The Mouseion at the JAXlibrary

1-8-2021

The Human Phenotype Ontology in 2021.

Sebastian Köhler
Michael Gargano
Nicolas Matentzoglu
Leigh Carmody
David Lewis-Smith

See next page for additional authors

Follow this and additional works at: https://mouseion.jax.org/stfb2021

Part of the Life Sciences Commons, and the Medicine and Health Sciences Commons
Authors
Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynarn, Amy M Brower, Tiffany J Callahan, Christopher G Chute, Johanna L Est, Peter D Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L Harris, Michael J Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A McMurry, Jillian A Miller, Monica C Munoz-Torres, Rebecca L Peters, Christina K Rapp, Ana M Rath, Shahmir A Rind, Avi Z Rosenberg, Michael M Segal, Markus G Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J Mungall, Melissa A Haendel, and Peter N Robinson
The Human Phenotype Ontology in 2021

Sebastian Köhler¹,², Michael Gargano²,³, Nicolas Matentzoglu²,⁴,⁵, Leigh C. Carmody²,³, David Lewis-Smith⁶,⁷, Nicole A. Vasilovsky²,⁸, Daniel Danis, Ganna Balagura⁹,¹⁰, Gareth Baynam¹¹,¹², Amy M. Brower¹³, Tiffany J. Callahan¹⁴, Christopher G. Chute¹⁵, Johanna L. Est¹⁶, Peter D. Galer¹⁷,¹⁸, Shiva Ganesan¹⁷,¹⁸, Matthias Griese¹⁶,¹⁹, Matthias Haimel²⁰,²¹, Julia Pazmandi²⁰,²¹,²², Marc Hanauer²³, Nomi L. Harris²,²⁴, Michael J. Hartnett¹³, Maximilian Hastreiter¹⁶, Fabian Hauck¹⁶,²⁵, Yongqun He²⁶, Tim Jeske²⁶, Hugh Kearney²⁷, Gerhard Kindle²⁸,²⁹, Christoph Klein¹⁶, Katrin Knoflach¹⁶,¹⁹, Roland Krause³⁰, David Lagorce²³, Julie A. McMurry²,³¹, Jillian A. Miller¹³, Monica C. Munoz-Torres²,³¹, Rebecca L. Peters¹³, Christina K. Rapp¹⁶,¹⁹, Ana M. Rath²³, Shahmir A. Rind³²,³³, Avi Z. Rosenberg³⁴, Michael M. Segal³⁵, Markus G. Seidel³⁶, Damian Smedley³⁷, Tomer Talmi³⁸,³⁹, Yariyau Thomas⁴⁰, Samuel A. Wiafe⁴¹, Julie Xian¹⁷,⁴², Zafer Yüksel⁴³, Ingo Helbig⁴⁴,⁴⁵, Christopher J. Mungall¹²,²⁴, Melissa A. Haendel²,⁸,³¹ and Peter N. Robinson²,³,²²,*

