Automated Identification of Missing IS-A Relations in the Human Phenotype Ontology

Maryamsadat Mohtashamian[1,†], Ran Hu, MS[1,†], Rashmie Abeysinghe, PhD[2], Xubing Hao[1], Hua Xu, PhD[1], Licong Cui, PhD[1,*]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX
[2]Department of Neurology, The University of Texas Health Science Center at Houston, Houston, TX

[†]Contributed equally to this work

*Corresponding author: licong.cui@uth.tmc.edu

Abstract

Auditing the Human Phenotype Ontology (HPO) is necessary to provide accurate terminology for its use in clinical research. We investigate an approach leveraging the lexical features of concepts in HPO to identify missing IS-A relations among HPO concepts. We first model the names of HPO concepts as sets of words in lower case. Then, we generate two types of concept-pairs which have at least a single common word: (1) Linked concept-pairs generated from concept-pairs having an IS-A relation; (2) Unlinked concept-pairs generated from concept-pairs without an IS-A relation. Concept-pairs generate Derived Term Pairs (DTPs) emphasizing unique lexical information of each concept. If a linked concept-pair and an unlinked concept-pair generate the same DTP, then we suggest a potential missing IS-A relation among the unlinked concept-pair. Applying our approach to the 2022-02-14 release of HPO, we uncovered 2,516 potential missing IS-A relations in HPO. We validated 59 missing IS-A relations leveraging the Unified Medical Language System (UMLS) by mapping the concept-pair to UMLS concepts and verifying whether UMLS records an IS-A relation between the pair of concepts.

1 Introduction

Biomedical ontologies or terminologies play an essential role in a wide variety of applications in healthcare and biomedicine such as information retrieval, information integration, data exchange, and natural language processing[1]. Missing and erroneous information in ontologies can lead to various errors in these downstream tasks. Therefore, it is necessary for a biomedical ontology to be frequently updated so that identified quality issues are fixed. The identification of quality issues may be done as part of internal management of the ontology and/or supported through external reporting by users akin to bug reports in software development. However, comprehensive quality assurance approaches targeting specific ontologies are lacking in most circumstances. Manual inspection where a reviewer audits the terminology against certain quality factors, can be impractical at times due to the complexity of modern biomedical ontologies. Therefore, automated approaches that can directly identify specific defects or semi-automated approaches that can direct human reviewers towards potential defects can be of enormous help.

The Human Phenotype Ontology (HPO) is a biomedical ontology that provides a standardized vocabulary for phenotypic abnormalities encountered in human diseases[2,3]. While HPO has an internal sophisticated quality control pipeline[3], external approaches to identify quality issues are extremely rare. In this paper, we investigate an automated lexical approach that detects potential missing IS-A relations in HPO. Our approach is based on the Derived Term Pairs (DTP) that indicates unique lexical information in a concept-pair. The existence of a DTP among a linked concept-pair and an unlinked concept-pair is considered to denote a potential missing IS-A relation. The potential missing IS-A relations uncovered by the approach are validated through external ontological information in the Unified Medical Language System (UMLS).

The rest of the paper is organized as follows. Section 2 introduces HPO and discusses related work in ontology quality assurance. Section 3 contains a detailed discussion on the approach utilized in this work. Section 4 contains the results of the work. Section 5 discusses the results obtained, the drawbacks of the work and some future directions. Finally, the conclusion section concludes the paper.

2 Background

2.1 Human Phenotype Ontology

The Human Phenotype Ontology (HPO) was introduced in 2008 as a comprehensive standard that systematically defines and logically organizes phenotypic abnormalities in human diseases. HPO supports combined genomic and phenotypic analyses by enabling computational inference. Along with other ontologies as well as standards, HPO enables semantic interoperability in digital medicine. Computational disease models utilizing HPO are used in most phenotype-driven genomic diagnostic software[2,3]. The 2022-02-14 release of HPO contains over 15,000 terms and more than 150,000 annotations to hereditary diseases. Figure 1 shows the general hierarchy of HPO. Each term in HPO represents a clinical abnormality and is assigned to one of the five subontologies:

a. Phenotypic abnormality, which contains the descriptions of clinical abnormalities. This is the main subontology of HPO.

b. Mode of inheritance, which contains classes describing the pattern in which a particular genetic trait or disorder is passed from one generation to the next.

c. Clinical modifier, which contains classes describing typical modifiers of clinical symptoms.

d. Clinical course, which includes classes describing the course of a disease.

e. Frequency, which represents frequency of phenotypic abnormalities.

