rich data representation, standard vocabularies (and hierarchies) and the mappings it provides to more than 200 commonly used clinical vocabularies. To increase the coverage of the resources and the potential of an automatic mapping, OMOP2OBO leverages the National Library of Medicine's UMLS (MRCONSO and MRSTY tables [2020AA version[105]])[70]. These data are used to annotate each OMOP concept with a UMLS CUI and a Semantic Type[71]. Additionally, the mappings provided by the MRCONSO table are used to enhance existing database cross-reference mappings provided by OMOP and the OBO Foundry ontologies (both described in detail in the "Input data used to create OMOP2OBO mappings" section).

OMOP2OBO algorithm: overview

The OMOP2OBO algorithm (Fig. 1) consists of the following three components: (1) Process Input Data; (2) Map OMOP Standard Vocabulary Concepts to OBO Foundry Ontology Concepts; and (3) Synthesize and Process Mapping Results and Output Mappings. Each component is described in detail below.

The first component is Processing Input Data. The algorithm takes as input a table of OMOP concepts and a list of one or more OBO Foundry ontologies. For both types of data, the algorithm expects concept or class identifiers, source codes or database cross-references, labels, synonyms, and ancestor concepts or classes. While the algorithm expects a table of input OMOP concepts (due to the private nature of clinical data, the algorithm does not assume a direct database connection is possible), it automatically downloads the OBO Foundry ontologies using OWLTools (April 06, 2020 release; https://github.com/owlcollab/owltools).

The second component is Mapping OMOP Vocabulary Concepts to OBO Foundry Ontology Concepts. This component is designed to automatically map or align OMOP concepts to OBO Foundry ontology concepts. The algorithm includes several different approaches, prioritizing those that result in high-confidence mappings. This component includes concept alignment and concept embedding.

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 database cross-references. 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[106], 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 database cross-references are normalized using a custom dictionary (source_code_vocab_map.csv[107]). This resource ensures that concepts referenced by the same code using different prefixes or symbols can be aligned (e.g., SNOMED-CT:1234567 and sctid:1234567).

Concept embedding: using scikit-learn[108], a bag-of-words (BoW)[109] vector space model with term-frequency inverse-document frequency (TF-IDF)[110] and L2 normalization is used to learn concept embeddings for all OMOP and OBO Foundry ontology concepts and concept ancestors label and synonym text strings. While the BoW model was used because it is easy to understand and has shown great success when applied to EHR data and when used to align biomedical ontologies[111,112], any language or embedding model could be utilized. The BoW model is implemented as an $N \times M$ document-term matrix with

one row per input string and one column for each tokenized word appearing in the universe of all input strings. The value of each cell in the matrix is the normalized frequency each word occurred in each input string (using TF-IDF normalization). Prior to building the model, all text fields are made lowercase, stop words are removed using the wordnet list from Python's NLTK library[113], white spaces are removed, and word-level tokenization and lemmatization are applied. After learning the model, a final embedding is constructed for each input string by aggregating the constituent concept embeddings. Cosine similarity is used to compute scores between all pairwise combinations of OMOP and OBO Foundry concept embeddings. Given that each OMOP and OBO concept can have a label and one or more synonyms, only the single highest-scoring pairwise comparison is selected for the final mapping. Cosine similarity scores range from 0–1, where a score of one indicates a greater match between the embedding pairs. 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. Concept embeddings are created for all OMOP concepts, regardless of whether or not they were automatically mapped by a prior Component. All remaining unmapped concepts require manual curation.

The third component is Synthesizing and Outputting Mapping Results. The mapping results from the prior component are post-processed to include a mapping category and human-readable evidence. The mapping category is constructed by combining the following elements: (1) one or more OBO Foundry ontology identifiers and labels; (2) mapping logic applied to specify semantics when there are multiple ontology concepts (i.e., "and", "or") or to denote negation (i.e., "not"); (3) a mapping category derived from the mapping approach (e.g., automatically determined using an algorithm or manually derived by a human annotator), cardinality (i.e., one-to-one aligning a single OMOP concept to a single OBO Foundry ontology concept or one-to-many aligning a single OMOP concept to one or more OBO Foundry ontology concepts), and level (i.e., mapping to the OMOP concept directly or to one of its ancestors); and (4) mapping evidence represented as a pipe-delimited string that denotes all resources that support the mapping (i.e., the exact string matches between labels and synonyms, source codes and database cross-reference alignments, and other sources supporting a mapping like scored heuristics and references from manual review). Supplementary Table 4 provides additional details on and examples of the mapping categories. Post-processed mappings are serialized and able to be output to a variety of file types, like flat file, database dump, or RDF/XML file.