¹Ada Health GmbH, Berlin, Germany, ²Monarch Initiative, ³The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA, ⁴Semanticly Ltd, London, UK, ⁵European Bioinformatics Institute (EMBL-EBI), ⁶Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK, ⁷Clinical Neurosciences, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK, ⁸Oregon Clinical & Translational Research Institute, Oregon Health & Science University, ⁹Department of Neurosciences, Rehabilitation, Ophthalmology, Genetics, and Maternal and Child Health, University of Genoa, Genoa, Italy, ¹⁰Pediatric Neurology and Muscular Diseases Unit, IRCCS ‘G. Gaslini’ Institute, Genoa, Italy, ¹¹Western Australian Register of Developmental Anomalies, King Edward memorial Hospital, Perth, Australia, ¹²Telethon Kids Institute and the Division of Paediatrics, Faculty of Helath and Medical Sciences, University of Western Australia, Perth, Australia, ¹³American College of Medical Genetics and Genomics (ACMG), Bethesda, MD, USA, ¹⁴Computational Bioscience Program, University of Colorado Anschutz Medical Campus, Colorado, USA, ¹⁵Johns Hopkins University Schools of Medicine, Public Health, and Nursing, ¹⁶Department of Pediatrics, Dr. von Hauner Children’s Hospital, University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany, ¹⁷Division of Neurology, Children’s Hospital of Philadelphia, Philadelphia, PA, USA, ¹⁸Department of Biomedical and Health Informatics (DBHi), Children’s Hospital of Philadelphia, Philadelphia, PA, USA, ¹⁹Ludwig-Maximilians University, German Center for Lung Research (DZL), Munich, Germany, ²⁰Ludwig Boltzmann Institute for Rare and Undiagnosed Diseases, Vienna, Austria, ²¹CoMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria, ²²Institute for Systems Genomics, University of Connecticut, Farmington, CT 06032, USA, ²³INSERM, US14—Orphanet, Plateforme Maladies Rares, Paris, France, ²⁴Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley CA, USA, ²⁵German Centre for Infection Research (DZIF), Munich, Germany, ²⁶Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA, ²⁷FutureNeuro, SFI Research Centre for Chronic and Rare Neurological Diseases, Ireland, ²⁸Institute for Immunodeficiency, Center for Chronic Immunodeficiency (CCI), Faculty of Medicine, Medical Center - University of Freiburg, Freiburg, Germany, ²⁹Centre for Biobanking FREEZE, Faculty of Medicine, Medical Center - University of Freiburg, Freiburg, Germany, ³⁰Luxembourg Centre for Systems Biomedicine, University of Luxembourg, L-4367 Belvaux, Luxembourg, ³¹Translational and Integrative Sciences Center, Department of Environmental and Molecular Toxicology, Oregon State University, OR, USA, ³²WA Register of Developmental Anomalies, ³³Curtin University, Western Australia, Australia, ³⁴Division of Kidney-Urologic Pathology, Johns Hopkins University, Baltimore, MD 21205, USA,

*To whom correspondence should be addressed. Tel: +1 860 837 2095; Email: peter.robinson@jax.org

© The Author(s) 2020. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
The Human Phenotype Ontology (HPO, https://hpo.jax.org) was launched in 2008 to provide a comprehensive logical standard to describe and computationally analyze phenotypic abnormalities found in human disease. The HPO is now a worldwide standard for phenotype exchange. The HPO has grown steadily since its inception due to considerable contributions from clinical experts and researchers from a diverse range of disciplines. Here, we present recent major extensions of the HPO for neurology, nephrology, immunology, pulmonology, newborn screening, and other areas. For example, the seizure subontology now reflects the International League Against Epilepsy (ILAE) guidelines and these enhancements have already shown clinical validity. We present new efforts to harmonize computational definitions of phenotypic abnormalities across the HPO and multiple phenotype ontologies used for animal models of disease. These efforts will benefit software such as Exomiser by improving the accuracy and scope of cross-species phenotype matching. The computational modeling strategy used by the HPO to define disease entities and phenotypic features and distinguish between them is explained in detail. We also report on recent efforts to translate the HPO into indigenous languages. Finally, we summarize recent advances in the use of HPO in electronic health record systems.

INTRODUCTION

The Human Phenotype Ontology (HPO) is a comprehensive resource that systematically defines and logically organizes human phenotypes. As an ontology, HPO enables computational inference and sophisticated algorithms that support combined genomic and phenotypic analyses. Broad clinical, translational and research applications using the HPO include genomic interpretation for diagnostics, gene-disease discovery, mechanism discovery and cohort analytics, all of which assist in realizing precision medicine.
Figure 1. HPO terms organized by organ system. (A) Counts for top-level phenotype terms (direct descendants of Phenotypic abnormality (HP:0000118) are shown. Counts of terms added to the ontology after the previous article in this series (19) are shown in dark blue (added between 25 July 2018 and 18 August 2020). (B) Examples of new terms added 2018–2020 and their parent terms, for selected organ systems. (C) An example text definition and synonyms for a new term.

