JAMIA Open, 6(1), 2023, ooad007
https://doi.org/10.1093/jamiaopen/ooad007
Brief Communications

Brief Communications

Linking rare and common disease vocabularies by mapping between the human phenotype ontology and phecodes

Evonne McArthur [1], Lisa Bastarache[2], and John A. Capra [3,4]

[1]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, USA, [2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [3]Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, California, USA and [4]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA

Corresponding Author: Evonne McArthur, PhD, Vanderbilt Genetics Institute, Vanderbilt University, 2201 West End Ave, Nashville, TN 37235, USA; evonne.mcarthur@vanderbilt.edu

Received 18 October 2022; Revised 14 December 2022; Editorial Decision 20 December 2022; Accepted 31 January 2023

ABSTRACT

Enabling discovery across the spectrum of rare and common diseases requires the integration of biological knowledge with clinical data; however, differences in terminologies present a major barrier. For example, the Human Phenotype Ontology (HPO) is the primary vocabulary for describing features of rare diseases, while most clinical encounters use International Classification of Diseases (ICD) billing codes. ICD codes are further organized into clinically meaningful phenotypes via phecodes. Despite their prevalence, no robust phenome-wide disease mapping between HPO and phecodes/ICD exists. Here, we synthesize evidence using diverse sources and methods—including text matching, the National Library of Medicine's Unified Medical Language System (UMLS), Wikipedia, SORTA, and PheMap—to define a mapping between phecodes and HPO terms via 38 950 links. We evaluate the precision and recall for each domain of evidence, both individually and jointly. This flexibility permits users to tailor the HPO–phecode links for diverse applications along the spectrum of monogenic to polygenic diseases.

Key words: electronic health record, Mendelian genetics, phenotype ontology, medical phenome

Lay Summary

Rare and common diseases are often described using different terminologies that have no available translations. This makes sharing knowledge across these domains challenging. For instance, the Human Phenotype Ontology (HPO) is a vocabulary used to characterize the symptoms and features of rare diseases, while common diseases are often described by phenotype codes (phecodes) that are derived from clinical visits. This study fills this critical gap by creating and evaluating a map between phecodes and HPO terminology. The mapping is curated from multiple data sources and methods, including text-matching, the Unified Medical Language System (UMLS), and Wikipedia. We outline how this translation between rare and common disease vocabularies can be tailored to fit different applications. In conclusion, this map helps build a foundation for bridging genome biology and medicine and enabling new biomedical discoveries by connecting existing resources and tools across the spectrum of rare and common disease.

INTRODUCTION

Genetic traits have traditionally been stratified into distinct Mendelian (ie, monogenic) and complex (ie, polygenic) categories. However, the genetic causes of traits exist on a continuum, and Mendelian and complex diseases often share symptoms and features.[1] Nonetheless, different vocabularies are commonly used to describe diseases and their features. For example, the Human Phenotype Ontology (HPO) is a standard lexicon for describing symptoms, signs, and features of rare Mendelian disease.[2] HPO terminology forms the basis for gene–phenotype relationships in the most comprehensive genetic disease database, Online Mendelian Inheritance in Man (OMIM).[3,4] In contrast, most clinical encounters and electronic health record (EHR) research use International Classification of Diseases (ICD) codes. To facilitate discovery, ICD codes are often grouped into phecodes, which are manually curated categories of clinically-meaningful phenotypes.[5–7] A wealth of rich clinical and research data are annotated with HPO terms or ICD/phecodes. Together, these have the potential to enable discovery and translation of insights between rare and common diseases, yet no phenome-wide mapping between these vocabularies exists.

Previous attempts to link HPO terminology and EHR-related codes demonstrate feasibility, but they have been incomplete or domain-specific.[2] For example, approaches have been fruitful in linking HPO terms to other annotation sets including clinical laboratory data via Laboratory Observation Identifier Names and Codes (LOINC).[2,8] Phenotypic symptoms have also been mapped to diseases using ICD and HPO codes; yet, the links were often too general or too specific to accurately reflect typical disease manifestations.[9] Semantic mapping between HPO terms and another international health terminology system—Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT)—demonstrated the ability to identify complete and partial mappings between ontologies.[10–12] However, this work was before the integration of HPO into the Unified Medical Language System (UMLS) and, thus, represents an opportunity for improvement. We hypothesize that synthesizing both foundational and novel resources will allow for a robust mapping between vocabularies.