Input data used to create OMOP2OBO mappings

The OMOP2OBO mappings were constructed from two data sources: (1) the CHCO OMOP Database and (2) OBO Foundry ontologies (both are described in detail below). Figure 2 includes example mappings and illustrates how the OBO Foundry ontologies were used to map OMOP concepts from each domain. Supplementary Table 4 provides additional details on and examples of the mapping categories. Supplementary Table 3 provides descriptions of the clinical data sources used to build and validate the OMOP2OBO mappings.

Input data used to create OMOP2OBO mappings: OMOP data

The OMOP2OBO mappings were constructed using data from the CHCO OMOP Database, a de-identified database that contained data from more than six million pediatric patients. The CHCO OMOP Database is stored within University of Colorado Anschutz Medical Campus Health Data Compass' Health Insurance Portability and Accountability Act compliant Google Cloud-based infrastructure (created in October 2018; https://www.healthdatacompass.org). The data conformed to the structure defined by the National Pediatric Learning Health System (PEDSnet) OMOP CDM v3.0, which is an adaptation of the OMOP CDM version 5.0[53,62]. Data were obtained from a de-identified database that was determined by the Colorado Multiple Institutional Review Board to be non-human subjects (#15-0445). Due to the broad scope of the projects approved to be performed on this database, a Waiver of consent was obtained as it was not practical to obtain consent from all patients.

Concept lists were derived from standard OMOP vocabularies (i.e., SNOMED-CT [https://www.snomed.org; v20180131], RxNorm[52] [v20180507], and LOINC[51] [v2.64]) from the Condition Occurrence, Drug Exposure (at the drug ingredient level), and Measurement tables, respectively. For each concept set, 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). Concept lists were organized into two data waves according to whether or not they had been used in clinical practice (i.e., Concepts Used in Practice and Concepts Not Used in Practice). As manual annotation requires significant resources, only concepts from the first data wave (i.e., Concepts Used in Practice) were manually mapped. Prior to constructing the concept lists, Condition and Measurement data were preprocessed to ensure the mapping process was robust and reproducible.

For condition concepts, 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 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.

For all measurement concepts, 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 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". 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 aimed to maximize the inclusion of concepts from all scale types. Mappings were created for each result type using the procedures defined by LOINC2HPO[60]; 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.

The initial set of measurement concepts was supplemented with LOINC2HPO annotations[60], which were downloaded on August 2, 2020 from the LOINC2HPO annotation Github repository[114]. 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. All LOINC concepts in the LOINC2HPO mappings that were not used at least once in clinical practice in the CHCO pediatric OMOP Database were categorized as a Concept Not Used in Practice.

Input data used to create OMOP2OBO mappings: 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[115] [v2020-09-14]), phenotypes (HPO[59] [v2020-08-11]), anatomical entities (CL[65] [v2020-05-21], Uberon[64] [v2020-06-30]), organisms (NCBITaxon[66] [v2020-04-18]), chemicals (ChEBI[67] [v191]), vaccines (VO[68] [v1.1.102]), and proteins (PRO[69] [v61.0]). Similar to the clinical concepts, each ontology was queried to obtain labels, definitions, synonyms (including synonym type), and database cross-references. All OBO Foundry ontologies were downloaded in September 2020 using OWLTools (April 06, 2020 release; https://github.com/owlcollab/owltools).

Mapping evaluation

The OMOP2OBO mappings were evaluated by assessing their accuracy, generalizability, and clinical utility.

Mapping evaluation: accuracy

Automatic mappings are created from exact alignments between resources available in the OMOP CDM and the OBO Foundry ontologies and thus are assumed to be accurate and high-confidence mappings. The goal of this evaluation was to examine the accuracy of a portion of the manually-derived mappings. For conditions and drug ingredients, of all manual mappings (including one-to-one and one-to-many), 20% were randomly selected for manual review ($n = 2000$ conditions; $n = 116$ drug ingredients) by a practicing resident physician (J.M.W.) and clinical pharmacist (J.S.), respectively.

Measurement mappings are significantly more complex as they require interpreting lab test results and annotating the source of the sample (e.g., bodily fluid, anatomical entity, or cell type), entity being measured (e.g., chemical or cell type), and organism of the measured entity. While annotating the samples and entities is straightforward, interpreting lab tests results and aligning them to HPO concepts can be challenging. As a result, only the HPO mappings were evaluated by domain experts. These mappings 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 (T.D.B., J.A.F., and B.M.), a PhD-level molecular biologist (A.L.S.), and a Computational Biology PhD candidate with Masters-level training in epidemiology and biostatistics (T.J.C.) using a Qualtrics Survey[116]. 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 (https://github.com/TheJacksonLaboratory/loinc2hpoAnnotation/issues) by members of the biocuration team. (2) **Biocurator validation**. A random subset of 1350 measurement results were manually verified by an OBO Foundry biocurator.

