Deep phenotyping for patients by patients: a plain-language medical vocabulary for precision diagnosis

A major goal of human genetics and precision medicine is to understand the relationship between genetic variants and human diseases. The advent of whole-exome sequencing (WES) and genome sequencing (WGS) has substantially improved our ability to comprehensively characterize genetic variants. Although WES/WGS has led to the discovery of many novel disease-associated genes, the diagnostic yield in patients without a clear clinical diagnosis has been 11% to 25% (for example, Yang 2013).[1] With approximately 100,000 called variants in a typical exome or ~4.5 million variants in a typical genome, multiple candidate genes remain after a bioinformatic analysis and additional methods and data are needed to refine the list of candidates. The Human Phenotype Ontology (HPO) was created to enable "deep phenotyping" i.e., capture of symptoms and phenotypic findings using a logically-constructed hierarchy of phenotypic terms.[2] The HPO has become the *de facto* standard for representing clinical phenotype data to inform diagnoses for rare genetic diseases by the 100,000 Genomes project, the NIH Undiagnosed Diseases Program (UDP)/Network (UDN), and thousands of other clinics, labs, tools, and databases.[2,3] The computable phenotypic profiles (sets of terms) of individual patients allow "fuzzy" matching against known diseases and model organisms based on proximity in the hierarchy and weighting according to term specificity.[2,4,5] Compared to using WES/WGS data alone, we have shown that HPO-based phenotyping improves molecular diagnosis in the UDP by 10-20%. [6 and personal communication]

For undiagnosed and rare disease patients, the patients themselves are an especially critical source of phenotyping information. As these patients live with their condition, they also develop explicit and implicit knowledge about it, whether from multiple clinician evaluations or from other families and patients experiencing a diagnostic odyssey for similar conditions. From these interactions, they develop a lexicon of relevant terms; these terms are frequently in plain language, but can also include clinical terms. In exceptional cases, patients' self-phenotyping and investigations have led to clinical diagnosis, for example, Jill Viles, who despite skepticism from doctors, managed to not only diagnose herself and family members but also to reveal fundamental biology and improved management of autosomal dominant Emery-Dreifuss muscular dystrophy.[7] Further, some phenotypes are not readily observed clinically but can be documented by the patient or family; for instance, the Might family (*NGLY1*) observed that their baby did not produce tears when crying, and they used this feature to help identify other families with the same disease. However, the "terminology gap" between medical professionals and patients has hindered patients from finding appropriate information; the lay synonyms would be very useful to improve information retrieval for patients from the literature. The terminology gap has also limited patient participation in both research studies and clinical phenotyping.[8] Current consumer health vocabularies (CHVs) provide broad consumer equivalents for clinical findings, medical procedures, and equipment,[9] but are not well integrated with research terminologies and provide neither the structure nor the coverage required for translational research and diagnostic tools. While CHVs and clinical survey instruments can be mapped to HPO terms, they tend to provide only high level phenotype terms that are unsuitable for use in clinical genetics contexts.

In order to address these issues, we therefore sought to make the HPO capable of capturing patient-generated phenotypic profiles for use in diagnostic and patient community settings (registries, forums, clinics, and patient websites). To achieve this, we have systematically added patient-centered synonyms throughout the HPO by drawing on many sources and knowledge bases across a spectrum, including Wikipedia, MedlinePlus, Mayo Clinic, OMIM the Elements of Morphology, patient forums such as the WebMD message board, as well as other ontologies, terminologies, and specialty texts. We also aimed for consistency in synonym creation across terms, for example *Increased bone density in 2nd toe bone* and *Increased bone density in 3rd toe bone*, etc., were added as synonyms for *Sclerosis of the 3rd toe phalanx and Sclerosis of the 2nd toe phalanx, respectively*. The synonyms are classified as exact, broad, narrow, or related. Synonyms are classified as exact if they are precise alternatives to the HPO term, and can be used interchangeably in computational algorithms. For 40% of the terms there exists no reasonable non-clinical language, for example, *Anomalous origin of left coronary artery from the pulmonary artery* (HP:0011638). Other terms, for example phenotypes involving the radius and ulna cannot be succinctly distinguished in layperson terminology (e.g. *forearm bone*). We also recognized relationships within the ontology when adding layperson synonyms, for example, *Yellowing of the skin* was added to the HPO class, *Jaundice*, but also to subclasses, for example, *Intermittent jaundice* would have the lay synonym *Intermittent yellowing of the skin*.

