Proceedings of the Joint Bio-Ontologies and BioLINK ISMB'2014 SIG session "Phenotype Day" 2014:41-44.

Coverage of Phenotypes in Standard Terminologies

*Rainer Winnenburg and Olivier Bodenreider**

National Library of Medicine, Bethesda, Maryland, USA

ABSTRACT

Objective: To assess the coverage of the Human Phenotype Ontology (HPO) phenotypes in standard terminologies. **Methods**: We map HPO terms to the UMLS and its source terminologies and compare these lexical mappings to HPO cross-references. **Results**: Coverage of HPO classes in UMLS is 54% and 30% in SNOMED CT. Lexical mappings largely outnumber cross-references. **Conclusions**: Our approach can support the development of cross-references to standard terminologies in HPO. **Supplementary file**: Our mapping to UMLS is available at: http://mor.nlm.nih.gov/-pubs/supp/2014-biolink_phenotype-rw/index.html

1 INTRODUCTION

While the past decades have seen unprecedented efforts directed towards genotyping, parallel efforts are required on the side of phenotyping in order to understand how genetic variation relates to clinical manifestations (Hennekam and Biesecker, 2012). Coarse phenotyping has been shown to be useful for some purposes and the potential of using phenotypes based on electronic health record (EHR) data for genomic studies has been demonstrated (e.g., Newton, et al., 2013). However, the study of rare syndromes will likely require detailed phenotyping.

Efforts such as PhenX (Hamilton, et al., 2011) are underway to facilitate the adoption of standards for phenotyping across domains, in particular for use in genome-wide association studies (GWAS). However, resources for phenotyping tend to vary between clinical data repositories used for translational research and in healthcare settings. For example, while somewhat overlapping, the Human Phenotype Ontology (HPO) used for annotation of research data and SNOMED CT used in EHRs are not developed in a coordinated fashion and are only partially interoperable.

The main objective of this work is to assess the coverage of (fine-grained) phenotypes in standard terminologies. More specifically, we study the extent to which phenotypes from HPO are covered in the UMLS and its source vocabularies, including SNOMED CT and MeSH. A secondary objective is to compare the cross-references to standard terminologies provided by HPO to mappings of HPO terms to and through the UMLS.

* To whom correspondence should be addressed.

2 BACKGROUND

2.1 Resources

HPO. The Human Phenotype Ontology (HPO) is an ontology of phenotypic abnormalities developed collaboratively and used for the annotation of databases such as OMIM (Online Mendelian inheritance in Man), Orphanet (knowledge base about rare diseases), and DECIPHER (RNAi screening project) (Kohler, et al., 2014). The current version of HPO contains 10,491 classes and 16,414 names for phenotypes, including 5,923 exact synonyms in addition to one preferred term for each class. HPO also provides a rich set of cross-references to standard terminologies such as the UMLS, MeSH and SNOMED CT (see below). Additionally, HPO distributes a database of annotations for over 7,000 human hereditary syndromes in reference to HPO classes. However, because this investigation focuses on phenotype terms, only the ontology part of HPO is used here. The version of HPO used in this investigation is the (stable) OWL version downloaded on April 16, 2014 from the HPO website (http://www.human-phenotype-ontology.org/).

UMLS. The Unified Medical Language System (UMLS) is a terminology integration system developed by the U.S. National Library of Medicine (Bodenreider, 2004). The UMLS Metathesaurus integrates many standard biomedical terminologies, including SNOMED CT, the Medical Subject Headings (MeSH), several versions of the International Classification of Diseases, the Medical Dictionary for Regulatory Activities (MedDRA), as well as several nursing terminologies and consumer health vocabularies. Although the UMLS does not currently integrate HPO, it is expected to provide a reasonable coverage of phenotypes through its source vocabularies. In the UMLS Metathesaurus, synonymous terms from various sources are assigned the same concept unique identifier, creating a mapping among these source vocabularies. Terminology services provided for the UMLS support the lexical mapping of terms to UMLS concepts. Additionally, each UMLS is assigned at least one semantic type from the UMLS Semantic Network. These semantic types are clustered into Semantic Groups, which provide a partition of the 3 million UMLS concepts into 15 broad domains, including *Disorders*, *Anatomy* and *Genes & Molecular Sequences*. The 2013AB version of the UMLS is used in this work.

2.2 Related work

HPO has been studied mostly for its applications (e.g., cross-species analysis of phenotypes (Robinson and Webber, 2014)). Besides, researchers have investigated the representation of phenotypes through pre- and post-coordinated terms (Oellrich, et al., 2013). However, except for the integration of HPO into the Health Terminology/Ontology Portal (HeTOP) (Grosjean, et al., 2013), relatively little attention has been devoted to the terminological characteristics of HPO and to the representation of phenotypes in standard terminologies. While the coverage of specific subdomains of medicine has been studied (e.g., Chute, et al., 1996; Kim, et al., 2006), to the best of our knowledge, this investigation is the first one to focus on phenotypes in standard terminologies.