All of the manual mappings were derived by a member of the research team who at the time of the analysis was a Computational Biology PhD candidate with Masters-level training in epidemiology and biostatistics (T.J.C.). As this individual does not have specialized medical or pharmacological training, it is assumed that these mappings may contain errors. Additional details are provided on GitHub (https://github.com/callahantiff/OMOP2OBO/wiki/Accuracy).

Mapping evaluation: generalizability

The generalizability of the OMOP2OBO mappings were examined using the OHDSI Concept Prevalence Study data[72–75]. The Concept Prevalence study provides data on the frequency of OMOP concept usage in clinical practice across several independent sites in the OHDSI network. In addition to the Concept Prevalence Study sites, data were obtained from two independent academic medical centers, bringing the total number of sites to 24. None of the 24 sites overlapped with the site that was used to generate the OMOP2OBO mappings. Consistent with the Concept Prevalence Study procedures, all concepts from the OMOP CHCO Database occurring fewer than 100 times were assigned a count of 100. For all other analyses, the true range of counts in the OMOP CHCO Database were utilized. 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). Institutional review board approval was not required to use these data as the dataset was completely de-identified and contained no patient-level information.

An error analysis was performed to examine the Concept Prevalence Only concept set. Three scenarios were examined: (1) **"Recovered in Newer Version of CDM"**: concepts that could be recovered using a newer version of the OMOP CDM (v5.3.1; 02/25/2022); (2) **"Purposefully Excluded"**: concepts without clear pathological or biological origin that were purposefully excluded from the OMOP2OBO mapping set; and (3) **"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. See GitHub for additional details (https://github.com/callahantiff/OMOP2OBO/wiki/Generalizability).

Mapping evaluation: clinical utility

The clinical utility of the OMOP2OBO mappings was compared to an existing set of validated manual mappings (ICD-HPO mappings[58]) when used to identify undiagnosed rare disease patients. For this analysis, AoU Data (https://www.researchallofus.org) was selected because it provides access to a large sample of EHR data

with genetic testing results. For this evaluation, the version 6 build was used, which contained data from ~630 sites on more than 528,000 patients. Five genetic diseases for which diagnosis codes have been found to be of high positive predictive value in EHRs[58] were selected, which included: Marfan syndrome, multiple endocrine neoplasia, neurofibromatosis, paraganglioma, and tuberous sclerosis. These diseases are associated with 11 of the 73 American College of Medical Genetics and Genomics (ACMG) secondary finding genes (ACMG-73; v3.0), which have specific mutations known to cause disorders, have well-defined phenotypes, and are clinically actionable[117]. The diseases and associated genes included: *FBN1* and *TGFBR1* (Marfan syndrome); *MEN1* and *RET* (multiple endocrine neoplasia); *NF2* (neurofibromatosis); *SDHAF2*, *SDHB*, and *SDHC* (paragangliomas); and *TSC1*, *TSC2* (tuberous sclerosis). Using the Online Mendelian Inheritance in Man (OMIM; https://www.omim.org) database and the HPO gene annotation table[118], each gene and its corresponding set of phenotypic features were aligned to the HPO. To calculate the phenotypic burden of each genetic disease, HPO mappings to OMOP condition concepts from OMOP2OBO (v2.0.0 beta) and ICD concepts from a validated set of ICD-HPO mappings[58] were queried against the AoU data. PheRS for each gene were then calculated for patients from each the OMOP2OBO and Phecode mapping sets. The PheRS[81] is an algorithm used to identify patients with phenotypic features that are clinically similar to OMIM Mendelian profiles but who lack formal diagnosis and has demonstrated utility for identifying underdiagnosed rare disease patients using only EHR data[58,81]. The standardized version of the PheRS was used because it is easier to interpret and reduces noise when it is suspected that a large number of phenotypes will overlap between cases and controls[81]. The OMOP2OBO and ICD-HPO mappings were compared and evaluated on time to complete the query against the AoU Data and differences in the returned patient cohorts. As validation, case-control studies were performed for each of the five diseases using the patients returned from the OMOP2OBO mappings. 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. For each disease, a one-sided Wilcoxon rank sum test was performed in order to determine if PheRS were significantly higher for cases than controls. Results were verified by a PhD-level Epidemiologist specializing in genetics (C.Z.).