The HPO provides annotations to diseases defined by Online Mendelian Inheritance in Man (OMIM) (17), nearly all of which are monogenic (Mendelian) diseases. Currently, 93 885 of a total of 108 580 such annotations were derived from mining the Clinical Synopsis section of the corresponding entry. 14 695 (13.5%) annotations were produced by curation by the HPO team and often contain additional information such as age of onset, affected sex, clinical modifiers, or overall frequency of the feature. A total of 7801 diseases are annotated in this way, corresponding to 108 580 annotations in all (with a mean of 13.9 annotations per disease). 296 curated annotations to 47 chromosomal diseases identified by DECIPHER (18) accessions were also generated by the HPO team (mean 6.2 annotations per disease).

In parallel, Orphanet uses the HPO to annotate rare diseases and has continued to develop annotations to a broad range of diseases (currently 96 612 annotations utilizing 7495 distinct HPO terms for 3956 diseases, with an average of 24.4 terms per disease). These annotations include information about the frequency (obligatory, very frequent, frequent, occasional, very rare or excluded) and whether the annotated HPO term is a major diagnostic criterion or a pathognomonic sign of the rare disease. These data are available at Orphadata.org and in the HPO-Orphanet Rare Disease Ontology (ORDO) ontological module called HOOM (See Data Availability section, below). While some of the annotated diseases overlap, Orphanet contains information about non-Mendelian rare diseases and defines diseases primarily based on clinical criteria, thereby providing a complementary resource. Both sets of annotations are available in a combined annotation file available on the HPO website. Figure 2 displays the growth in annotations to the OMIM entries.

Abnormal phenotypic features or manifestations of human disease stored in HPO are also employed for medical research projects such as SOLVE-RD. Funded by the European Commission, SOLVE-RD aims to solve large numbers of rare diseases for which a molecular cause is not known.

The HPO has a sophisticated quality control pipeline. In addition to custom software, we make extensive use of the quality control checks implemented in ROBOT (‘ROBOT is an OBO Tool’) (47). We have added descriptions of our quality control processes to the HPO website under the Help menu.

COMMUNITY COLLABORATIONS TO EXTEND THE COVERAGE OF HPO

The UK’s National Institute for Health Research (NIHR) Rare Disease initiatives extensively use the HPO in their RD-TRC (Rare Disease-Translational Research Collaboration) and NIHR BioResource, in wide-ranging studies. Following an HPO workshop with members of the NIHR-RD-TRC in 2017, the NIHR-RD-TRC assessed the maturity of the HPO across different disease areas and organ systems. Disorders of the immune system, central nervous system, the respiratory system, and the kidney were among the areas where additional work was deemed desirable (3).
Epilepsy

The epilepsies are a group of diverse disorders that share a predisposition to seizures (20). They are phenotypically complex with constellations of clinical features indicating different age-specific syndromes, broad epilepsy types, and etiologies that guide clinical management (21). We have recently demonstrated that phenotypic similarity approaches based on HPO-related phenotypes in the epilepsies can be used to identify novel genetic etiologies such as AP2M1 (22), to map the natural history of genetic epilepsies over time from electronic medical records (23), and to identify patterns of gene-phenotype associations (Figure 3) (24).

An important challenge in seizure classification is that seizures are paroxysmal, and often incompletely characterized or observed. In order to maximize the available information, the revised subontology includes terms independent of some of the dimensions of seizure description. For example, the terms Focal aware seizure (HP:0002349) and Focal motor seizure (HP:0011153) allow a true instance of Focal aware motor seizure (HP:0020217) to be coded as precisely as possible when knowledge of either the initial manifestation or the preservation of awareness is unknown. These concepts provide a way to categorize high-level, incomplete information that often makes disease classification difficult. Where possible, pre-existing terms were retained for the benefit of legacy HPO data. A few inconsistencies with contemporary seizure concepts were identified and corrected, such as the previous relationship of Bilateral tonic-clonic seizure with focal onset (HP:0007334) as a type of Generalized-onset seizure (HP:0002197) rather than Focal-onset seizure (HP:0007359). The new seizure subontology currently contains 347 terms, which significantly increases the detail with which seizures can be described (Figure 4).