Despite these challenges, small-scale manual mappings between HPO and EHR annotations have demonstrated potential to facilitate innovation in biomedical informatics methods. For example, phenotype risk scores (PheRS) enable the recognition of undiagnosed Mendelian disease from EHR data using weighted aggregates of clinical phenotypes linked to diseases.[13] Manual mapping of HPO and phecodes for 16 Mendelian disorders allowed the identification of a genetic etiology for previously undiagnosed individuals in the EHR.[14] Notably, an HPO-focused mapping has a distinct advantage over OMIM disease-ICD code mapping, because it allows for the detection of individuals with rare disease symptomatology without an established ICD-coded diagnosis. In a domain-specific example, manually linking neurology-related problems in the EHR with HPO codes via Intelligent Medical Objects (IMO) terms uncovered longitudinal patterns underlying specific genetic causes of epilepsy.[15] These previous maps addressed only a fraction of the phenome, yet their approaches and downstream applications—including the discovery of genetic associations through phenome-wide association tests (PheWAS)[16]—would be applicable at the population-scale if a complete mapping were available.

Here, we define maps between phecodes and HPO terms to enable translation of insights between datasets annotated with these different disease ontologies. We integrate evidence from diverse complementary sources including text matching, the National of Library Medicine's Unified Medical Language System (UMLS),[17] existing software and knowledge bases that map ontologies—for example, SORTA[18] and PheMap[19]—and tools that leverage shared knowledge in Wikipedia articles—for example, WikiMedMap.[20] Using manual curation and review, we evaluate the precision and recall of each piece of evidence, both individually and jointly. This flexibility permits future users to select HPO–phecode links that are appropriate for diverse research questions along the spectrum of monogenic to polygenic diseases to advance precision medicine.

MATERIALS AND METHODS

Strategy for creating a map between phecodes and HPO terms

To create a map between phecodes and HPO terms, we use 2 kinds of evidence for linking these terminologies. First, we considered evidence that "directly" links phecodes with HPO codes. Second, we linked phecodes with HPO terms "indirectly" via established ICD code mappings. Because phecodes are manually curated hierarchies of ICD codes, any HPO term linked to an ICD code could then be mapped to the corresponding phecode. We identify these indirect links via both ICD-9-CM and ICD-10-CM codes. These direct and indirect links are graphically illustrated in Figure 1A.

HPO terms are from the release on April 14, 2022.[2] We exclude the ontology root (HP:0000001), root of all phenotypic abnormalities (HP:0000118), and terms that are modes of inheritance (HP:0000005), clinical modifiers (HP:0012823, eg, "mild"), frequency (HP:0040279, eg, "very rare"), blood groups (HP:0032223), and health status (HP:0032319, eg, "Unaffected"). These exclusions removed 246 total terms. Phecodes are from version 1.2 available in the PheWAS catalog.[16] We also replicated all analyses with Phecode X (Supplementary Figure S1), a new extended version with increased granularity and coverage of terms related to pregnancy, congenital anomalies, and neonatology. However, it is currently unpublished and undergoing continued quality control; thus, results from Phecode X are in the Supplementary Material.

Types of evidence

Building off the methods outlined by the previous, but incomplete, phecode–HPO mapping from Bastarache et al.,[13,14] we integrated several sources of evidence. Here, we outline each evidence type (Figure 1), but we provide further technical details and implementation examples in the Supplementary Text.

1. *String or sub-string match*: We include links where the HPO term—or any specified synonym of the HPO term—is an exact string match or a sub-string match of a phecode name (direct) or ICD code name (indirect). Medical spelling terminology was Americanized, punctuation was removed, and the case was not considered when matching. When sub-string matching, we permit word order permutations. To reduce false matches, we required that the sub-string was longer than 5 characters and at least two-thirds the length of the target string.
2. *UMLS*: The UMLS Metathesaurus organizes medical concepts with Concept Unique Identifier (CUI) strings. If an HPO code and an ICD code are annotated by the same CUI, we link them together because they describe the same medical concept (indirect only). We use UMLS version 2022AA.[17] We describe further details of the UMLS mapping, including synonymous