All analyses were performed in the AoU Researcher Workbench (https://www.researchallofus.org/data-tools/workbench) by an authorized researcher (C.Z.). Informed consent is obtained from all participants who enroll in the AoU Research program[119]. Because the authors were not directly involved with the participants and all data were de-identified, the use of these data was exempt from institutional review. For additional details, see "Do I need my project reviewed by the AoU Institutional Review Board (IRB) in order to access this data using the Researcher Workbench?" for more information (https://www.researchallofus.org/frequently-asked-questions/#workbench-faqs).

Statistics and 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[120–122]. 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 (https://www.researchallofus.org/data-tools/workbench) using R (v4.1.2) and Python (v3.7). Analyses were performed on a machine with 16 CPUs and 60GB of memory.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Supplementary Table 2 lists the resources used by the OMOP2OBO algorithm. The MRCONSO and MRSTY tables (2020AA) require a license and are available through the UMLS (https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html). The data used to build and validate the OMOP2OBO mappings (v1) are described in Supplementary Table 3. The OMOP concepts are available for download through Athena (https://athena.ohdsi.org). The UMLS data require a license to use (https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html) and some of the OMOP CDM vocabularies also require a license. All users should read the licensing agreements for both resources and consult their institution before developing new mappings. The OBO Foundry ontologies are publicly available (https://obofoundry.org). The OMOP2OBO (v1.0) mappings are publicly available and can be downloaded from Zenodo: Conditions (https://doi.org/10.5281/zenodo.6774363); Drugs (https://doi.org/10.5281/zenodo.6774401); and Measurements (https://doi.org/10.5281/zenodo.6774443).

CODE AVAILABILITY

OMOP2OBO is publicly available through GitHub (https://github.com/callahantiff/OMOP2OBO) and PyPI (https://pypi.org/project/omop2obo). The interactive dashboard code is also available on GitHub (https://github.com/callahantiff/OMOP2OBO_Dashboard).

Received: 9 September 2022; Accepted: 28 April 2023;
Published online: 19 May 2023

REFERENCES

1. Adler-Milstein, J. & Jha, A. K. HITECH act drove large gains In hospital electronic health record adoption. *Health Aff.* **36**, 1416–1422 (2017).
2. Atasoy, H., Greenwood, B. N. & McCullough, J. S. The digitization of patient care: a review of the effects of electronic health records on health care quality and utilization. *Annu. Rev. Public Health* **40**, 487–500 (2019).
3. Dexter, P. R. et al. A computerized reminder system to increase the use of preventive care for hospitalized patients. *N. Engl. J. Med.* **345**, 965–970 (2001).
4. King, J., Patel, V., Jamoom, E. W. & Furukawa, M. F. Clinical benefits of electronic health record use: national findings. *Health Serv. Res.* **49**, 392–404 (2014).
5. Evans, R. S. Electronic health records: then, now, and in the future. *Yearb. Med. Inform.* **Suppl 1**, S48–S61 (2016).
6. Hulsen, T. et al. From big data to precision medicine. *Front. Med.* **6**, 34 (2019).
7. Robinson, P. N. Deep phenotyping for precision medicine. *Hum. Mutat.* **33**, 777–780 (2012).
8. Richesson, R. L., Sun, J., Pathak, J., Kho, A. N. & Denny, J. C. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif. Intell. Med.* **71**, 57–61 (2016).
9. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-wide association studies as a tool to advance precision medicine. *Annu. Rev. Genomics Hum. Genet.* **17**, 353–373 (2016).
10. Rossi, R. L. & Grifantini, R. M. Big data: challenge and opportunity for translational and industrial research in healthcare. *Front. Digit. Humanit.* **5**, 13 (2018).
11. Jha, S. & Topol, E. J. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA* **316**, 2353–2354 (2016).

