Proceedings of the ISMB'2015 SIG session "Phenotype Day" 2015:8-11

Investigating the lexico-syntactic properties of phenotype terms – Application to interoperability between HPO and SNOMED CT

Ferdinand Dhombres and Olivier Bodenreider[*]

National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA.

ABSTRACT

Objective: To investigate the lexico-syntactic properties of clinical phenotype terms in order to identify partial lexical mappings between HPO and SNOMED CT. **Methods**: We identify modifiers HPO terms and attempt to map demodified terms to SNOMED CT through UMLS. **Results**: We identified partial mappings to SNOMED CT for 20% of HPO concepts with no complete mapping to SNOMED CT. **Conclusions**: Through complete and partial mappings, 50% of the HPO concepts can be mapped to SNOMED CT.

1 INTRODUCTION

In parallel to the deep sequencing effort enabled by Next Generation Sequencing technologies, a need for deep phenotyping has emerged (Robinson, 2012). Clinical phenotypes can be recorded in reference to multiple terminologies, including the Human Phenotype Ontology (HPO) and the Standardized Nomenclature of Medicine Clinical Terms (SNOMED CT). Therefore, there is also a need for interoperability between datasets annotated with different terminologies, especially electronic health record data (Frey, et al., 2014).

The interoperability between HPO and SNOMED CT can be addressed in several complementary ways, through lexical mappings (complete or partial) and by leveraging the logical definitions of phenotypes.

Complete lexical mappings identify exact and normalized matches between existing ("pre-coordinated") terms in HPO and SNOMED CT and denote equivalent relations between the corresponding concepts. In previous work, we showed that only 30% of HPO concepts could map to pre-coordinated SNOMED CT concepts (Winnenburg and Bodenreider, 2014). For example, *Multicystic dysplastic kidney* [HP:0000003] maps to *Multicystic renal dysplasia* [SNCTID:204962002] (through synonymy).

Partial lexical mappings identify matches similar to complete lexical mappings, but allow some words of the HPO terms to be omitted in the mapping to SNOMED CT. Such mappings denote subsumption (subclass) relations between the more specific HPO concept and the more general SNOMED CT concept mapped to. For example, *Bilateral renal atrophy* [HP:0012586] maps to the more general concept *Atrophy of kidney* [SCTID:197659005] (ignoring the modifier *bilateral*). Leveraging the compositional features of HPO terms for mapping purposes had already been suggested by (Beck, et al., 2012).

Mappings leveraging the logical definitions of phenotypes. Since both HPO and SNOMED CT are developed using description logics, it would be possible to compare the logical definitions of phenotype concepts in the two terminologies. However, given the differences in modeling choices in HPO and SNOMED CT, few matches would be expected. Instead, we analyzed the logical definitions of existing phenotype concepts in SNOMED CT and created patterns ("post-coordinated expressions") from these definitions that could be applied to HPO phenotypes not represented in SNOMED CT as pre-coordinated concepts. Through this approach, 1617 additional mappings could be identified between HPO and SNOMED CT (Dhombres, et al., 2015). For example, *Aplastic clavicle* [HP:0006660] would be equivalent to the following post-coordinated expression in SNOMED CT: '*Disease* and (***Role group*** some ((***Associated morphology*** some *Hypoplasia*) and (***Occurrence*** some *Congenital*) and (***Finding site*** some *Clavicle*)))'.

The objective of this paper is to investigate the lexico-syntactic properties of clinical phenotype terms in order to identify partial lexical mappings between HPO and SNOMED CT.

2 BACKGROUND

2.1 Resources

SNOMED CT is developed by the International Health Terminology Standard Development Organization (IHTSDO) (IHTSDO, 2015). It is the world's largest clinical terminology and provides broad coverage of clinical medicine, including diseases and phenotypes. SNOMED CT includes pre-coordinated concepts (with their terms ("descriptions")) and supports post-coordination, i.e., the principled creation of expressions (logical definitions) for new concepts. The U.S. edition of SNOMED CT dated March 2015 used in this work includes about 300,000 active concepts, of which 103,748 correspond to clinical findings.

[*] Corresponding author: olivier@nlm.nih.gov.

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) and Orphanet (knowledge base about rare diseases) (Kohler, et al., 2014). The version of HPO used in this investigation is the (stable) OWL version downloaded on January 21, 2015 (build #1337) from the HPO website (http://www.human-phenotype-ontology.org/). It contains 10,589 classes (concepts) and 16,608 names (terms) for phenotypes, including 6019 exact synonyms in addition to one preferred term for each class.