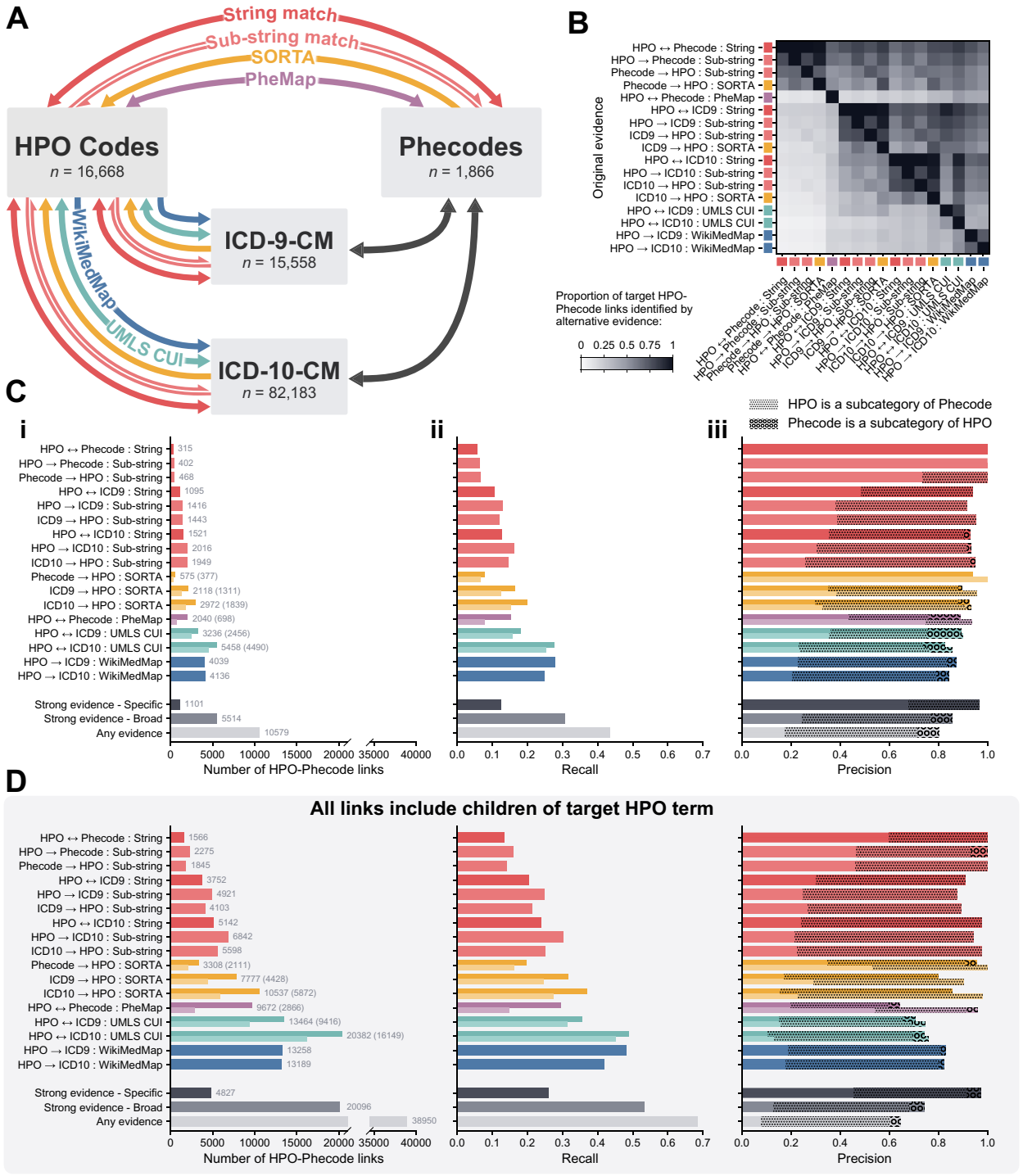

Figure 1. Creating and evaluating a map between phecodes and HPO terms. (A) HPO codes and phecodes can be linked by a variety of evidence (arrows). Some evidence links codes directly, indirectly via ICD codes, or both. Counts for ICD codes indicate how many codes have phecode links. (B) Many links identified by one piece of evidence (rows) are also identified using alternative evidence (columns). The proportion of links identified by the alternative evidence is shown. Supplementary Figure S3 is a larger annotated version of this figure and Supplementary Figure S4 replicates this with phecode X. (C) (i) We evaluate the number of links established by each evidence type and the corresponding (ii) recall and (iii) precision of these links. Evaluation of recall is based on 4200 manually defined HPO–phecode links, and evaluation of precision is based on manual review of 300 randomly selected links. For evidence that has a quantitative score, we report these measures at 2 thresholds (more stringent in the lighter bottom bars). Precision is evaluated using 3 designations—"exact" match (solid), HPO is a subcategory (dotted), and phecode is a subcategory (open circles)—and represented with stacked bars. (D) We depict the same analyses from C with the map that includes links that are children of the target HPO term (Materials and Methods). Counts, recall, and precision for the phecode X mapping are in Supplementary Figures S5–S7, respectively. Definitions for strong specific and broad evidence were curated based on the precision and recall and are discussed in the Conclusions.

relationships and how the mappings encompass evidence supported by SNOMED CT mappings, in the Supplementary Text.

3. *SORTA*: SORTA is a system for encoding free text to a formal coding system or ontology.[18] SORTA provides similarity scores between the query and the candidate match. We report links between HPO terms and phecodes (direct) or ICD codes (indirect) with scores above 80%, but output the raw score so users can define stricter thresholds if desired. For our analysis, we defined 100% similarity as the most stringent threshold.
4. *PheMap*: PheMap is a knowledge base that incorporates multiple online resources (Mayo Clinic Patient Care & Health Information website, MedlinePlus, MedicineNet, WikiDoc, and Wikipedia) to estimate the strength of relationships between phenotypes and medical concepts (encoded by CUI)[19] (Supplementary Text). We map phecodes to HPO codes (direct only) if PheMap (v1.1) identifies that the phecode is related to a CUI that is shared with an HPO code. This indicates the phecode and HPO code describe the same medical concept. We report links with scores in the top 95th percentile but also output raw scores. For our analysis, we define the 99th percentile as the most stringent threshold.