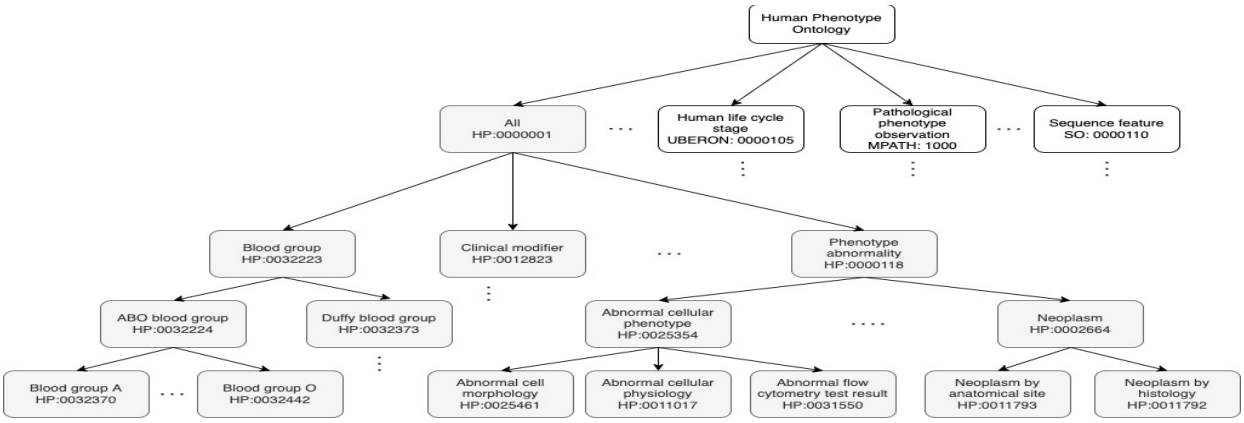

Figure 1. Human Phenotype Ontology hierarchy.

2.2 Auditing methods for biomedical terminologies

Different attributes of an ontology can be audited such as its: (1) terms and concepts; (2) semantic classification of concepts; and (3) semantic relationships among concepts. Various approaches have been proposed to audit biomedical terminologies[4]. Abstraction networks have been widely employed to identify various quality defects in biomedical ontologies[5]. Abstraction networks are a type of summary graphs of an ontology where certain characteristics denote quality issues of an ontology. Campbell et al. have initially investigated the effectiveness of lexical processing algorithms to address errors of omissions in biomedical terminologies. Agrawal et al. have extensively investigated modelling inconsistencies among concepts with lexically similar names in SNOMED CT[7-10]. Bodenreider has proposed an approach where logical definitions are generated through lexical features of concept names in SNOMED CT[11]. The resulting concept hierarchy by reasoning on these logical definitions is compared with the original SNOMED CT hierarchy to identify missing hierarchical relations[11]. In previous studies, we have investigated non-lattice-based approaches for auditing a number of biomedical ontologies[12-16]. Non-lattice subgraphs indicate potentially inconsistent fragments of an ontology as they violate the lattice property, a desirable structural property for a well-formed ontology[17]. In addition, we investigated a lexical-based inference approach where partially matching concept-pairs were identified and further processed to obtain inferred term-pairs, which were leveraged to identify hierarchical relation inconsistencies in Gene Ontology, National Cancer Institute thesaurus, and SNOMED CT[18,19].

3 Methods

We use the OWL (Web Ontology Language) file of the 2022-02-14 release of HPO in this work. Using Owlready2, a python library for ontology programming[20], we extract the concept names and IS-A relations between concepts from the OWL file. We identify linked and unlinked concept-pairs based on the information extracted through OwlReady2. Both linked and unlinked concept-pairs would further yield Derived Term-Pairs that are used to identify potential missing IS-A relations. The approach is discussed in-detail in the following subsections.