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. Although the UMLS does not yet 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 by the UMLS support the lexical mapping of terms to UMLS concepts. The 2014AB version of the UMLS is used in this work.

2.2 Related work

Particularly relevant to this investigation where we attempt to find partial mappings for HPO concepts in SNOMED CT by removing some of modifiers that specialize phenotype terms in HPO is work done on the compositional aspects of biomedical terms. Terminologies, such as the Gene Ontology, have been shown to be highly compositional (Ogren, et al., 2004) in that some of their more complex terms are derived from simpler terms by addition of modifiers. Moreover, it has been reported that the compositional structure of Gene Ontology impacts its usage (Ogren, et al., 2005). Similarly, the compositional structure of SNOMED terms has been exploited for assessing the consistency of its hierarchical structure (Bodenreider, et al., 2002). Recent work based on the compositionality of phenotype terms investigated skeletal abnormalities (Groza, et al., 2013) and clinical phenotypes across species (Oellrich, et al., 2013). However in the latter study, the Entity-Quality decomposition strategy yielded better results on the Mammalian Phenotype Ontology than on HPO.

2.3 Specific contribution

The specific contribution of this work is to extend the mapping of clinical phenotypes from HPO to SNOMED CT through partial mappings, leveraging the lexico-syntactic properties of HPO terms.

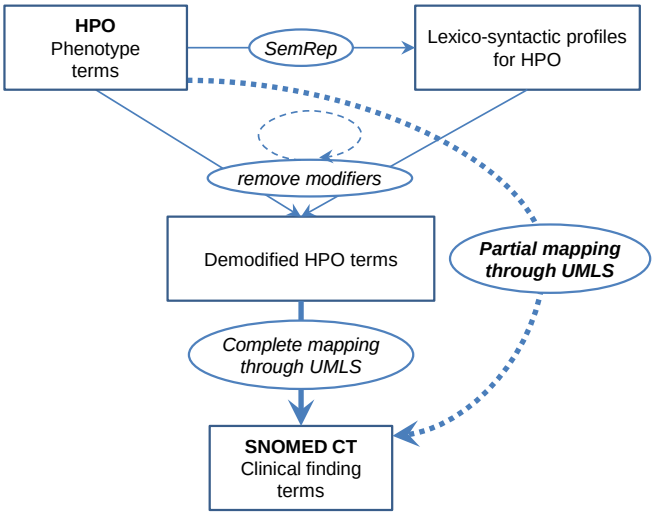

Figure 1. Workflow for term demodification and mappings between HPO and SNOMED CT through UMLS (partial mapping of the original HPO term to SNOMED CT)

3 MATERIAL AND METHODS

Our investigation of the lexico-syntactic properties of phenotype terms for mapping purposes is illustrated in Figure 1 and can be summarized as follows. We extracted phenotype concepts (along with their terms) from HPO and SNOMED CT. We identified lexico-syntactic profiles of interest among the corresponding terms. We then performed increasingly aggressive demodification of HPO terms until the demodified HPO terms could be mapped to SNOMED CT, resulting in a partial mapping of the original HPO term. Finally, we analyzed the modifiers that had to be removed for the mappings to happen and evaluated the partial mappings we obtained.

3.1 Extracting phenotypes terms

From HPO, we selected the concept *Phenotypic abnormality* [HP:0000118] and all its descendants with their corresponding terms (preferred and synonyms). In order to restrict SNOMED CT to phenotypes and disorders, we selected the concept *Clinical Findings* [SCTID:404684003] and all its descendants, along with their terms (referred to as "descriptions" in SNOMED CT).

3.2 Identifying lexico-syntactic profiles

In order to identify modifiers in HPO terms, we performed a lexico-syntactic analysis ("shallow parsing") of these terms using the minimal commitment parser available as part of natural language processing tool *SemRep* (Rosemblat, et al., 2013). For example, the HPO term *Rudimentary uterus* is analyzed as the adjectival modifier *rudimentary* followed by the head noun *uterus*. Its lexico-syntactic profile would therefore be recorded as [MOD-HEAD].

3.3 Demodifying phenotype terms

Since our intuition is that modifiers in specialized HPO terms prevent mapping to the more general terms found in SNOMED CT, we attempted to remove the modifiers identified in HPO terms through lexico-syntactic analysis and to map the demodified terms to SNOMED CT through the UMLS, thereby creating a partial mapping of the original HPO term to SNOMED CT. For example, after removing the modifier *bilateral* from the HPO term *Bilateral renal atrophy* [HP:0012586]with lexico-syntactic profile [MOD-MOD-HEAD], the demodified term *renal atrophy* mapped to SNOMED CT through the UMLS.