5. *WikiMedMap*: WikiMedMap is a tool that queries Wikipedia to extract ICD code references found in Wikipedia pages. It leverages a large database of alternative names and abbreviations for string normalization.[20] We map HPO terms to ICD codes (indirect only) if an ICD code is found within a page identified by an HPO term search. We note that links generated by WikiMedMap may update if run at another time because Wikipedia pages are living documents.

We report this map as a large table of phecode–HPO links and their corresponding supporting evidence. An example of this is shown in Table 1. We also create and evaluate a second set of phecode–HPO links that include all child HPO terms. For example, all phecodes linked to the HPO term "Anemia" would also be linked to more specific terms like "Microcytic anemia." If an HPO term had more than 100 children, we did not complete this linking process to avoid very broad terms getting linked to hundreds of narrow terms (eg, "Abnormality of limbs" has 2857 children). Traversal of the HPO ontology was aided by pyHPO.[2]

Evaluating phecode and HPO links

Each of the HPO–phecode maps was evaluated on their recall (sensitivity) and precision (positive predictive value). To test the recall in a realistic research scenario, we manually linked 4200 HPO codes that describe known Mendelian diseases in OMIM to their best-fitting phecode.[2,4] We quantify how many true positives are identified by each type of evidence over the total positives in our manual set. To test precision, we manually reviewed 300 randomly selected HPO–phecode links per map (1200 over 4 maps) to determine if the mapping was accurate using 3 designations: HPO and phecode are an "exact" match; the HPO term is a subcategory of the phecode; or, the phecode is a subcategory of the HPO term. We report precision for each of these designations of a "true match."

RESULTS

Five domains of evidence identify 38 950 phecode–HPO links

Our mapping identifies 10 579 links between phecodes and HPO terms, and an additional 28 371 links when including the children of HPO terms for a total of 38 950 unique links. On average, phecodes are linked to 6.2 HPO terms and HPO terms are linked to 3.1 phecodes which reflects the more granular nature of HPO terms. As expected, this increases when considering HPO term children: each phecode links to 22.9 HPO terms on average. Distributions of the link counts between HPO terms and phecodes are in Supplementary Figure S2.