Inborn errors of immunity (IEI)

Inborn errors of immunity (IEI), previously referred to as primary immunodeficiencies (PID), involve a variable, disorder-specific predisposition towards infections, immune dysregulation (including autoimmunity, autoinflammation,
Figure 3. HPO-based analyses demonstrate the clinical features associated with diagnostic variants in SCN1A in published cohorts with developmental and epileptic encephalopathies of various known, or unknown but presumed genetic, etiologies. Fisher’s exact test p-value for each term indicates the significance of the association between the HPO term and the presence of a diagnostic SCN1A variant in the cohort. (A) The frequency of HPO terms in SCN1A variant carriers versus non-carriers regardless of age. (B) The same data presented to demonstrate the conceptual relationships between associated features within the structure of the HPO. (A) and (B) modified from (24) with only a selection of terms labeled for legibility.

Granuloma formation, lymphoproliferation, etc.), and malignancies. Phenotypes of IEI are often complex, making it difficult to distinguish primary disease-specific features from secondary unspecific, infection- or inflammation-related, or merely randomly occurring clinical manifestations. However, unequivocal phenotypic descriptions are needed for semantic interoperability to enable the use of defining, cross-referencing, and/or filtering algorithms during the process of diagnosing these rare diseases. For the purpose of data verification of entries into the large international registry of the European Society for Immunodeficiencies (ESID) that includes data from >30 000 patients, either a known genetic diagnosis or the fulfillment of working definitions for the clinical diagnosis of IEI is required. Together with a group of international collaborators, the ESID registry working group designed a comprehensive list of obligatory and optional criteria for 92 entities that lack a genetic diagnosis (e.g. common variable immunodeficiency) that were cross-validated by other experts in a two-phase process (35). To enhance the catalog of clinical working definitions of IEI, we recently added HPO terms and the frequencies of phenotypes observed, derived from HOOM. For most other IIEIs that are included in the genotypic classification of the International Union of Immunological Societies (36), complete HPO term annotations are still lacking. To improve the available vocabulary and annotated diseases, a targeted expansion of IEI relevant HPO terms and re-annotation of currently known IEIs was launched by representatives of the ESID genetics working party and of ERN-RITA (European Network on Rare Primary Immunodeficiency, Autoinflammatory and Autoimmune diseases) with input from the International Society of Systemic Autoinflammatory Diseases (ISSAID) in 2018. The systematic review involved expert clinicians, geneticists, researchers (working on IEI) and bioinformaticians combining an ontology-guided machine-learning approach (37) with expert clinical immunologists’ reviews (M. Haimel, et al., manuscript in preparation). The HPO-classification of IEI is part of The Medical Informatics Initiative Germany (MII) founded by the Federal Ministry of Education and Research, which has launched the Collaboration on Rare Diseases (CORD) project. Aided by the national TRANSLATE-NAMSE project, this initiative plays a key role in the development of digitalized patient data allowing clinicians and scientists to make use of stan-
Newborn screening technologies or genetic sequencing (42). A range of organisms including Drosophila, Dictyostelium discoideum, Drosophila, fission yeast, planarian, Xenopus, mammals (MP) and zebrafish (ZP), as well ontologies for glycophenotypes (47) and pathogen–host interactions. The goal is to enable meaningful and reliable mapping of phenotype data such as gene-to-phenotype associations across databases that are specific to particular modalities or organisms, and leverage this data for a variety of important applications including clinical diagnosis and variant prioritization. For example, Exomiser (15) leverages the semantic associations between HPO, MP and ZP to prioritize variants effectively by matching human phenotypic abnormalities with phenotypes observed in animal models with knockouts of genes orthologous to human disease-associated genes.