The specific contribution of this work is to investigate the coverage of HPO phenotypes in standard terminologies and to propose approaches for increased operability between terminological resources.

3 MATERIALS AND METHODS

Our approach to assessing the coverage of HPO phenotypes in standard terminologies can be summarized as follows. We start by extracting HPO terms and cross-references from the OWL file. We map HPO terms to the UMLS, and through UMLS concepts, to concepts from the source vocabularies in the UMLS, including SNOMED CT and MeSH, and assess the proportion of HPO classes represented in each source. Finally, we compare the cross-references to UMLS provided by HPO to the lexical mappings of HPO terms to UMLS concepts. Similarly, we compare the cross-references to standard terminologies provided by HPO to the mappings derived through the UMLS.

3.1 Extracting HPO terms and cross-references

For each HPO class, we extracted its identifier (`oboInOwl:id`), along with all preferred terms (`rdfs:label`) and synonyms (`oboInOwl:hasExactSynonym`). Synonyms other than "exact synonyms" were not extracted. We also extracted the cross references of HPO classes to UMLS and standard terminologies (`oboInOwl:hasDbXref`). For example, the class identified by *HP:0003419* has *Low back pain* as its preferred term, has *Lower back pain* as an exact synonym, and has cross-references to UMLS (*C0024031*) and MeSH (*D017116*). In this work, we ignore the cross-references that are not in the OWL file.

3.2 Lexical mapping of HPO terms to UMLS

We map each HPO term, preferred term or synonym, to the UMLS using increasingly aggressive methods, namely exact match (case insensitive) and normalization. Normalization abstracts away from minor differences in terms, including case, punctuation, inflectional variants (e.g., singular vs. plural), and stop words. It also ignores word order. For example, the term *Low back pain* maps to UMLS concept *C0024031* through an exact match. (Although not used in this mapping, the normalized form of *Low back pain* would be "*back low pain*".) We consider as lexical mappings for a given HPO class the set of UMLS concepts obtained from the mapping of each term in the class (preferred term and synonyms). Here, the synonym *Lower back pain* also maps to *C0024031*, so there is only one UMLS concept mapped to for the HPO class *HP:0003419*.

In order to avoid false positive mappings, we add semantic restrictions to the mapping. More specifically, we ignore mappings to UMLS semantic groups other than *Disorders*, *Anatomy*, *Phenomena* and *Physiology*. While most phenotypes are expected to map to concepts from the *Disorders* group (including signs and symptoms, in addition to diseases and syndromes), we also allow mappings to these other semantic groups to cover, for example, anatomical structures, whose pathological persistence can correspond to a phenotype (e.g., *Ductus arteriosus*). The semantic constraints prevent the mapping of some HPO terms to a gene name, when the gene name matches the name of the phenotype (e.g., the HPO class *Insulin resistance* (*HP:0000855*) maps to two UMLS concepts, one for the pathologic function, i.e., a phenotype, the other corresponding to an allelic variant, i.e., a genotype. The mapping to the latter is ignored through semantic filtering.)

3.3 Deriving mappings to standard terminologies through UMLS

Through the mapping to a UMLS concept, we can derive a mapping to the vocabularies integrated in the UMLS, more precisely to those vocabularies, whose terms have been found synonymous with *Low back pain* and assigned the same identifier *C0024031*. Such terms include *Low Back Pain* from MeSH (*D017116*), *Low back pain* from MedDRA (*10024891*), and *Low back pain* from SNOMED CT (*279039007*), among others.

3.4 Assessing the coverage of HPO phenotypes in UMLS and standard terminologies

In order to assess the coverage of HPO phenotypes in the UMLS and standard terminologies, we simply compute the proportion of HPO classes for which we find a cross-reference provided by HPO or a lexical mapping to or through the UMLS.

3.5 Comparing HPO cross-references to lexical mappings to and through UMLS

Having extracted the cross-references provided by HPO for a given class and mapped all terms for this class to the UMLS, we can compare the set of identifiers obtained with each method for a given target. For example, the HPO class *HP:0003419* maps to the same UMLS concept (*C0024031*)

through both cross-references and lexical mapping. Similarly, HPO provides a cross-reference to the MeSH descriptor *D017116*, which happens to be the same MeSH descriptor to which a mapping can be derived through UMLS. However, a mapping to MedDRA (*10024891*) and to SNOMED CT (*279039007*) can also be established through the UMLS, whereas no cross-reference is provided by HPO (in the OWL file) to these target terminologies.