Different sources of evidence identify both shared and unique phecode–HPO links (Figure 1B, Supplementary Figures S3 and S4). The most sharing occurs between string and sub-string matches (ie, by definition, all string matches are also sub-string matches). The most unique links are established by PheMap, UMLS (especially via ICD-10), and WikiMedMap which, respectively, identify 36%, 21%, and 17% of links that are not found by a second source of evidence. Considering all phecode–HPO links, 39% are identified by only one type of evidence and, on average, each link is supported by 3.3 types of evidence (Supplementary Figure S2).

Different sources of evidence identify different numbers of phecode–HPO links. Exact phecode–HPO term string matches establish the fewest links ($n = 315$) and UMLS CUI matches and synonyms (via ICD-10) identify the most ($n = 5458$) (Figure 1Ci, Supplementary Figure S5). This increases to 1566 and 20 382 when including HPO term children (Figure 1Di). In summary, 1698 phecodes (91%) are mapped to at least one HPO term and 7957 HPO terms (48%) are linked to at least one phecode (4522 HPO terms come from including children, Supplementary Figure S2).

Precision–recall tradeoff varies by the evidence used

Recall

Using a manually-curated map of 4200 Mendelian-disease HPO terms linked to phecodes, we evaluated recall for each type of evidence both individually and jointly. UMLS and WikiMedMap independently recall the most links (27.6%–27.9%) and string matching recalls the fewest (5.7%). Using all evidence, we can recall up to 43% of links (Figure 1Cii). However, given the higher granularity of HPO terms compared to phecodes we did not expect to recapitulate most links because there is unlikely to be a perfect semantic match in the more general phecode space. Accordingly, when we expand our map to include all HPO term children, we are able to recall 68% of the manual links (Figure 1Dii). Phecode X achieves a slightly increased recall up to 70%, likely due to its increased coverage of terms related to Mendelian disease (eg, congenital anomalies) (Supplementary Figure S6). We detail reasons for failing to identify manually curated links in the Supplementary Text, including a lack of reasonable link, a suitable identified alternative, or a more specific mapping.

Precision

We evaluate the precision of the links generated with manual review. As expected, evidence like string-matching—although it has low recall—is highly precise (100%). The lowest precision of 82.7% and 84.1% are, respectively, from UMLS and WikiMedMap via ICD-10 (Figure 1Ciii, Supplementary Figure S7). Some evidence is highly precise in matching terms at the same semantic level (eg, HPO–phecode string, sub-string, and SORTA). Other types of evidence correctly identify links but are more likely to do so at different levels. For example, ICD string, sub-string, SORTA, and PheMap links all have precision above 90%, but in many of their matches, the HPO term is a sub-category of the phecode match (up to 71% of links). Linking a broader HPO term to a narrower phecode is substantially less common but happens most frequently with UMLS-

Table 1. Phecode–HPO map is annotated by evidence type

			Evidence type	String	Sub-string	Sub-string	String	Sub-string	Sub-string	String	Sub-string	Sub-string	SORTA	SORTA	SORTA	PheMap	UMLS CUI	UMLS CUI	WikiMedMap	WikiMedMap
			Method of link	HPO phecode	HPO phecode	phecode HPO	HPO ICD9	HPO ICD9	ICD9 HPO	HPO ICD10	HPO ICD10	ICD10 HPO	phecode HPO	ICD9 HPO	ICD10 HPO	HPO phecode	HPO ICD9	HPO ICD10	HPO ICD9	HPO ICD10
Phecode	Phecode term	HPO code	HPO term																	
038	Septicemia	HP:0100806	Sepsis														✓	✓		✓
010	Tuberculosis	HP:0032262	Pulmonary tuberculosis				✓	✓	✓					100			✓	✓	✓	
070	Viral hepatitis	HP:0006562	Viral hepatitis	✓	✓	✓	✓	✓	✓	✓	✓	✓	100	100	100		✓	✓	✓	✓
230	Kaposi's sarcoma	HP:0002664	Neoplasm														✓	✓		
284.1	Pancytopenia	HP:0001915	Aplastic anemia													20				
345.1	Epilepsy	HP:0002069	Bilateral tonic-clonic seizure															✓	✓	✓

Note: We reproduce a subset of phecode–HPO mappings here to illustrate the data structure. The first 4 columns define the phecode and HPO terms linked. Subsequent columns describe the evidence used to link them. "✓" indicates that there was a link indicated by that evidence. Columns that report evidence from SORTA and PheMap have scores assigned to the match which can further be used to filter the mapping. We provide examples of similar or exact semantic matches (eg, septicemia-sepsis or viral hepatitis), an example where the HPO term is a child concept of the phecode (eg, pulmonary tuberculosis is a sub-category of tuberculosis), an example where the phecode is a child concept of the HPO term (eg, Kaposi's sarcoma is a sub-category of neoplasm), and an example of related concepts (eg, pancytopenia and aplastic anemia).