3.1 Representing concepts names

The names of the concepts extracted through Owlready2 is further processed as follows. We obtain a set-of-words for each concept by converting the concept name to lowercase, tokenizing the concept name to words, and then removing duplicates. The resultant set-of-words is considered as lexical features representing the concept. For instance, the name of the HPO concept *HP:0000532* is "*Abnormal chorioretinal morphology"*. The set-of-words for this concept would be {*abnormal*, *chorioretinal, morphology*}. Note that since this is an unordered set of words, there is a possibility that another concept would have the same set-of-words.

3.2 Generating linked concept-pairs

A pair of concepts in HPO is said to form a linked concept-pair if they satisfy the following conditions:

a) If their set-of-words have at least a single common word; and

b) If they are connected by a direct or indirect IS-A relation.

For instance, the concept *Colonic atresia (HP:0010448)* with the set-of-words {*colonic, atresia*} and the concept *Intestinal atresia (HP:0011100)* with the set-of-words {*intestinal, atresia*} in Figure 2 form a linked concept-pair as they have a common word {*atresia*} and they are directly connected by an IS-A relation (i.e., *HP:0011100* is the parent of *HP:0010448*). Similarly, the concepts *Hand pain (HP:0046505)* and *Pain (HP:0012531)* form a linked concept-pair, however in this instance the IS-A relation between the concepts is indirect.

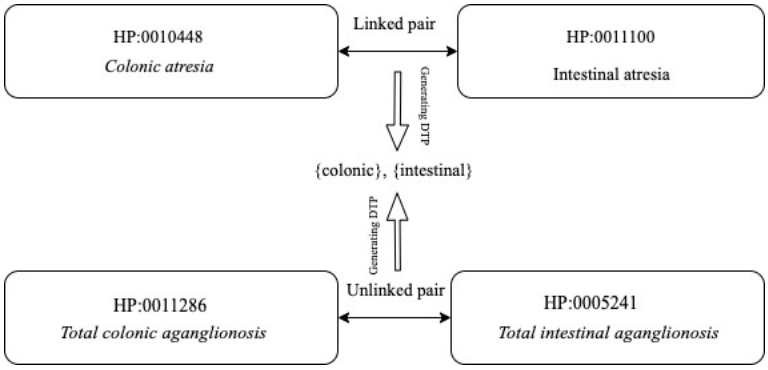

Figure 2. A linked concept-pair *Colonic atresia (HP:0010448)* and *Intestinal atresia (HP:0011100)* as well as an unlinked concept-pair *Total colonic aganglionosis (HP:0011286)* and *Total intestinal aganglionosis (HP:0005241)* generating the same Derived Term Pair (DTP): ({*colonic*}, {*intestinal*}). This suggests a potential missing IS-A relation between the unlinked concept-pair.

Note that HPO reuses some external ontology concepts such as *collagen-containing extracellular matrix (GO:0062023) from* Gene Ontology (GO)[21,22] and *hormone receptor modulator (CHEBI:51061) from Chemical Entities of Biological Interest (ChEBI)*[23,24]. We do not consider such external ontology concepts to generate linked concept-pairs. In addition, deprecated and obsolete concepts are not considered in this work.

3.3 Generating unlinked concept-pairs

A pair of HPO concepts will form an unlinked concept-pair if they satisfy the following conditions:

a) If their sets-of-words have at least a single common word;

b) If both of them fall under the same subontology in HPO; and

c) If they are not connected by either a direct or indirect IS-A relation in HPO.

For instance, the concepts *Total colonic aganglionosis (HP:0011286)* with the set-of-words {*total, colonic, aganglionosis*} and *Total intestinal aganglionosis (HP:0005241)* with the set-of-words in Figure 2 form an unlinked concept-pair as they contain the common words {*total, aganglionosis*}, both are under the subontology *Phenotypic abnormality*, and the two concepts do not have a direct or indirect IS-A relation.