For each HPO class, we compare the set of target concepts (to UMLS or any of the standard terminologies under investigation), obtained through the cross-references provided by HPO, to the lexical mappings to the UMLS and to standard terminologies through the UMLS. In addition to the terminologies targeted by HPO cross-references, we also explore a variety of source vocabularies in the UMLS, including clinical vocabularies, nursing vocabularies and consumer health vocabularies, in order to assess whether phenotypes can be annotated with these resources in clinical repositories and in consumer health information sources.

4 RESULTS

4.1 Coverage of HPO phenotypes

We extracted for 10,491 HPO phenotypes (classes) their preferred terms and 5,923 synonyms and mapped them to UMLS concepts. In total, some cross-reference or lexical mapping to UMLS was found for 5,858 HPO classes (56%). In a second step, we used the lexical mappings to UMLS we identified for 5,650 HPO classes (54%) to derive mappings to concepts from several source vocabularies in the UMLS. Through these UMLS concepts, 3,116 classes (30%) mapped to SNOMED CT concepts and 1,970 (19%) to MeSH descriptors and supplementary concepts (see Table 1). Finally, for 4,633 HPO classes (44%), there are neither cross-references nor lexical mappings to UMLS. Differences in the representation of phenotypes across sources are sometimes responsible for the failure to link HPO classes to standard terminologies. For example, the class *Third toe clinodactyly* has no correspondence in any UMLS source vocabulary, because a similar notion is represented there as *3rd-4th toe clinodactyly* (*C1858040*).

4.2 Cross-references vs. lexical mappings

We compared the cross-references to UMLS provided by HPO to the lexical mappings of HPO terms to UMLS concepts. While HPO provides cross-references to the UMLS for 36% of their classes, we were able to identify lexical mappings for 54% of the classes. As shown in Figure 1, the coverage provided by lexical mappings is systematically and often largely (e.g., SNOMED CT) superior to that of the HPO cross-references.

The various types of differences observed between HPO cross-references and lexical mappings to UMLS concepts are presented in Table 1. The largest category (38%) corresponds to HPO phenotypes with identical sets of UMLS concepts through cross-references and lexical mappings. An example form this category is the HPO class *Low back pain* as (*HP:0003419*) presented earlier. Phenotype classes for which lexical mappings were obtained but for which no cross-references are provided in HPO represent 36% of the cases. For example, HPO does not provide a cross-reference for the phenotype *Subcutaneous hemorrhage*, for which the lexical mapping obtains *Haemorrhage subcutaneous (C0854107)*. Conversely, our method failed to obtain lexical mappings for 168 classes (3%) with cross-references in HPO. For example, because of terminological variation beyond what is absorbed by normalization, no lexical mapping is identified for the HPO term *Increased circulating cortisol level*, while a cross-reference to *Serum cortisol increased* (*C0241003*) is provided by HPO.

Similarly, we compared the cross-references to standard terminologies provided by HPO to those derived through the UMLS. In Table 2 we present the comparison for MeSH. By and large, the lexical mappings are either identical to the cross-references provided in HPO (for 46% of HPO classes) or they supplement the cross-references to MeSH (48%).

Table 1. Relations between HPO classes and UMLS concepts

HPO classes to UMLS concepts	#	%
Classes with identical sets of UMLS concepts cross-referenced in HPO and through lexical mapping	2206	37.7
Classes with identical sets of UMLS concepts (each UMLS concept from the cross-references set is identical to or hierarchically related to a UMLS concept in the lexical mapping set)	189	3.2
Classes with additional UMLS concepts in the cross-references set only	84	1.4
Classes with additional UMLS concepts in the lexical mapping set only	976	16.7
Classes with additional UMLS concepts in both the HPO cross-references and the lexical mapping set	117	2.0
Classes with cross-references only (no lexical mappings)	168	2.9
Classes with lexical mappings only (no cross-references)	2118	36.2
Total number of classes related to UMLS concepts	**5858**	**100.0**

Table 2. Relations between HPO classes and MeSH descriptors and supplementary concepts ("MeSH terms")

HPO classes to MeSH terms	#	%
Classes with identical sets of MeSH terms cross-referenced in HPO and through lexical mapping	922	46.2
Classes with identical sets of MeSH terms (each MeSH term from the cross-references set is identical to or hierarchically related to a MeSH terms in the lexical mapping set)	51	2.6
Classes with additional MeSH terms in the cross-references set only	0	0.0
Classes with additional MeSH terms in the lexical mapping set only	32	1.6
Classes with additional MeSH terms in both the HPO cross-references and the lexical mapping set	3	0.2
Classes with cross-references only (no lexical mappings)	24	1.2
Classes with lexical mappings only (no cross-references)	963	48.3
Total number of classes related to MeSH terms	**1995**	**100.0**

5 DISCUSSION

5.1 Coverage of HPO phenotypes

The coverage of HPO phenotypes in the UMLS as a whole is 54% and is only 30% in the best individual standard terminology, SNOMED CT. This proportion is likely to be insufficient for fine-grained phenotyping in EHR data. In contrast to nursing vocabularies, consumer health vocabularies show a relatively high coverage of phenotypes. This suggests that they could be used to annotate phenotypes in consumer health information resources.