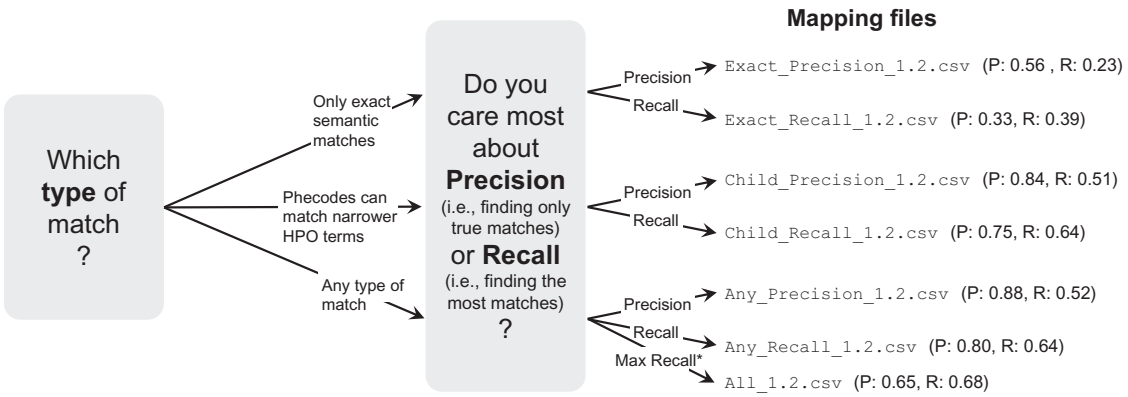

Figure 2. Subsets of HPO–phecode links provide flexibility for diverse applications. This flowchart guides selection of the appropriate HPO–phecode mapping for an application. First, some mappings prioritize only exact semantic matches versus matches that include the children HPO terms (eg, an anemia phecode would link to HPO term iron-deficiency anemia) versus all identified relationships. Second, mappings can prioritize precision or recall. The evidence types used in each of these mappings are described in Supplementary Table S1. For each map, evidence types were selected by identifying the combination of evidence that resulted in the maximal F0.5 score (to prioritize precision [P]) or F1 score (to prioritize recall [R]). The precision and recall are based on the manual review described in the Materials and Methods (Supplementary Figure S8). The "Max Recall" mappings (indicated by *) include links from any evidence. Note that these have the lowest precision and include up to 35% false positive links; therefore, we recommend evidence filters for most uses. A flowchart and evidence types used for Phecode X are described in Supplementary Figure S9 and Table S2.

facilitated mappings (10%–15%) and PheMap links (14%). Considering all links from any evidence, the overall precision is 80.3%. We detail examples of incorrect links via inaccurate evidence and related terms or symptoms in Supplementary Text.

CONCLUSION

In summary, we present and evaluate a phecode–HPO map to enable the translation of insights across a spectrum of diseases and data sources. To enable different use cases, we annotate each phecode–HPO link with different levels of evidence, so that users can tailor the mapping to their use. This will enable a broad range of applications—from PheRS to identify patients likely to benefit from genetic testing to PheWAS to explore molecular mechanisms of disease at a population-scale in the EHR.[13,21] Our map helps build a foundation for bridging genome biology and medicine by enabling connection to existing large disease biology networks linking genes, phenotypes, cross-species diseases, and model organisms (eg, Monarch Initiative[22,23]) and it broadens the scope of tools or resources that use the language of HPO (eg, Phenomizer,[24] Exomiser,[25] SimulConsult,[26] Matchmaker Exchange,[27] PhenoTips,[28] GeneNetwork Assisted Diagnostic Optimization [GADO],[29] and GWAS Central[30]).[31]

In evaluating the phecode–HPO links, we find that some evidence is more precise, while other types of evidence provide higher recall (Figure 1, Supplementary Figures S5–S7). Direct string and sub-string matching, SORTA candidate matches with a high similarity score, and strong PheMap links are the most high-quality and semantically exact. We define these as "strong specific evidence." In addition to these, we find that indirect string, substring, and UMLS matches provide strong evidence, although they do not necessarily indicate a match at the same level of granularity; thus, we define these as "strong broad evidence."