12. Butte, A. J. Big data opens a window onto wellness. *Nat. Biotechnol.* **35**, 720–721 (2017).
13. Beam, A. L. & Kohane, I. S. Big data and machine learning in health care. *JAMA* **319**, 1317–1318 (2018).
14. Hinton, G. Deep learning—a technology with the potential to transform health care. *JAMA* **320**, 1101–1102 (2018).
15. Leopold, J. A. & Loscalzo, J. Emerging role of precision medicine in cardiovascular disease. *Circ. Res.* **122**, 1302–1315 (2018).
16. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
17. Freimer, N. & Sabatti, C. The human phenome project. *Nat. Genet.* **34**, 15–21 (2003).
18. Basile, A. O. & Ritchie, M. D. Informatics and machine learning to define the phenotype. *Expert Rev. Mol. Diagn.* **18**, 219–226 (2018).
19. Delude, C. M. Deep phenotyping: the details of disease. *Nature* **527**, S14–S15 (2015).
20. Weng, C., Shah, N. H. & Hripcsak, G. Deep phenotyping: embracing complexity and temporality-towards scalability, portability, and interoperability. *J. Biomed. Inform.* **105**, 103433 (2020).
21. Dorsey, E. R. et al. Deep phenotyping of Parkinson's disease. *J. Parkinsons. Dis.* **10**, 855–873 (2020).
22. Georgiou, M. et al. Deep phenotyping of PDE6C-associated achromatopsia. *Invest. Ophthalmol. Vis. Sci.* **60**, 5112–5123 (2019).
23. Fassihi, H. et al. Deep phenotyping of 89 xeroderma pigmentosum patients reveals unexpected heterogeneity dependent on the precise molecular defect. *Proc. Natl Acad. Sci. USA* **113**, E1236–E1245 (2016).
24. Russo, R. S. et al. Deep phenotyping in 3q29 deletion syndrome: recommendations for clinical care. *Genet. Med.* **23**, 872–880 (2021).
25. Daich Varela, M. et al. The peroxisomal disorder spectrum and Heimler syndrome: deep phenotyping and review of the literature. *Am. J. Med. Genet. C. Semin. Med. Genet.* **184**, 618–630 (2020).
26. Mei, C. et al. Deep phenotyping of speech and language skills in individuals with 16p11.2 deletion. *Eur. J. Hum. Genet.* **26**, 676–686 (2018).
27. Droogmans, G., Swillen, A. & Van Buggenhout, G. Deep phenotyping of development, communication and behaviour in Phelan-McDermid syndrome. *Mol. Syndromol.* **10**, 294–305 (2020).
28. Fernandes, S. A., Cooper, G. E., Gibson, R. A. & Kishnani, P. S. Benign or not benign? Deep phenotyping of liver glycogen storage disease IX. *Mol. Genet. Metab.* **131**, 299–305 (2020).
29. Mak, E. et al. Longitudinal trajectories of amyloid deposition, cortical thickness, and tau in down syndrome: a deep-phenotyping case report. *Alzheimers Dement.* **11**, 654–658 (2019).
30. Mishra, R. et al. Robinow syndrome and brachydactyly: an Interplay of high-throughput sequencing and deep phenotyping in a kindred. *Mol. Syndromol.* **11**, 43–49 (2020).
31. Welsink-Karssies, M. M. et al. Deep phenotyping classical galactosemia: clinical outcomes and biochemical markers. *Brain Commun.* **2**, fcaa006 (2020).
32. Shim, Y. et al. Deep phenotyping in 1p36 deletion syndrome. *Ann. Child Neurol.* **28**, 131–137 (2020).
33. Spedicati, B. et al. Natural human knockouts and mendelian disorders: deep phenotyping in Italian isolates. *Eur. J. Hum. Genet.* **29**, 1272–1281 (2021).
34. Yurkovich, J. T., Tian, Q., Price, N. D. & Hood, L. A systems approach to clinical oncology uses deep phenotyping to deliver personalized care. *Nat. Rev. Clin. Oncol.* **17**, 183–194 (2020).
35. Papadimitriou, K. et al. Deep phenotyping reveals distinct immune signatures correlating with prognostication, treatment responses, and MRD status in multiple myeloma. *Cancers* **12**, 3245 (2020).
36. Christopoulos, P. 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* **21**, 743 (2021).
37. Sirinukunwattana, K. et al. Improving the diagnosis and classification of Ph-negative myeloproliferative neoplasms through deep phenotyping. *bioRxiv* 762013 https://doi.org/10.1101/762013 (2019).
38. Nagaoka, K. 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* **8**, e001358 (2020).
39. Song, T. H. et al. Deep learning-based phenotyping of breast cancer cells using lens-free digital In-line holography. *bioRxiv* 2021.05.29.446284. https://doi.org/10.1101/2021.05.29.446284 (2021).
40. Kuai, R., Ochyl, L. J., Bahjat, K. S., Schwendeman, A. & Moon, J. J. Designer vaccine nanodiscs for personalized cancer immunotherapy. *Nat. Mater.* **16**, 489–496 (2017).