5.2 Cross-references vs. lexical mappings

Overall, as shown in Figure 1 (light gray bars), HPO provides cross-references for a limited proportion of its classes. The lexical mapping to and through UMLS provides systematically and largely more links to concepts in standard terminologies, demonstrating the potential of our approach for increasing the interoperability between resources. Moreover, we noted the presence of 127 cross-references to obsolete UMLS concepts, which reflects a maintenance issue.

5.3 Limitations and future work

The analysis presented here is essentially quantitative. A detailed qualitative analysis should be performed in order to investigate terminological variants and differences in concept representation. Another limitation is that, except for semantic filtering, no validation of the lexical mappings was performed. Finally, the cross-references to MedDRA provided in an ancillary file should also be considered.

ACKNOWLEDGEMENTS

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine (NLM).

REFERENCES

Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Res*, **32**, D267-270.

Chute, C.G., *et al.* (1996) The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures, *J Am Med Inform Assoc*, **3**, 224-233.

Grosjean, J., *et al.* (2013) Integrating the human phenotype ontology into HeTOP terminology-ontology server, *Stud Health Technol Inform*, **192**, 961.

Hamilton, C.M., *et al.* (2011) The PhenX Toolkit: get the most from your measures, *Am J Epidemiol*, **174**, 253-260.

Hennekam, R.C. and Biesecker, L.G. (2012) Next-generation sequencing demands next-generation phenotyping, *Hum Mutat*, **33**, 884-886.

Kim, H., *et al.* (2006) Content coverage of SNOMED-CT toward the ICU nursing flowsheets and the acuity indicators, *Stud Health Technol Inform*, **122**, 722-726.

Kohler, S., *et al.* (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data, *Nucleic Acids Res*, **42**, D966-974.

Newton, K.M., *et al.* (2013) Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network, *J Am Med Inform Assoc*, **20**, e147-154.

Oellrich, A., Grabmuller, C. and Rebholz-Schuhmann, D. (2013) Automatically transforming pre- to post-composed phenotypes: EQ-lising HPO and MP, *J Biomed Semantics*, **4**, 29.

Robinson, P.N. and Webber, C. (2014) Phenotype Ontologies and Cross-Species Analysis for Translational Research, *PLoS Genet*, **10**, e1004268.

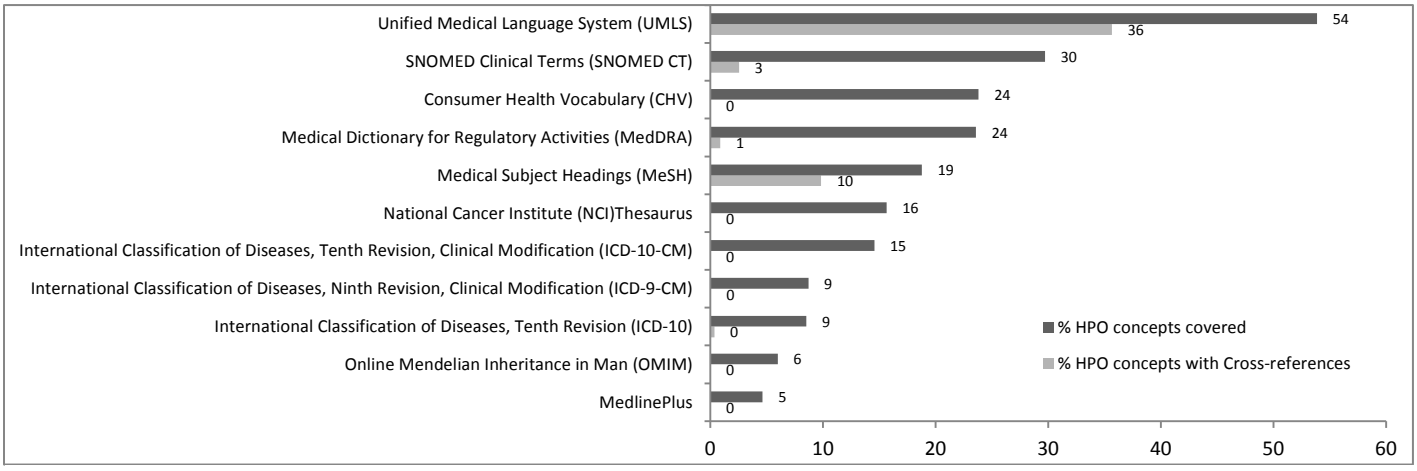

Figure 1. Coverage of HPO phenotypes in the UMLS and in standard terminologies through lexical mappings (dark gray) and cross-references provided in HPO (light gray).