We then evaluated the quality of the lay-HPO to ensure lay profiles could be effectively used by variant prioritization tools such as Exomiser[10]. We found that the major fraction of layperson synonyms were added to very specific terms (e.g. "clinodactyly" rather than "Abnormality of the digits"), suggesting that they would indeed be specific enough to be used diagnostically. In total, 36% of the HPO terms from the most recent release (2017-06-21) have at least one layperson synonym (4547 of 12,623). 89% of the diseases (8666 of 9657) in the HPO database have at least one HPO annotation with a layperson synonym and 60% of all disease annotations (73,932 of 122,120) are referring to HPO terms with lay translations. This further confirms the diagnostic utility of the synonyms despite incomplete coverage. This suggests that patient-generated HPO profiles will be diagnostically informative. Further, because the lay-translation of the HPO uses the same logical infrastructure as the HPO, patient-generated phenotyping data can be readily be combined with clinical phenotyping data in the context of variant prioritization to improve diagnostic rates as well as other analytics such as examining expressivity and penetrance, disease progression.

We envision the layperson HPO as a tool that will allow patients and families to become more effective partners in translational research, empowering families to achieve an accurate diagnosis, as well as providing opportunities for people to improve the lives of others by increasing medical knowledge through their personal journeys. Through a recently funded partnership with the Patient-Centered Outcomes Research Institute (PCORI), we will evaluate the use of the layperson HPO for genomic diagnostics in patient cohorts and enable development of new patient-friendly phenotyping tools. Given the great need to make rare disease phenotyping data available in a computationally useful and as open-as-possible, manner we urge clinical geneticists, rare disease patient communities, phenotyping tools, registries, and consumer testing labs to help evaluate, adopt, and contribute towards what we think is a critical resource for engaging patients in their own deep phenotyping. Finally, we

envision the layperson HPO will enable rare disease patients to share their phenotyping profiles openly on the web, even in places such as Facebook. This will allow the use of informatics to support open querying for similar patients to improve diagnosis and for cohort and community identification globally.

Funding: National Institutes of Health (NIH) Monarch Initiative [OD #5R24OD011883].