3.4 Generating Derived Term Pairs

Based on the linked or unlinked concept-pairs, we further generate Derived Term Pairs (DTPs) highlighting the unique lexical information in each concept. Let the sets-of-words of a concept-pair C_1 and C_2 be S(C_1) and S(C_2) respectively, the DTP generated by this concept-pair is defined as follows,

$$DTP(C_1, C_2) = (\{x \in S(C_1) \mid x \notin S(C_2)\}, \{x \in S(C_2) \mid x \notin S(C_1)\}).$$

In other words, the common words are removed from the set-of-words of each concept to construct Derived Term Pairs (DTP). Note that if both the sets of words in a DTP are all stop words, such DTPs are ignored. The stop words in consideration are: 'with', 'of', 'and', 'or', 'and/or', 'no', 'not', 'without', 'due to', 'secondary to', 'except', 'by', 'after', 'able', 'removal', 'replacement', 'NOS'. In addition, if both the sets in a DTP are empty, such DTPs are also ignored.

For instance, the linked concept-pair *Colonic atresia (HP:0010448)* with the set-of-words {*colonic, atresia*} and *Intestinal atresia (HP:0011100)* with the set-of-words {*intestinal, atresia*} in Figure 2 generate the Derived Term Pair (DTP) ({colonic}, {intestinal}). Similarly, the unlinked concept-pair *Cervical clear cell adenocarcinoma (HP:0031522)* with the set-of-words {*cervical, clear, cell, adenocarcinoma*} and *Cervical neoplasm (HP:0032241)* with the set-of-words *{cervical, neoplasm}* in Figure 3 generate the DTP ({*clear, cell, adenocarcinoma*}, {*neoplasm*}).

Note that the Derived Term Pair (DTP) is directional, meaning that ({*clear, cell, adenocarcinoma*}, {*neoplasm*}) is not the same as ({*neoplasm*}, {*clear, cell, adenocarcinoma*). In some circumstances, one set of the DTP could be an empty set. For instance, in the linked concept-pair *Hand pain (HP:0046505)* with the set-of-words {*hand, pain*} and *Pain (HP:0012531)* with the set-of-words {*pain*}, the DTP generated is ({*hand*}, {}). This is because the set-of-words of the concept *HP:0012531* is a subset of that of the concept *HP:0046505*.

3.5 Identifying missing IS-A relations

Let C_1 and C_2 be a linked concept-pair, and C_3 and C_4 be an unlinked concept-pair. If $DTP(C_1, C_2) = DTP(C_3, C_4)$, then we suggest that there is a potential missing IS-A relation between the unlinked pair of concepts C_3 and C_4. In other words, if the same Derived Term Pair (DTP) is obtained by a linked concept-pair and an unlinked concept-pair, then we suggest a potential missing IS-A between the unlinked concept-pair.

For instance, the linked concept-pair *Colonic atresia (HP:0010448)* and *Intestinal atresia (HP:0011100)* in Figure 2 generate the DTP ({colonic}, {intestinal}), which can be also generated by the unlinked concept-pair *Total colonic aganglionosis (HP:0011286)* and *Total intestinal aganglionosis (HP:0005241)*. Therefore, we suggest a potential missing IS-A relation: *Total colonic aganglionosis (HP:0011286)* IS-A *Total intestinal aganglionosis (HP:0005241).*

Similarly, the linked concept-pair *Vaginal clear cell adenocarcinoma (HP:0031521)* and *Vaginal neoplasm (HP:0100650)* in Figure 3 generate the Derived Term Pair (DTP) ({*clear, cell, adenocarcinoma*}, {*neoplasm*}) which the unlinked concept-pair *Cervical clear cell adenocarcinoma (HP:0031522)* and *Cervical neoplasm (HP:0032241)* also can generate. Therefore, we suggest a potential missing IS-A relation: *Cervical clear cell adenocarcinoma (HP:0031522)* IS-A *Cervical neoplasm (HP:0032241)*.