41. Paquette, A. G., Hood, L., Price, N. D. & Sadovsky, Y. Deep phenotyping during pregnancy for predictive and preventive medicine. *Sci. Transl. Med.* **12**, eaay1059 (2020).
42. Davidson, L. & Boland, M. R. Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. *Brief. Bioinform.* **22**, bbaa369 (2021).
43. Kennedy, S. H. et al. Deep clinical and biological phenotyping of the preterm birth and small for gestational age syndromes: The INTERBIO-21 st Newborn Case-Control Study protocol. *Gates Open Res.* **2**, 49 (2018).
44. Alterovitz, G. et al. SMART on FHIR genomics: facilitating standardized clinico-genomic apps. *J. Am. Med. Inform. Assoc.* **22**, 1173–1178 (2015).
45. Sperber, N. R. et al. Challenges and strategies for implementing genomic services in diverse settings: experiences from the Implementing GeNomics In pracTicE (IGNITE) network. *BMC Med. Genomics* **10**, 35 (2017).
46. Haendel, M. A. et al. A census of disease ontologies. *Annu. Rev. Biomed. Data Sci.* **1**, 305–331 (2018).
47. Hammond, W. E. Call for a standard clinical vocabulary. *J. Am. Med. Inform. Assoc.* **4**, 254–255 (1997).
48. Cornet, R. & Chute, C. G. Health concept and knowledge management: twenty-five years of evolution. *Yearb. Med. Inform.* **25**, S32–S41 (2016).
49. Haendel, M. A., Chute, C. G. & Robinson, P. N. Classification, ontology, and precision medicine. *N. Engl. J. Med.* **379**, 1452–1462 (2018).
50. Knibbs, G. H. The International Classification of Disease and Causes of Death and its revision. *Med. Dent. J.* **1**, 2–12 (1929). & Others.
51. McDonald, C. J. et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin. Chem.* **49**, 624–633 (2003).
52. Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T. & Moore, R. Normalized names for clinical drugs: RxNorm at 6 years. *J. Am. Med. Inform. Assoc.* **18**, 441–448 (2011).
53. Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G. & Stang, P. E. Validation of a common data model for active safety surveillance research. *J. Am. Med. Inform. Assoc.* **19**, 54–60 (2012).
54. Kho, A. N. et al. Practical challenges in integrating genomic data into the electronic health record. *Genet. Med.* **15**, 772–778 (2013).
55. Hoehndorf, R., Schofield, P. N. & Gkoutos, G. V. The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.* **16**, 1069–1080 (2015).
56. Smith, B. et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
57. Jackson, R. et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database* **2021**, baab069 (2021).
58. Bastarache, L. et al. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J. Am. Med. Inform. Assoc.* **26**, 1437–1447 (2019).
59. Köhler, S. et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
60. Zhang, X. A. et al. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *npj Digital Med.* **2**, 1–9 (2019).
61. Dhombres, F. & Bodenreider, O. Interoperability between phenotypes in research and healthcare terminologies—investigating partial mappings between HPO and SNOMED CT. *J. Biomed. Semant.* **7**, 3 (2016).
62. Forrest, C. B. et al. PEDSnet: a National Pediatric Learning Health System. *J. Am. Med. Inform. Assoc.* **21**, 602–606 (2014).
63. Reich, C. & Ostropolets, A. Chapter 5 standardized vocabularies. in *The Book of OHDSI* (ed. Observational Health Data Sciences) Online Edition (2021).
64. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* **13**, R5 (2012).
65. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
66. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **36**, D13–D21 (2008).
67. Hastings, J. et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219 (2016).
68. Xiang, Z. et al. VIOLIN: vaccine investigation and online information network. *Nucleic Acids Res.* **36**, D923–D928 (2008).
69. Natale, D. A. et al. The Protein Ontology: a structured representation of protein forms and complexes. *Nucleic Acids Res.* **39**, D539–D545 (2011).
70. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
71. McCray, A. T. Representing biomedical knowledge in the UMLS semantic network. in *High Performance Medical Libraries: Advances in Information Management for the Virtual Era* 45–55 (Meckler Corporation, 1993).
72. Ostropolets, A., Ryan, P. B. & Hripcsak, G. OHDSI network study: concept prevalence. https://forums.ohdsi.org/t/network-study-concept-prevalence/6562 (2019).