Figure 6 illustrates the extent to which phenotype ontologies adhere to phenotype DOSDP patterns (‘uPheno out the world. In the United States, the Newborn Screening Translational Research Network (NBSTRN) provides tools and resources to researchers working to discover novel screening technologies and interventions (43). An important goal for the NBSTRN is to understand health outcomes and the natural history of rare diseases by capturing longitudinal genomic and phenotypic information on the estimated 22,000 infants diagnosed through newborn screening (NBS) each year. A US federal advisory committee recommends conditions for NBS resulting in the Recommended Uniform Screening Panel, and in 2018, screening for Spinal Muscular Atrophy was endorsed. As a case study of HPO in NBS and rare disease, a REDCap™ data dictionary of 4757 data elements in the SPOT SMA Longitudinal Pediatric Data Resource was reviewed to identify existing terms and suggest new terms. The aim of this effort is to develop HPO as a resource for the longitudinal follow-up of NBS identified individuals with the goal of advancing understanding of rare disease.

Kidney Precision Medicine Project (KPMP)

The Kidney Precision Medicine Project (KPMP) aims to understand and find ways to treat chronic kidney disease (CKD) and acute kidney injury (AKI). KPMP has contributed over 100 kidney-related phenotype terms; clinical nephrologists, pathologists and ontologists worked together over multiple workshops to propose new terms and modifications to HPO and underlying ontologies such as Uberon (38). Two new major HPO branches were generated, one focusing on pathology-related terms, and the other on clinical phenotype terms (Figure 5).

Pulmonology

The category of respiratory disorders is not only underrepresented in the HPO; it is rapidly expanding with the ongoing molecular definition of rare to ultra-rare novel diseases. Therefore, substantial effort was undertaken to improve the foundation and formulation of terms and disease associations. However, gaps remain—for example, for most rare and common pulmonary disorders included in the current classification of children's interstitial lung diseases (40), comprehensive HPO term annotations still need to be completed. To this end, representatives of the European research collaboration for Children's Interstitial Lung Disease (chILD-EU) consortium have called for community participation and initiated a low barrier approach to facilitate contribution to the HPO for newcomers (see section on contributing to the HPO in the Data Availability section, below). To facilitate sharing knowledge about rare respiratory disorders, information is collected in international registers like the Kids Lung Register, operating through the chILD-EU management platform. The chILD-EU network utilizes the HPO, which significantly improved the categorization of novel diseases and the annotation of cases included for long term investigation (41).

Pharmacogenomics

HPO has introduced several terms to describe drug response phenotypes. The new terms added to HPO are branched under the term Abnormal drug response (HP:00020169) and aim to encompass a spectrum of clinical phenotypes with regards to drug metabolism. The underlying HPO terms refer to abnormal blood concentration of drugs, altered efficacy and adverse drug response. As pharmacogenomic research makes its way into routine clinical applications, such terms may be valuable in describing variance in drug metabolism as ascertained by laboratory investigation or genetic sequencing (42).

Newborn screening

Screening of newborns to facilitate the early identification, diagnosis and treatment of rare diseases occurs through standardized phenotypic patient information. Digital recording of HPO terms will facilitate genetic research to identify disease-causing variants; it will also support large-scale studies aiming to associate genetic variance with a plethora of risks that can disrupt immune homeostasis.

