

Enhancing the Human Phenotype Ontology for Use by the Layperson

Nicole A. Vasilevsky, Mark E. Engelstad, Erin D. Foster, Melissa A. Haendel
Ontology Development Group, Library,
Oregon Health & Science University
Portland, OR, USA
vasilevs@ohsu.edu

Christopher J Mungall
Environmental Genomes and Systems Biology,
Lawrence Berkeley National Laboratory
Berkeley, CA, USA

Peter Robinson, Sebastian Köhler
Charité - Universitätsmedizin Berlin
Berlin, Germany
sebastian.koehler@charite.de

Abstract—In rare or undiagnosed diseases, physicians rely upon genotype and phenotype information in order to compare abnormalities to other known cases and to inform diagnoses. Patients are often the best sources of information about their symptoms and phenotypes. The Human Phenotype Ontology (HPO) contains over 12,000 terms describing abnormal human phenotypes. However, the labels and synonyms in the HPO primarily use medical terminology, which can be difficult for patients and their families to understand. In order to make the HPO more accessible to non-medical experts, we systematically added new synonyms using non-expert terminology (i.e., layperson terms) to the existing HPO classes or tagged existing synonyms as layperson. As a result, the HPO contains over 6,000 classes with layperson synonyms.

Keywords—*Human Phenotype Ontology, Synonyms, Rare Disease, Patient phenotypes*

I. INTRODUCTION

Every person has a unique collection of phenotypes, or physical and physiological characteristics or traits. Diseases can be characterized by symptoms and abnormal phenotypes and many diseases are caused by underlying genetic variations. Use of genetic analyses like whole genome sequencing can help inform disease diagnosis, as well as analysis of the corresponding patient phenotypes. However, although the cost and ease of collecting and analyzing genomic data has improved rapidly [1], collecting the phenotypic data has not become more standardized, convenient, or less expensive [2], limiting algorithmic approaches. Thus a major challenge in clinical care and research aimed at understanding genetic diseases is phenotyping patients accurately, yet efficiently.

This is a particular challenge for patients with rare or undiagnosed diseases. In these cases, the patients themselves are a valuable resource and may be the best source of phenotyping information on their condition. Not only do patients live with their condition, but they often have a wealth of knowledge about their condition, especially those who have

been evaluated by multiple clinicians. In fact, the only person who may have all of the information about a patient's phenotype is the patient him/herself. A few remarkable stories exist highlighting cases where patients' phenotyping and investigations have led to a diagnosis, such as for NGLY1 [3], or Jill Viles [4] who despite skepticism from her doctors, managed to not only diagnose herself but also to reveal fundamental biology of the Lamin protein. While these particular cases are exceptional, many patients could further their own diagnoses with improved phenotyping.

In order to maximize the usefulness of accurate phenotyping for clinical diagnosis, and to build cohorts of patients for gene discovery, a standard vocabulary is essential. The use of a standardized vocabulary can ensure proper understanding of terminology across different users, such as patients and healthcare professionals. Therefore, using a controlled vocabulary that provides synonyms and definitions for the medical terminology is valuable. To this end, the Human Phenotype Ontology (HPO) (<http://www.human-phenotype-ontology.org/>) was developed for describing phenotypic abnormalities encountered in human disease to facilitate "deep phenotyping", whereby symptoms and characteristic phenotypic findings (a phenotypic profile) are captured using a logically constructed hierarchy of phenotypic terms [5].

In a clinical setting, these phenotypes are defined using medical terminology, which can be difficult for patients to understand. The terminology gap between medical professionals and non-medical experts has long been recognized in many areas of medical practice. The degree to which patients understand the terminology used in medical encounters has been evaluated through various methods and across different disciplines [6-9]. This research has consistently acknowledged and expressed the importance of making the terminology used in medical encounters more accessible to patients. Numerous organizations have developed term lists that align medical terms with lay

language as well as provide guidance on communicating with the public about health issues [10-13]. Additionally, there are information resources, such as MedlinePlus, that provide access to curated, quality health information on a variety of topics in patient-friendly language [14].