73. Ostropolets, A., Ryan, P. & Hripcsak, G. OHDSI network study: concept prevalence. https://github.com/ohdsi-studies/ConceptPrevalence (2020).
74. Ostropolets, A., Ryan, P. & Hripcsak, G. Concept Prevalence Study Protocol. https://github.com/ohdsi-studies/ConceptPrevalence/blob/master/extras/ConceptPrevalenceStudyProtocol_v1.0.docx (2020).
75. Ostropolets, A., Ryan, P. & Hripcsak, G. Phenotyping in distributed data networks: selecting the right codes for the right patients. *AMIA Annu. Symp. Proc.* **2022**, 826–835 (2022).
76. Lin, M. C., Vreeman, D. J., McDonald, C. J. & Huff, S. M. Auditing consistency and usefulness of LOINC use among three large institutions—using version spaces for grouping LOINC codes. *J. Biomed. Inform.* **45**, 658–666 (2012).
77. Kremer, L. S. et al. Genetic diagnosis of mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
78. Splinter, K. et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N. Engl. J. Med.* **379**, 2131–2139 (2018).
79. Groopman, E. E. et al. Diagnostic utility of exome sequencing for kidney disease. *N. Engl. J. Med.* **380**, 142–151 (2019).
80. Yang, Y. et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
81. Bastarache, L. et al. Phenotype risk scores identify patients with unrecognized mendelian disease patterns. *Science* **359**, 1233–1239 (2018).
82. Morley, T. J. et al. Phenotypic signatures in clinical data enable systematic identification of patients for genetic testing. *Nat. Med.* **27**, 1097–1104 (2021).
83. Ganesan, S. et al. A longitudinal footprint of genetic epilepsies using automated electronic medical record interpretation. *Genet. Med.* **22**, 2060–2070 (2020).
84. Movaghar, A. et al. Artificial intelligence-assisted phenotype discovery of fragile X syndrome in a population-based sample. *Genet. Med.* **23**, 1273–1280 (2021).
85. Kafkas, Ş. et al. PathoPhenoDB, linking human pathogens to their phenotypes in support of infectious disease research. *Sci. Data* **6**, 79 (2019).
86. Thompson, R. et al. Increasing phenotypic annotation improves the diagnostic rate of exome sequencing in a rare neuromuscular disorder. *Hum. Mutat.* **40**, 1797–1812 (2019).
87. 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.* **316**, 79–86 (2020).
88. Edgren, H., Mano, B. & Laaksonen, M. Efficient curation and ontology mapping of clinical and phenotypic data. *Cancer Res.* **78**, 2276–2276 (2018).
89. Gourdine, J.-P. F. et al. Representing glycophenotypes: semantic unification of glycobiology resources for disease discovery. *Database* **2019**, baz114 (2019).
90. 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.* **245**, 925–929 (2017).
91. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
92. Rando, H. M. et al. Challenges in defining long COVID: striking differences across literature, electronic health records, and patient-reported information. *medRxiv* https://doi.org/10.1101/2021.03.20.21253896 (2021).
93. Reese, J. et al. Generalizable long COVID subtypes: findings from the NIH N3C and RECOVER programs. *bioRxiv* https://doi.org/10.1101/2022.05.24.22275398 (2022).
94. Deer, R. R. et al. Characterizing long COVID: deep phenotype of a complex condition. *EBioMedicine* **74**, 103722 (2021).
95. Coleman, 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* https://doi.org/10.1101/2022.07.08.22277388 (2022).
96. Callahan, T. J., Hunter, L. E. & Kahn, M. G. Leveraging a neural-symbolic representation of biomedical knowledge to improve pediatric subphenotyping. https://doi.org/10.5281/zenodo.5746187 (2021).
97. Jacobsen, J. O. B. et al. The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol.* **40**, 817–820 (2022).
98. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G. & Rindflesch, T. C. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* **28**, 3158–3160 (2012).
99. Hoyt, C. T. et al. Unifying the identification of biomedical entities with the Bioregistry. *Sci. Data* **9**, 714 (2022).
100. Matentzoglu, N. et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database* **2022**, baac035 (2022).
101. Matentzoglu, N. et al. Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. *Database* **2022**, baac087 (2022).
102. Amith, M., He, Z., Bian, J., Lossio-Ventura, J. A. & Tao, C. Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *J. Biomed. Inform.* **80**, 1–13 (2018).
103. Vrandečić, D. Ontology evaluation. in *Handbook on Ontologies* (eds Staab, S. & Studer, R.) 293–313 (Springer, 2009).
104. Gómez-Pérez, A. Ontology evaluation. in *Handbook on Ontologies* (eds Staab, S. & Studer, R.) 251–273 (Springer, 2004).
105. National Library of Medicine. UMLS release file archives: 2020AA. https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html (2020).
106. Banda, J. M. OHDSI Ananke—a tool for mapping between OHDSI Concept Identifiers to Unified Medical Language System (UMLS) identifiers. https://github.com/thepanacealab/OHDSIananke (2020).
107. Callahan, T. J. OMOP2OBO Code Normalization Dictionary. OMOP2OBO: Initial Release. https://github.com/callahantiff/OMOP2OBO/blob/master/resources/mappings/source_code_vocab_map.csv; https://doi.org/10.5281/zenodo.5655853 (2020).
108. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
109. Harris, Z. S. Distributional structure. *Word World* **10**, 146–162 (1954).
110. Rajaraman, A. & Ullman, J. D. Data mining. in *Mining of Massive Datasets* 1–17 (Cambridge University Press, 2011).
111. Zhan, X., Humbert-Droz, M., Mukherjee, P. & Gevaert, O. Structuring clinical text with AI: old vs. new natural language processing techniques evaluated on eight common cardiovascular diseases. *bioRxiv* https://doi.org/10.1101/2021.01.27.21250477 (2021).
112. Kolyvakis, P., Kalousis, A., Smith, B. & Kiritsis, D. Biomedical ontology alignment: an approach based on representation learning. *J. Biomed. Semant.* **9**, 21 (2018).
113. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (O'Reilly Media, Inc., 2009).
114. Aaron, Z. X. et al. LOINC2HPO Annotations. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. https://github.com/monarch-initiative/loinc2hpo/annotations.tsv (2020).
115. Mungall, C. J. et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **45**, D712–D722 (2017).
116. Callahan, T. J. Survey to evaluate OMOP2OBO measurement mappings. *Qualtrics* https://survey.az1.qualtrics.com/jfe/form/SV_cAZvVBV7LU0YVa5?Q_CHL=qr (2018).
117. Miller, D. T. 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.* **23**, 1381–1390 (2021).
118. The Human Phenotype Ontology. Gene to phenotype annotations. http://purl.obolibrary.org/obo/hp/hpoa/genes_to_phenotype.txt (2022).
119. Ramirez, A. H., Gebo, K. A. & Harris, P. A. Progress with the All of Us Research Program: opening access for researchers. *JAMA* **325**, 2441–2442 (2021).
120. Callahan, T. J. et al. OMOP2OBO Condition Occurrence Mappings. https://doi.org/10.5281/zenodo.6949688 (2020).
121. Callahan, T. J. et al. OMOP2OBO Drug Exposure Ingredient Mappings. https://doi.org/10.5281/zenodo.6949696 (2020).
122. Callahan, T. J. et al. OMOP2OBO Measurement Mappings. https://doi.org/10.5281/zenodo.6949858 (2020).