Finally, we explored precision–recall space for all possible combinations of evidence types to produce data-driven recommendations for mappings depending on the research scenario (Supplementary Figure S8, Tables S1 and S2). Figure 2 provides a flowchart to identify the suggested mapping based on the preference for the type of match and prioritization of precision versus recall. We make these filtered maps available at the phecode–HPO-map Github and Dryad repositories.[32]

While this mapping is not comprehensive or without incorrect links, it represents significant advantages over manual curation in efficiency, avoiding user error, and minimizing potential bias from curators' preferences. Despite challenges in mapping across vocabularies with different goals and granularities, this mapping provides an accessible framework for both future research and mappings. Achieving the promise of precision medicine requires integrating knowledge across diverse domains from molecular mechanisms to clinical practice. Ultimately, as more complex and rich healthcare data become available it will be paramount to unite previously siloed knowledge using flexible computational solutions.

FUNDING

This work was supported by the National Institutes of Health (NIH) General Medical Sciences award R35GM127087 to JAC, NIH National Human Genome Research Institute award F30HG011200 to EM, National Library of Medicine R01LM010685 to LB, and T32GM007347.

AUTHOR CONTRIBUTIONS

Conceptualization, Methodology: EM, LB, and JAC; Validation, Investigation, Data Curation: EM and LB; Formal Analysis, Visualization, Writing—Original Draft: EM; Writing—Reviewing & Editing: EM, LB, and JAC; Supervision: JAC.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

ACKNOWLEDGMENTS

The authors would like to thank Patrick Wu and members of the Capra lab for helpful discussions and manuscript comments. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The publicly available data used for analysis are available in the following repositories: Phecodes and their ICD9/10 mappings (https://phewascatalog.org/phecodes)[5,6]; HPO codes (https://hpo.jax.org/app/data/ontology) v2022-04-14[2]; 2022AA UMLS (files `MRCONSO.RRF`, `MRREL.RRF`) (https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html)[17]; and PheMap knowledge base (https://www.vumc.org/cpm/phemap) (`PheMap_UMLS_Concepts_1.1.csv`).[19]

All final mappings described in Figure 2, Supplementary Figure S9, and all other data we generated are available in the Dryad digital repository[32] and GitHub repository "phecode-HPO-map" (https://github.com/emcarthur/phecode-HPO-map). This includes intermediate tables for each source of evidence and maps that were manually curated and reviewed. It also includes intermediate mappings to ICD-9 and ICD-10. It contains pre-filtered maps based on the "strong specific" and "strong broad" evidence. We include a table that also has pre-calculated precision and recall for all combinations of evidence. Finally, we provide maps formatted for use with the PheRS R package.[33]

CODE AVAILABILITY

The publicly available code and software for analysis are available in the following repositories: SORTA phenotype mapping software on the Molgenis Cloud (https://sorta.molgeniscloud.org/menu/main/home)[18,34]; WikiMedMap code that we modified for our mapping (https://github.com/Linasulieman/WikiMedMap/).[20]

The custom code we generated are available in the GitHub repository "phecode-HPO-map" (https://github.com/emcarthur/phecode-HPO-map). This includes all code used to generate, evaluate, and visualize the map.