While these resources increase accessibility and comprehension of medical terminology for health consumers, other structured vocabularies have been developed to enable cross communication, and comprehension, between non-specialists and medical professionals. These “consumer health vocabularies”, or CHVs, provide patient-friendly terms that are often mapped (or aligned) to established medical terminologies [15-17]. For example, the Unified Medical Language System (UMLS) aims to include lay terms as synonyms or quasi-synonyms in their Metathesaurus, through various efforts (quasi synonyms are terms that are not precisely the same) [15]. To this end, the UMLS Metathesaurus was enhanced with the Dictionary of American Regional English extension to map consumer terms for diabetes to medical terms [14]. These vocabularies are generally broad, containing layperson equivalents for clinical findings as well as medical procedures and equipment. Mapping to standardized terminologies promotes interoperability between disparate sources of health information as well as enables development of informatics tools that assist patients with aspects of their medical care, such as filling out family histories [18].

The terms for CHVs are frequently sourced from online forums and patient-friendly websites focused on health information and medical conditions. An example of these types of online forums are patient registries. A patient registry is a researcher-generated platforms that are “an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves a predetermined scientific, clinical, or policy purpose(s)” [19]. Patient registries are valuable resources for patients who share or are affected by a disease to learn more about their disease and connect with community members. *Inspire* (<http://corp.inspire.com/>) is a platform for patients to engage and share amongst disease-specific communities. With patient permission, Inspire promotes primary and secondary research and analyses based on community contributions. *PatientsLikeMe* (<https://www.patientslikeme.com/>) is a health data sharing platform where patients can share information and connect. In addition, these platforms capture how patients refer to their diseases and symptoms, which is how these forums most directly contribute to the development of consumer health vocabularies [20]. These online platforms can also reveal the developing health literacy of patients, particularly in regards to their specific conditions [21]. Depending on the condition and timeframe, health consumers can become quite proficient in understanding and using medical terminology as it pertains to their particular condition or disease. In many ways, patients can become adept in recognizing and applying medical terminology to symptoms or other aspects of their condition over time.

Recognizing that patients are experts in their medical history and at keeping track of their genetic information, GenomeConnect

(<https://connect.patientcrossroads.org/?org=GenomeConnect>) was conceived by ClinGen (Clinical Genome Resource, <http://clinicalgenome.org/>), a NIH-funded resource of clinical and laboratory geneticists and genetic counselors at over 24 institutions, as a registry to empower patients to help researchers and clinicians understand the genetic contributions to health and disease. GenomeConnect was built on the premise that: “As the utility of genetic and genomic testing in healthcare grows, there is need for a high-quality genomic knowledge base to improve the clinical interpretation of genomic variants. Active patient engagement can enhance communication between clinicians, patients, and researchers, contributing to knowledge building. It also encourages data sharing by patients and increases the data available for clinicians to incorporate into individualized patient care, clinical laboratories to utilize in test interpretation, and investigators to use for research” [22]. To this end, GenomeConnect developed a self-phenotyping survey that generates HPO phenotype profiles. Patients use GenomeConnect to enter their information for researchers and clinicians to use, facilitating the diagnostic evaluation as well as research. Not only may “self-phenotyping” be an accurate and comprehensive source of data on patients, it also empowers patients, which may be particularly beneficial to the undiagnosed disease population.

To make the HPO more accessible to patients in GenomeConnect and other patient registries, we aimed to add non-expert terminology the HPO in the form of synonyms for phenotype classes, as patients are often unfamiliar with technical terminology or may misinterpret meanings without a proper definition or explanation. Similarly, health care providers may be unfamiliar with the colloquial expressions. The goal of this project was to systematically review the current terminology in the Human Phenotype Ontology and to 1) apply lay synonyms to current classes and 2) tag existing classes as layperson where applicable. This resulted in the addition of 6,240 synonyms or primary labels marked as layperson. The layperson classes are available in the current release of HPO, and 44% of synonyms are classified as layperson. Addition of the layperson synonyms to the HPO will increase accessibility for patients to use the HPO, enhance interoperability for clinicians, and enable crowdsourcing by citizen scientists.

II. METHODS

While an initial review of the HPO OWL file was done in Protégé, to expedite the process and make it easier to evaluate patterns in the labels, the entirety of the HPO was downloaded to a collaborative spreadsheet and manually evaluated by members of the HPO development team. The work was divided amongst curators with clinical and biomedical expertise who cross-reviewed each other's work.