More specifically, we focused on terms with a [MOD]*[HEAD] profile (i.e., one or more adjectival or noun modifiers followed by a head noun). We also considered terms containing prepositional attachments, in which we treated each element of the prepositional phrase as a modifier for the purpose of this analysis. For example, the term *Congenital absence of uvula* [HP:0010292] has a the lexico-syntactic profile of [MOD HEAD][PREP HEAD]. Except for head noun of the first noun phrase (*absence*), all the other lexical items are treated as modifiers (*congenital*, *of*, and *uvula*). In practice, we iteratively removed any combination of modifiers from an original HPO term, in increasing order of aggressiveness, i.e., first removing one modifier at the time, then, two modifiers, etc. until only the head noun remained. For example, from the HPO term *Bilateral renal atrophy*, where the head noun *atrophy* is modified by *bilateral* and *renal*, we generated the following three demodified terms, at different levels corresponding to the number of modifiers removed: level 1: *bilateral atrophy*; *renal atrophy*; level 2: *atrophy*.

3.4 Mapping through UMLS

We attempted a complete lexical mapping of the demodified HPO terms to SNOMED CT through the UMLS, as was done for the original HPO terms in (Winnenburg and Bodenreider, 2014). Note that the complete mapping of a demodified term corresponds to the partial mapping of the original term prior to demodification. In order to select the closest mappings, we only recorded the mapping for the less demodified term(s). For example, there is no complete mapping to SNOMED CT for *Bilateral renal atrophy* [HP:0012586], but a "level-1" partial mapping is found to *Atrophy of kidney* [SCTID:197659005] after removing one modifier, *bilateral*.

3.5 Evaluation

We evaluated the quality of the partial mappings by manual review of 5% of the mappings. One of the authors (FD), a physician, classified the mappings as clinically relevant or too broad to be clinically useful.

4 RESULTS

4.1 Extracting phenotypes terms

In HPO, we selected 10,454 concepts specifically representing phenotypic abnormalities and their 16,572 terms. From SNOMED CT, we selected 103,748 concepts for clinical findings, along with 167,986 terms.

4.2 Identifying lexico-syntactic profiles

The lexico-syntactic analysis of the HPO terms produced 542 distinct lexico-syntactic profiles, the most frequent of which being [MOD-HEAD] (28%). The list of the 8 most frequent lexico-syntactic profiles (accounting for 72% of the HPO terms) is shown in Table 1. The 13,494 modifiers extracted from HPO terms include 1416 distinct adjectives and 1405 distinct nouns.

Table 1. Most frequent lexico-syntactic profiles of HPO terms, with indication of mapping to SNOMED CT

Lexico-syntactic profile	Freq.	%	Mapping	%
[MOD – HEAD]	4722	28	2898	61
[MOD – MOD – HEAD]	2416	15	1095	45
[HEAD]	2294	14	1989	87
[HEAD] [PREP – DET – HEAD]	767	5	163	21
[MOD – MOD – MOD – HEAD]	549	3	149	27
[HEAD] [PREP – MOD – HEAD]	432	3	11	3
[MOD – HEAD] [PREP – HEAD]	392	2	151	39
[HEAD] [PREP – HEAD]	383	2	83	22

4.3 Demodification of phenotype terms

Ignoring the 2294 terms consisting of a single head noun ([HEAD]), the majority of HPO terms (8177) were amenable to demodification. We did not process the 6101 terms with complex lexico-syntactic profiles (e.g., with multiple prepositional attachments). In HPO, the most frequent head nouns were *hypoplasia*, *abnormality*, *atrophy*, *weakness* and *ossification*. Excluding prepositions, the most frequent modifiers were *abnormal*, *increased*, *decreased* and *absent*.

4.4 Mapping through UMLS

Replicating our previous study, we identified a complete mapping to clinical findings in SNOMED CT for (at least one term of the) 3090 HPO concepts (30%). Of the HPO concepts with no complete mapping to SNOMED CT, we identified a partial mapping for (at least one term of the) 2136 HPO concepts (20%). A majority of the partial mappings occurred at level 1 (i.e., after removing a single modifier). An analysis of the lowest level at which the mapping occurred is presented in Figure 2. Also, as shown in Table 1, terms with simpler lexico-syntactic profiles have higher rates of mappings to SNOMED CT (after demodification). The most frequently removed modifiers include *progressive*,

recurrent, *abnormal*, *generalized*, *bilateral*, *unilateral*, *congenital*, *episodic*, *severe*, and *multiple*.