REFERENCES

1. Katsanis N. The continuum of causality in human genetic disorders. *Genome Biol* 2016; 17 (1): 233.
2. Köhler S, Gargano M, Matentzoglu N, *et al.* The human phenotype ontology in 2021. *Nucleic Acids Res* 2021; 49 (D1): D1207–D1217.
3. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005; 33 (Database Issue): D514–D517.
4. *Online Mendelian Inheritance in Man, OMIM®*. Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University; 2022.
5. Wei WQ, Bastarache LA, Carroll RJ, *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017; 12 (7): e0175508.
6. Wu P, Gifford A, Meng X, *et al.* Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Med Inform* 2019; 7 (4): e14325.
7. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci* 2021; 4: 1–19.
8. Zhang XA, Yates A, Vasilevsky N, *et al.* Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *NPJ Digit Med* 2019; 2 (1): 32.
9. Kafkas S, Althubaiti S, Gkoutos GV, Hoehndorf R, Schofield PN. Linking common human diseases to their phenotypes; development of a resource for human phenomics. *J Biomed Semant* 2021; 12 (1): 1–15.
10. Winnenburg R, Bodenreider O. Coverage of phenotypes in standard terminologies. In: *Proceedings of the Joint Bio-Ontologies and BioLINK ISMB'2014 SIG session "Phenotype Day"*; 2014: 41–44; Boston, MA.
11. Dhombres F, Winnenburg R, Case JT, Bodenreider O. Extending the coverage of phenotypes in SNOMED CT through post-coordination. *Stud Health Technol Inform* 2015; 216: 795–799.
12. Dhombres F, Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies—investigating partial mappings between HPO and SNOMED CT. *J Biomed Semantics* 2016; 7 (1): 3.
13. Bastarache L, Hughey JJ, Hebbring S, *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* 2018; 359 (6381): 1233–1239.
14. Bastarache L, Hughey JJ, Goldstein JA, *et al.* Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J Am Med Inform Assoc* 2019; 26 (12): 1437–1447.
15. Ganesan S, Galer PD, Helbig KL, *et al.* A longitudinal footprint of genetic epilepsies using automated electronic medical record interpretation. *Genet Med* 2020; 22 (12): 2060–2070.
16. Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenomewide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013; 31 (12): 1102–1110.
17. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32 (Database issue): D267–D270.
18. Pang C, Sollie A, Sijtsma A, *et al.* SORTA: a system for ontology-based recoding and technical annotation of biomedical phenotype data. *Database* 2015; 2015: bav089.
19. Zheng NS, Feng Q, Eric Kerchberger V, *et al.* PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. *J Am Med Inform Assoc* 2020; 27 (11): 1675. 1687.
20. Sulieman L, Wu P, Denny JC, Bastarache L. WikiMedMap: expanding the phenotyping mapping toolbox using wikipedia. *bioRxiv* 2019: 727792. doi:10.1101/727792.
21. Denny JC, Bastarache L, Roden DM. Phenome-wide association studies as a tool to advance precision medicine. *Annu Rev Genomics Hum Genet* 2016; 17: 353–373.
22. Mungall CJ, McMurry JA, Kohler S, *et al.* The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res* 2017; 45 (D1): D712–D722.
23. McMurry JA, Köhler S, Washington NL, *et al.* Navigating the phenotype frontier: the monarch initiative. *Genetics* 2016; 203 (4): 1491–1495.
24. Köhler S, Schulz MH, Krawitz P, *et al.* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* 2009; 85 (4): 457–464.
25. Smedley D, Jacobsen JO, Jäger M, *et al.* Next-generation diagnostics and diseasegene discovery with the Exomiser. *Nat Protoc* 2015; 10 (12): 2004–2015.
26. Fuller G. Simulconsult: www.simulconsult.com. *J Neurol Neurosurg Psychiatry* 2005; 76 (10): 1439.
27. Philippakis AA, Azzariti DR, Beltran S, *et al.* The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat* 2015; 36 (10): 915–921.
28. Girdea M, Dumitriu S, Fiume M, *et al.* PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat* 2013; 34 (8): 1057–1065.
29. Deelen P, van Dam S, Herkert JC, *et al.* Improving the diagnostic yield of exomesequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nat Commun* 2019; 10 (1): 1–13.
30. Beck T, Rowlands T, Shorter T, Brookes AJ. GWAS Central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies. *Nucleic Acids Res* 2022; 51 (D1): D986–D993.

31. Köhler S, Oien NC, Buske OJ, *et al*. Encoding clinical data with the human phenotype ontology for computational differential diagnostics. *Curr Protoc Hum Genet* 2019; 103 (1): e92.
32. McArthur E, Capra JA. Mapping between Human Phenotype Ontology and phecode terminologies. *Dryad Digital Repository, Dataset* 2023. https://doi.org/10.7272/Q6H70D20.
33. Aref L, Bastarache L, Hughey JJ. The phers R package: using phenotype risk scores based on electronic health records to study Mendelian disease and rare genetic variants. *Bioinformatics* 2022; 38 (21): 4972–4974.
34. Van Der Velde KJ, Imhann F, Charbon B, *et al*. MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. *Bioinformatics* 2019; 35 (6): 1076–1078.