Synonyms in the HPO are classified as exact, broad, narrow or related. Exact synonyms are precise alternatives to the HPO term, broad synonyms are more general than the HPO term, narrow synonyms are more specific than HPO term, and related synonyms are associated with the HPO term.

In order to find appropriate synonyms, several methods were used. First we checked online knowledge bases such as Wikipedia, MedlinePlus, Mayo Clinic (<http://www.mayoclinic.org/>), Online Mendelian Inheritance of Man (OMIM, <http://www.omim.org/>) and the Elements of Morphology (<https://elementsofmorphology.nih.gov/>). Next we referred to other ontologies, terminologies, and texts such as Uberon (for anatomical site synonyms), SNOMED CT browsers (e.g., IHTSDO), and specialty medical texts like Gorlin's Syndromes of the Head and Neck, or other similar sources [23]. We made attempts to reuse synonym sub-strings for similar terms, such as layperson terms for terms such as 'absent' for classes using the quality aplasia (PATO_0001483) or for anatomical classes, for example the synonym for 'tailbone' was added to all classes using 'coccyx' (UBERON_0001350).

The terms were scripted into the HPO OWL file. Automated quality checks on the ontology were performed, such as checking for classes with the same label or exact synonym; character encoding; and formatting in title-case. We integrated these into our workflow using Travis CI (<https://travis-ci.org/>). The curation team also performed an exhaustive manual review for consistency across the hierarchy, and checked for errors or inconsistencies. The file is available at: <http://www.human-phenotype-ontology.org> (under Downloads).

III. OUTCOMES

The inclusion of these plain language synonyms will support patient-driven applications for deep phenotyping that can be utilized clinically and computationally, as depicted in Figure 1.




Apert's syndrome		
	Plain language	Medical term
	Webbed toes	Syndactyly
	Deformity due to premature fusing of skull bones	Cranio-synostosis
	Wide-set eyes	Ocular hypertelorism

Figure 1: Patients and medical providers can search HPO for phenotypes of medical conditions such as Apert's syndrome using layperson or medical terms. Apert's syndrome image is provided for illustration only and credits are available from monarchinitiative.org.

As a result of this effort, the HPO now contains a total of 14,253 synonyms for all of the existing classes. Of these synonyms, 6,240 are marked as lay synonyms (Table I). Synonyms were either added to existing classes, or existing classes were tagged as layperson. New synonyms were typed either as exact, broad, related or narrow. The final numbers of each type are reported in Table I.

Table I: Layperson synonyms in HPO

	All synonyms in HPO	Layperson synonyms marked as lay
All synonyms	14253	6240
Exact	12167	5357
Broad	441	298
Related	1236	419
Narrow	409	166

Information content of HPO classes

We aimed to understand the impact of adding lay synonyms to the HPO if they were to be used for disease diagnostics or patient-led cohort discovery. To this end, we performed an evaluation of the information content (IC) content for HPO classes that were tagged as layperson or those that contain layperson synonyms. Mathematically this is expressed as the negative logarithm of the frequency with which the class is used to describe a disease, i.e. more general classes (such as 'Abnormality of the nervous system') have a low IC and very specific classes (such as 'Spinal cord posterior columns myelin loss') have high IC. Figure 2 shows the distribution of the IC for the HPO classes with a label or synonym marked as layperson. The analysis shows that major

fraction of layperson synonyms were added to very specific HPO classes. This could substantially help in the differential diagnostic process for HPO users. This is due to the fact that searching and identifying diseases with specific HPO classes is now easier in case users do not know the specific medical terms.

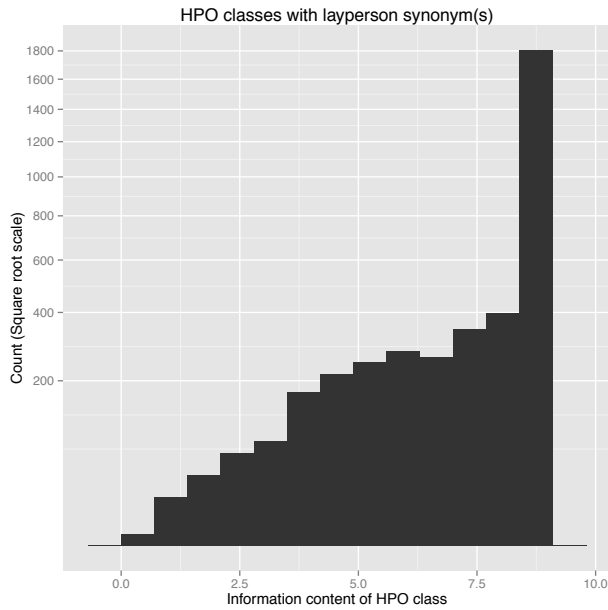


Figure 2: The distribution the HPO classes with layperson synonyms. The average IC of these classes is 7.4.