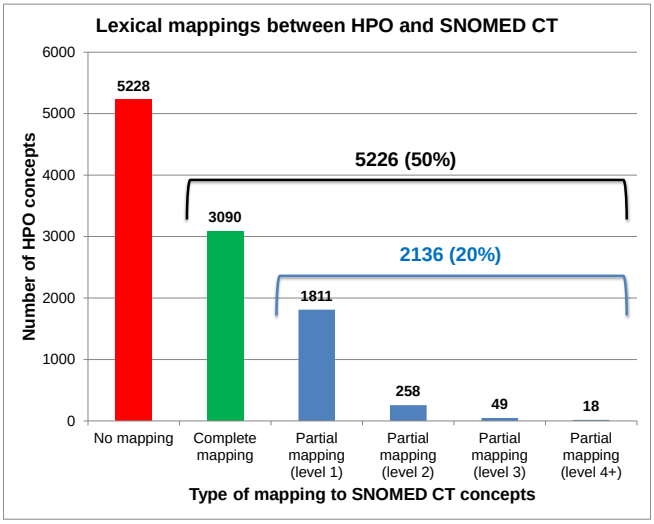

Figure 2. Lexical mappings between HPO and SNOMED CT (type of mapping and minimal level of demodification at which the mapping occurred)

4.5 Evaluation

Our limited review of the partial mappings suggests that most level-1 and level-2 mappings are clinically relevant, while partial mappings at higher levels are usually too broad. In practice, since 97% of the partial mappings occur at level 1 or 2, a vast majority of the partial mappings are potentially useful.

5 DISCUSSION AND CONCLUSIONS

Findings. In addition to the 30% of HPO concepts that can be mapped to SNOMED CT through complete lexical mapping through UMLS, we assessed that 20% of HPO concepts have partial mappings to SNOMED CT concepts, bringing to 50% the proportion of HPO concepts mapped lexically to SNOMED CT (Figure 2). Moreover, we determined that most of the partial mappings we identified by leveraging the lexico-syntactic properties of HPO terms occurred after removing one or two modifiers and were clinically relevant. This investigation also confirmed that HPO concepts tend to be more specialized than phenotype concepts in SNOMED CT.

Limitations and future work. This investigation focuses on a small number of lexico-syntactic profiles. In the future, we plan to explore the potential contribution of the 6101 terms with complex lexico-syntactic profiles. Along the same lines, we want to revisit the terms characterized as "single head nouns" ([HEAD]), some of which actually correspond to multi-word terms identified by the SPECIALIST lexicon as a single lexical unit, but also potentially demodifiable (e.g., *choanal atresia* and *multiple epiphyseal dysplasia*). Finally, we need to carefully assess the overlap between the partial mapping approach presented here and the use of post-coordination in SNOMED CT (Dhombres, et al., 2015).

ACKNOWLEDGEMENTS

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine, the French Gynecology and Obstetrics Association (Collège National des Gynécologues et Obstétriciens Français), and the Philippe Foundation.

REFERENCES

Beck, T., *et al.* (2012) Semantically enabling a genome-wide association study database, *J Biomed Semantics*, **3**, 9.

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

Bodenreider, O., Burgun, A. and Rindflesch, T.C. (2002) Assessing the consistency of a biomedical terminology through lexical knowledge, *Int J Med Inform*, **67**, 85-95.

Dhombres, F., *et al.* (2015) Extending the coverage of phenotypes in SNOMED CT through post-coordination, *MEDINFO Proceedings*, **in press**, 5p.

Frey, L.J., Lenert, L. and Lopez-Campos, G. (2014) EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group, *Yearb Med Inform*, **9**, 206-211.

Groza, T., Hunter, J. and Zankl, A. (2013) Decomposing phenotype descriptions for the human skeletal phenome, *Biomed Inform Insights*, **6**, 1-14.

SNOMED CT: http://www.ihtsdo.org/snomed-ct

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

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

Ogren, P.V., *et al.* (2004) The compositional structure of Gene Ontology terms, *Pac Symp Biocomput*, 214-225.

Ogren, P.V., Cohen, K.B. and Hunter, L. (2005) Implications of compositionality in the gene ontology for its curation and usage, *Pac Symp Biocomput*, 174-185.

Robinson, P.N. (2012) Deep phenotyping for precision medicine, *Hum Mutat*, **33**, 777-780.

Rosemblat, G., *et al.* (2013) A methodology for extending domain coverage in SemRep, *J Biomed Inform*, **46**, 1099-1107.

Winnenburg, R. and Bodenreider, O. (2014) Coverage of Phenotypes in Standard Terminologies. *Joint Bio-Ontologies and BioLINK ISMB'2014 SIG session "Phenotype Day."*. Boston, USA, pp. 41-44.