Interoperability with other phenotype ontologies

We have developed templated ontology design patterns to structure OWL definitions, encoded as Dead Simple OWL Design Patterns (DOSDPs) (44). DOSDPs provide a number of advantages, including standardized patterns for the logical definitions and automatic classification. As coordinators of the Phenotype Ontologies Reconciliation Effort (45, 46), HPO developers contributed to the definition of 207 DOSDP templates for the consistent definition of phenotypes across species and modalities (44). The Unified Phenotype Ontology (uPheno) integrates multiple phenotype ontologies into a harmonized cross-species phenotype ontology. uPheno enables the comparison and grouping of species-specific phenotypes under species-neutral categories, and links phenotypes from one species with comparable phenotypes from other species. Using templates generates phenotype terms that are not only consistently structured, but also enriched with associations to, for example, biological processes (Gene Ontology), anatomical entities, and molecular entities. For example, an abnormal level of chemical entity with role in location provides a template for terms such as Abnormal circulating hormone level (HP:0003117). Reconciliation is ongoing and is improving the alignment between phenotype ontologies for a range of organisms including C. elegans, Dictyostelium discoideum, Drosophila, fission yeast, planarian, Xenopus, mammals (MP) and zebrafish (ZP), as well ontologies for glycophenotypes (47) and pathogen–host interactions. The goal is to enable meaningful and reliable mapping of phenotype data such as gene-to-phenotype associations across databases that are specific to particular modalities or organisms, and leverage this data for a variety of important applications including clinical diagnosis and variant prioritization. For example, Exomiser (15) leverages the semantic associations between HPO, MP and ZP to prioritize variants effectively by matching human phenotypic abnormalities with phenotypes observed in animal models with knockouts of genes orthologous to human disease-associated genes.
**Figure 5.** One major goal of KPMP is to refine classification of kidney diseases in molecular, cellular, and phenotypic terms and thereby identify novel targeted therapies. The kidney-related HPO terms are being used in multiple ways in KPMP. For example, KPMP has used the HPO terms for clinical and pathological phenotype annotations, integrative Kidney Tissue Atlas Ontology (KTAO) (39) development, and systematic data integration software development.

**Figure 6.** Proportions of terms defined in the HPO, the Mammalian Phenotype Ontology (MP) (48), the Drosophila Phenotype Ontology (DPO) (49), the Worm Phenotype Ontology (WPO) (50), the Xenopus Phenotype Ontology (XPO) (51) and the Zebrafish Phenotype Ontology (ZP) (52).

conformant’). Currently, the HPO has 6154 OWL-defined terms (41% of the total number of 15 029 terms), out of which 4139 (67%) adhere to an existing template. While some phenotypes may be too complex to define using a general template, we hope to increase our coverage to ~50% of the terms.

**Indigenous languages**

For equity and scale of precision medicine and precision public health, it is critical to advance methods to improve the diagnosis and treatment of rare diseases. Communication is critical to healthcare and methods to deliver and incorporate translations, community narratives and family-based approaches are important to advancing culturally appropriate care. Lyfe Languages (lyfelanguages.com) is improving communication between indigenous patients, families, and medical professionals, in part by delivering indigenous language translations of the HPO. This started with a focus on rare diseases, then expanded to also include COVID-19 and is being extended into mental health. Currently, HPO terms are being translated to 11 Australian Aboriginal and Torres Strait Islander Languages and 6 Ghanian indigenous languages. The latter project is being performed together with the Rare Disease Ghana Initiative.

**HPO for medical education & crowdsourcing**

One of the advantages of the structured knowledge contained in the HPO is that it can be utilized as a teaching tool. One recent example of using HPO in this way is Phenotate, a portal that allows the annotation of OMIM and Orphanet disorders with HPO terms to be formulated as assignments for students (53). Phenotate has been used in five undergraduate courses, allowing for the collection of annotations for 22 diseases, including six where previously structured annotations were not available. Interestingly, the annotations generated by Phenotate, while sourced from untrained undergraduate students, were equal to curated gold standards in terms of allowing clinicians to identify rare disorders.

**EHR INTEGRATION**

Electronic health records (EHRs) have been widely adopted and offer an unprecedented opportunity to accelerate translational research because of advantages of scale and cost-efficiency as compared to traditional cohort-based studies. Textual data within EHRs can describe phenotypic features that are not encoded within the structured fields of the EHR, but natural language processing (NLP) is required to transform such data into terminological entities (ontology terms) for downstream analysis. NLP of phenotypic data is becoming a mature field that can be used to improve clinical care, and HPO has been used by a number of groups as a resource for EHR analysis (54). For example, EHRs spanning individuals’ entire childhoods can be mapped to the HPO, yielding longitudinal patterns of phenotypic features associated with particular genetic etiologies (Figure 7) (23).
Figure 7. Analysis of time-stamped EHRs of children with epilepsy demonstrates the association of HPO terms with diagnostic SCN1A variants at different ages (modified from (23) with only a selection of terms labeled for legibility).