IV. CHALLENGES

The process of adding layperson synonyms gave rise to several challenges.

Layperson terms were not added to all the HPO classes. As exemplified in Figure 2, some HPO classes already used layperson terminology, so they were tagged as layperson, and an additional layperson synonym was not added. In some cases, a layperson term simply does not exist; for example, it is difficult to describe a joint contracture using non-medical terminology. In some instances, the layperson version of an HPO class might be the literal definition in the HPO, which we tried to avoid. For example, the term ‘Vasculitis’ (HP_0002633) is defined as ‘Inflammation of blood vessel’, which would be a likely addition as a layperson synonym. In adding synonyms, questions emerged as to whether or not certain synonyms were useful to add. An example is the bones in the body - many of these have assigned names (e.g., radius, coccyx). In some instances, as with coccyx, ‘tailbone’ has emerged as a widely used synonym; however, in other cases, a potential synonym not only strongly resembles the definition of the term (like using ‘short bone in forearm’ as a synonym for ‘radius’), it also may not be a term widely used amongst laypeople or clinicians. In the case of radius and ulna, these are both forearm bones, but there is not a way to differentiate them in layperson terminology.

Another challenge was ensuring that the application of a layperson synonym aligned with the definition of the assigned HPO class. For example, colorblindness could be broadly used to describe many classes such as HP_0007641 ‘Dyschromatopsia’ or HP_0007803 ‘Monochromacy’, but there are specific differences between these two classes, with dyschromatopsia being defined as ‘A form of colorblindness in which only two of the three fundamental colors can be distinguished due to a lack of one of the retinal cone pigments’ and monochromacy defined as ‘Complete color blindness, a complete inability to distinguish colors. Affected persons cannot perceive colors, but only shades of gray’. These two classes were therefore given more specific layperson synonyms, ‘colorblindness’ and ‘total colorblindness’, respectively. The subclasses of dyschromatopsia were assigned more specific layperson synonyms as well, such as HP_0011521 Deuteranopia, layperson synonym: Green-blind, and HP_0011522 Protanopia, layperson synonym: Red-blind, even though these classes may be more broadly referred to as colorblindness.

It was also necessary to recognize the relationships within the ontology and applying proper consistency across classes/sub-classes when adding layperson synonyms. For example, the layperson synonym, ‘Yellowing of the skin’, was added to the HPO class, ‘Jaundice’. In order to maintain consistency in the application of layperson synonyms, ‘Yellowing of the skin’ also needed to be added to sub-classes, ‘Intermittent jaundice’ and ‘Prolonged neonatal jaundice’.

V. NEXT STEPS

A next step is to develop a method of validating the added layperson synonyms in order to determine whether or not they are reflective of terms actually used and recognized by patients and clinicians alike. This will be done by the HPO development team and via a crowd sourcing approach. We will encourage crowd sourcing for requests for additional layperson synonyms, as well as validating the existing layperson terms. Validation would also assist with determining which layperson synonym is marked as ‘primary’ within the HPO, so that a lay version of the HPO can be used in software applications and surveys geared towards patients.

VI. CONCLUSIONS

The addition of layperson synonyms increases the usability of the HPO, making it useful for data interoperability across clinicians and patients. Additionally, this work will enable crowdsourcing by citizen scientists. The layperson synonyms are available in the current release of the HPO and are available at www.purl.obolibrary.org/obo/hp.owl. Additionally, community contributions are welcome by submitting to our issues tracker: <https://github.com/obophenotype/human-phenotype-ontology>.