ACKNOWLEDGEMENTS

This work was primarily supported by funding from the National Library of Medicine (NLM T15LM009451 and T15LM007079) to T.J.C. and in part by the National Center for Advancing Translational Sciences (NCATS U24TR002306) to M.A.H. and P.N.R., the National Human Genome Research Institute (NHGRI 5RM1HG010860) to M.A.H., P.N.R., N.A.V., and N.A.M., the NLM (R01LM013400) to L.E.H. and (R01LM006910) G.H., the Medical Research Council (MR/P02002X/1) to J.H.C., the National Heart, Lung, and Blood Institute (NHLBI 1K23HL161352) to K.E.T., the NHGRI (5U24HG011449-02) to P.N.R., and the Intramural Research Program of the NHGRI (ZIA HG200417) to J.C.D. and C.Z. The authors thank colleagues at the Health Data Compass warehouse, Children's Hospital Colorado Research Informatics team, and the OMOP2OBO and Machine Learning Working Groups at the National COVID Cohort Collaboration for piloting testing, extending, and improving the mappings. The authors would also like to thank Drs. Paul Schofield (University of Oxford) and members of Dr. Robert Hoehndorf's (King Abdullah University of Science and Technology) lab for their feedback on the mappings.

AUTHOR CONTRIBUTIONS

M.G.K. and L.E.H. served as primary supervisors of this work. T.J.C., M.G.K., and A.L.S. conceived and developed the analyses. T.J.C. and W.A.B. developed the OMOP2OBO algorithm with feedback from N.A.V. and J.M.B., A.L.S., and J.M.W. helped develop documentation. R.D.B., A.O., P.B.R., G.H., J.C.D., D.M., S.J.D.D., and A.E.W. provided data for the evaluation. P.N.R., X.A.Z., M.A.H., N.A.M., S.B.T., E.C., B.D.C., B.M., J.S., A.Y.L., J.H.C., J.R., J.M.W., A.L.S., J.A.F., T.D.B., N.A.V., K.E.T., and C.Z. reviewed, evaluated or aided in pilot testing the mappings and/or assisted with the error analysis. C.Z.

performed the "Clinical Utility" evaluation and J.C.D. reviewed the results. T.J.C. drafted the manuscript and all authors reviewed it, provided feedback, and approved the final version. During the publication process, the affiliation for X.A.Z. changed to Regeneron Genetics Center, Regeneron Pharmaceuticals Inc., Tarrytown, NY 10591, USA.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-023-00830-x.

Correspondence and requests for materials should be addressed to Tiffany J. Callahan.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.