However, EHR data are often incomplete or incorrect, and EHR systems are generally billing instruments rather than tools to improve patient care, much less allow secondary research.

LOINC (Logical Observations Identifiers, Names, Codes) is a clinical terminology for laboratory test orders and results that is widely used in EHRs (55). We developed a mapping strategy (LOINC2HPO) to transform laboratory data in EHR records to HPO terms. For instance, if the result of the test LOINC:6298-4 (potassium in blood) is above normal limits, our library would call the HPO term Hyperkalemia (HP:0002153). Many common tests in medicine can be performed in multiple ways, so there can be multiple LOINC codes for tests that measure the same biological quantity. For instance, currently, there are four different LOINC terms for different tests of urine nitrite. Our library maps these terms to the same HPO term. Additionally, the hierarchy of the HPO can be used to roll up related results (e.g. reduced concentrations of different B vitamins in the blood). In a pilot study, we investigated EHR data from 15 681 patients with respiratory complaints and identified known biomarkers for asthma (56). However, the absence of an ontological structure in LOINC, a known issue, impeded optimal information capture and coding. Members contributing to last year’s paper have secured funding to partner with the LOINC developer to address this challenge, which will enhance the community’s ability to categorize clinical laboratory findings into HPO terms.

The diagnostic decision support system SimulConsult uses a controlled list of 9871 findings chosen for their importance in diagnosis (12). As part of a project to use machine-assisted chart review to flag which of those findings are discussed in the EHR, hundreds of new findings were added to HPO in a collaboration between HPO and SimulConsult. Since HPO is one of the key inputs to the UMLS concept codes, adding terms to HPO is an efficient workflow for adding terms to UMLS as well.

Enabling large scale integration of biomedical knowledge with clinical patient data requires robust and accurate mappings between standardized clinical terminology concepts and ontologies, like the HPO. Existing work has demonstrated the power of the HPO to enrich clinical data including craniofacial and oral phenotypes (57), rare and Mendelian disease (58, 59), and infectious disease (60). There have also been more generalized mapping efforts aimed at aligning different clinical terminologies to the HPO including free-text narratives (61) and structured data like diagnosis codes (62, 63). While this work is very promising, it has largely been limited to specific clinical domains (i.e. only diagnosis codes from structured data or only phenotype mentions in free-text). Additionally, the vast majority of prior work focused on mapping clinical codes from standardized terminologies has exclusively focused on mapping only specific terminologies (e.g. SNOMED-CT or ICD-9). Mapping to a single terminology limits the generalizability of the mappings. One solution is to generate mappings to common data models (CDM) as well as tools that integrate different EHR data, such as Informatics for Integrating Biology and the Bedside (i2b2) (64) and Observational Health Data Sciences and Informatics’s Observational Medical Outcomes Partnership (OMOP) (65).

Currently, there exist no large-scale mappings spanning multiple clinical domains (e.g. diagnosis, medications, laboratory measurements) to the HPO and other biomedical ontologies. In collaboration with researchers from the University of Colorado Anschutz Medical Campus, a new framework, OMOP2OBO (66), is being developed to map several ontologies, including the HPO, to standardized clinical terminologies in the OMOP CDM. The mappings are generated using a combination of manual and automatic approaches and validated by a panel of clinical and biological domain experts. To date, the mappings cover over 29 000 diagnosis codes (over 20 000 diagnosis codes map to a total of over 4000 HPO codes), 1700 medication ingredients, and
over 11 000 laboratory test results including and extending current LOINC2HPO annotations.