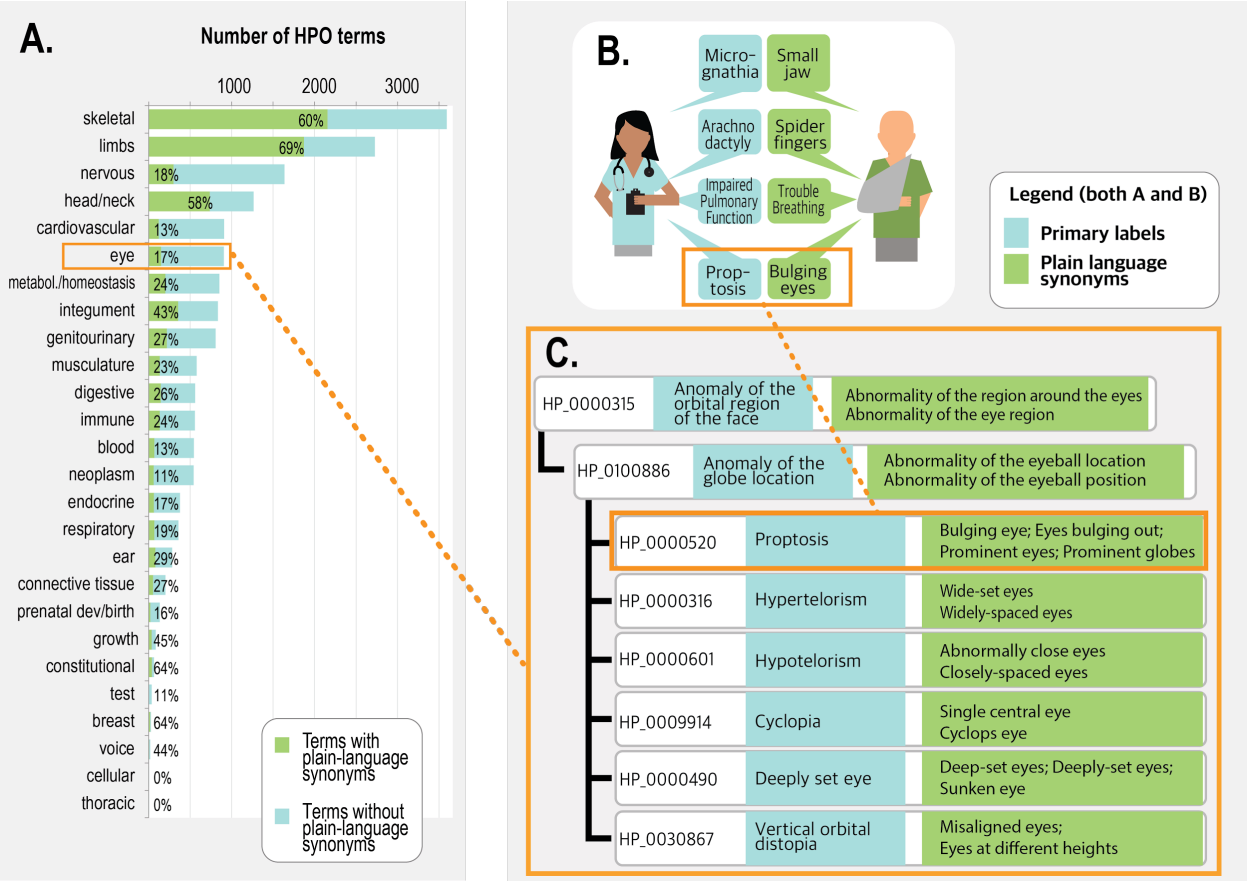

Figure 1. A) Coverage of HPO terms with plain language synonyms (terms broken down by anatomical system). B) A physician and a patient describe a patient's phenotype profile in different ways but with the same meaning. *This constellation of diverse phenotypes is common in Marfan Syndrome; each has a plain-language equivalent.* ***C) A sub-branch of eye phenotypes within the HPO.*** *Terms are structured rigorously, not only in terms of hierarchy (as shown) but also in terms of logical definitions (not shown).*

References

1. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013;369: 1502–1511. doi:10.1056/NEJMoa1306555

2. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human

Phenotype Ontology in 2017. Nucleic Acids Res. 2017;45: D865–D876. doi:10.1093/nar/gkw1039

3. Posey JE, Rosenfeld JA, James RA, Bainbridge M, Niu Z, Wang X, et al. Molecular diagnostic experience of whole-exome sequencing in adult patients. Genet Med. 2016;18: 678–685. doi:10.1038/gim.2015.142

4. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85: 457–464.

5. Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2016; doi:10.1093/nar/gkw1128

6. Gall T, Valkanas E, Bello C, Markello T, Adams C, Bone WP, et al. Defining Disease, Diagnosis, and Translational Medicine within a Homeostatic Perturbation Paradigm: The National Institutes of Health Undiagnosed Diseases Program Experience. Frontiers in Medicine. 2017;4: 62. doi:10.3389/fmed.2017.00062

7. Epstein D. The Muscular Dystrophy Patient and Olympic Medalist with the Same Genetic Disorder. In: ProPublica [Internet]. 15 Jan 2016 [cited 6 Jun 2017]. Available: https://www.propublica.org/article/muscular-dystrophy-patient-olympic-medalist-same-genetic-mutation

8. Park MS, He Z, Chen Z, Oh S, Bian J. Consumers' Use of UMLS Concepts on Social Media: Diabetes-Related Textual Data Analysis in Blog and Social Q&A Sites. JMIR Med Inform. 2016;4: e41. doi:10.2196/medinform.5748

9. Zielstorff RD. Controlled vocabularies for consumer health. J Biomed Inform. 2003;36: 326–333. doi:10.1016/j.jbi.2003.09.015

10. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. Nature Research; 2015;10: 2004–2015. doi:10.1038/nprot.2015.124