ACKNOWLEDGMENT

This work is supported by NIH Office of Director grant: 1R24OD011883. Thank you to Tudor Groza and Julie McMurry for their help. Apert's syndrome image credits available from monarchinitiative.org

- [22] Kirkpatrick BE, Riggs ER, Azzariti DR, et al. GenomeConnect: matchmaking between patients, clinical laboratories, and researchers to improve genomic knowledge. *Hum Mutat.* 2015;;36(10):974-978. doi: 910.1002/humu.22838. Epub 22015 Aug 22836.
- [23] Raoul Hennekam, Judith Allanson, Ian Krantz. *Gorlin's Syndromes of the Head and Neck.* Oxford University Press; 5 edition (February 5, 2010)

REFERENCES

- [1] Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! *Genome Biol.* 2011;;12(8):125. doi: 110.1186/gb-2011-1112-1188-1125.
- [2] Hunter L. Computational challenges of mass phenotyping. *Pac Symp Biocomput.* 2013:454-455.
- [3] Might M, Wilsey M. The shifting model in clinical diagnostics: how next-generation sequencing and families are altering the way rare diseases are discovered, studied, and treated. *Genet Med.* 2014; 16(10):736-737. doi: 710.1038/gim.2014.1023. Epub 2014 Mar 1020.
- [4] Epstein D. The DIY Scientist, the Olympian, and the Mutated Gene. 2016; <https://www.propublica.org/article/muscular-dystrophy-patient-olympic-medalist-same-genetic-mutation>. Accessed January 17, 2016, 2016.
- [5] Kohler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* 2014;;42(Database issue):D966-974. doi: 910.1093/nar/gkt1026. Epub 2013 Nov 1011.
- [6] Spiro, D. and F. Heidrich (1983). "Lay understanding of medical terminology." *Journal of Family Practice* 17(2): 277-279.
- [7] Pieterse, A. H., et al. (2013). "Lay understanding of common medical terminology in oncology." *Psycho-Oncology* 22(5): 1186-1191.
- [8] Chapman, K., et al. (2003). "Lay understanding of terms used in cancer consultations." *Psycho-Oncology* 12(6): 557-566.
- [9] Barker, K. L., et al. (2009). "Divided by a lack of common language? A qualitative study exploring the use of language by health professionals treating back pain." *BMC Musculoskeletal Disorders* 10: 123.
- [10] http://www.cdc.gov/other/pdf/everydaywordsforpublichealthcommunication_final_11-5-15.pdf
- [11] <http://www.portland.va.gov/research/documents/hrpp/glossary-of-lay-terms.pdf>
- [12] <http://hso.research.uiowa.edu/medical-terms-lay-language>
- [13] https://humansubjects.stanford.edu/new/docs/glossary_definitions/lay_language.pdf
- [14] Miller, N., et al. (2000). "MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service." *Bulletin of the Medical Library Association* 88(1): 11-17.
- [15] Tse T, Soergel D. Exploring Medical Expressions Used by Consumers and the Media: An Emerging View of Consumer Health Vocabularies. *AMIA Annual Symposium Proceedings.* 2003;2003:674-678.
- [16] Patrick TB, Monga HK, Sievert MC, Hall JH, Longo DR. Evaluation of Controlled Vocabulary Resources for Development of a Consumer Entry Vocabulary for Diabetes. *Journal of Medical Internet Research.* 2001;3(3):e24. doi:10.2196/jmir.3.3.e24.
- [17] Sedorff, M., et al. (2013). Incorporating expert terminology and disease risk factors into consumer health vocabularies. *Pacific Symposium on Biocomputing:* 421-432.
- [18] Hulse, N. C., et al. (2010). Deriving consumer-facing disease concepts for family health histories using multi-source sampling. *J Biomed Inform* 43(5): 716-724.
- [19] Gliklich R, Dreyer N. *Registries for Evaluating Patient Outcomes: A User's Guide.* Rockville, MD: Agency for Healthcare Research and Quality; 2010. AHRQ Publication No. 10-EHC049.
- [20] Smith, C. A. and P. J. Wicks (2008). PatientsLikeMe: Consumer health vocabulary as a folksonomy. *AMIA Annu Symp Proc:* 682-686.
- [21] Fage-Butler, A. M. and M. Nisbeth Jensen (2015). Medical terminology in online patient-patient communication: evidence of high health literacy? *Health Expectations.*