The distinction between diseases and phenotypes
The community uses the word phenotype with multiple meanings. The HPO defines a disease as an entity that has all four of the following attributes:

- an etiology (whether identified or as yet unknown)
- a time course
- a set of phenotypic features
- if treatments exist, there is a characteristic response to them

A phenotype phenotypic feature is a part of a disease. The phenotype of an individual with a disease can be said to be the sum of all of the phenotypic features manifested by that individual. HPO terms can be used to describe the phenotypic features that occur in individuals with a disease. For instance, if the disease entity is the common cold, then the cause is a virus; the phenotypic features include fever, cough, runny nose, and fatigue; the time course usually is a relatively acute onset with manifestations dragging on for days to about a week; and the treatment may include bed rest, aspirin, or nasal sprays. In contrast, a phenotypic feature such as fever is a manifestation of many diseases. There is a grey zone between diseases and phenotypic features. For instance, diabetes mellitus can be conceptualized as a disease, but it is also a feature of other diseases such as Bardet Biedl syndrome. The HPO takes a practical stance and provides terms for such entities. In the future, the HPO will develop tighter integration with the Mondo Disease Ontology (67) in order to define this category of HPO terms based on the corresponding diseases. A related issue is the fact that phenotypic features are analyzed and reported at different levels of granularity. For instance, the evaluation of a liver biopsy in an individual with hepatitis C would usually involve an assessment of focal lobular necrosis, portal inflammation, piecemeal necrosis, and bridging necrosis, each of which could be classified into one of several levels, each of which would be specified in the pathology report. If the findings are sufficiently abnormal, the pathologist may make a diagnosis such as chronic hepatitis. For the purposes of precision medicine, it would be preferable to have all the information available in electronic form, but in many settings, not all of this information is available. The HPO takes a practical stance, providing terms at different levels of granularity; for example, Hepatic bridging fibrosis (HP:0012852) and Chronic hepatitis (HP:0200123).

CONCLUSION
The HPO has continued to benefit from the support of domain experts from multiple areas of clinical medicine. We will expand our work on extending the HPO terminology to several additional subontologies including those for behavioral abnormalities, various areas related to prenatal and perinatal medicine, as well as to common diseases. We are designing an online collaboration portal for domain experts to submit new disease annotations.

DATA AVAILABILITY
Human Phenotype Ontology: https://hpo.jax.org/. Files available for download include the main ontology file in OBO, OWL, and JSON formats (See Download/Ontology); the main HPOA file, genes_to_phenotype.txt and phenotype_to_genes.txt (See Download/Annotation).
- GitHub: https://github.com/obophenotype/human-phenotype-ontology
- Change logs: https://github.com/obophenotype/human-phenotype-ontology/tree/master/src/ontology/reports
- Instructions for contributing to the HPO are available at https://hpo.jax.org/app/help/collaboration
- chILD-EU management platform: (www.childevu.net)
- Collaboration on Rare Diseases (CORD): https://www.medizininformatik-initiative.de/en/CORD
- DOSDP: https://github.com/obophenotype/upheno/tree/master/src/patterns/dosdp-dev
- ESID registry https://esid.org/Working-Parties/Registry-Working-Party/Diagnosis-criteria
- Kidney Precision Medicine Project (KMP) https://kpmp.org/
- Lyfe languages: http://www.lyfelanguages.com/About.html
- Monarch Initiative: https://monarchinitiative.org/
- Newborn Screening Translational Research Network (NBSTRN): www.nbstrn.org
- NIH CDE Repository: https://cde.nlm.nih.gov/
- OMOP2OBO: https://github.com/callahantiff/OMOP2OBO
- Orphadata (including HOOM): http://www.orphadata.org/
- Orphanet: http://www.orpha.net
- ORPHApackets: https://github.com/Orphanet/orphapacket
- Rare Disease Ghana Initiative (https://www.rarediseaseghanai.org/)
- Zooma: https://www.ebi.ac.uk/spot/zooma/

FUNDING
Monarch R24 [2R24OD011883-05A1]; NHGRI Phenomics [1RM1HG010860]; NHGRI/NCI Forums in Phenomics [5U13CA221044]; Solve-RD [79725]; HIPBI [643578]; DFG [Gr 970/9-1]; E-Rare-3; HCQ4Surfdefect; Cost CA [16125 ENTe-r-chILD]. Funding for open access charge: NIH.
Conflict of interest statement. None declared.

REFERENCES