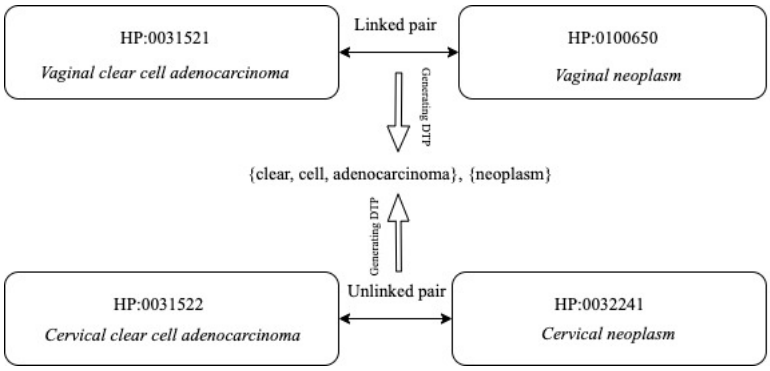

Figure 3. A linked concept-pair *Vaginal clear cell adenocarcinoma (HP:0031521)* and *Vaginal Neoplasm (HP:0100650)* as well as an unlinked concept-pair *Cervical clear cell adenocarcinoma (HP:0031522)* and *Cervical neoplasm (HP:0032241)* generating the same Derived Term Pair (DTP): ({*clear, cell, adenocarcinoma*}, {*neoplasm*}). This suggests a potential missing IS-A relation between the unlinked concept-pair.

3.6 Validating the suggested missing IS-A relations

We leverage external knowledge from the terminologies in the Unified Medical Language System (UMLS) to validate the potential missing relations identified by our approach. The UMLS is an integration of many biomedical vocabularies including SNOMED CT, Gene Ontology, and Medical Subject Headings (MeSH)[25]. The 2021AA release of the UMLS Metathesaurus contains over 4 million concepts from 218 source vocabularies[26]. A key goal of the UMLS is to aggregate and link concepts from all source vocabularies that convey the same meaning[27]. Basic building blocks of the UMLS are concept names from different source vocabularies, which are referred to as atoms. Every atom in the UMLS is assigned an Atom Unique Identifier (AUI). A concept in the UMLS represents a single meaning and aggregates all the atoms from any source that conveys the meaning in any way. All of the atoms within a concept are synonyms. A UMLS concept is also assigned a Concept Unique Identifier (CUI) which uniquely identifies that single meaning. Every UMLS concept is linked to at least one atom[25]. For example, UMLS concept *Back Injuries (*with CUI *C0004601)* is linked to atom *Back Injuries (*with AUI *A0874990)* from MeSH and atom *Injury of back (A33288745)* from SNOMED CT.

In this work, we use the 2021-AA-full version of the UMLS and only leverage the concepts that are in English. We first normalize the names of all atoms in the UMLS. We also normalize the names of the two concepts in each potential missing IS-A relation identified. Normalization involves lowercase conversion, lemmatization, stop word removal, and synonym replacement as performed in our previous work[15]. We use the open-source python library Natural Language Toolkit (NLTK) for the normalization tasks[28]. Then we check whether the names of the two HPO concepts in the missing IS-A relation exist as UMLS atoms by leveraging the Concept Names and Sources file (MRCONSO.RRF) that records each atom in the UMLS. If so, we further check whether there exists a direct or indirect hierarchical relation between the mapped atoms. We leveraged UMLS's Related Concepts file (MRREL.RRF) that records the relationship between concepts or atoms[29]. Note that the IS-A relationship is recorded as CHD (i.e., has child) in this file. If there exist two mapped atoms that have such an IS-A relation recorded in the UMLS, we say that the suggested missing IS-A relation has been validated by the UMLS.

4 Results

We applied this approach to the 2022-02-14 release of HPO. From 16,480 HPO concepts, we identified 128,651 linked concept-pairs, which generated 52,923 distinct Derived Term Pairs (DTP). Out of these DTPs, 1,123 were observed among unlinked-pairs leading to 2,516 potential missing IS-A relations. One of the missing IS-A relations was in the *Clinical modifier* subontology while the rest of them were in the *Phenotypic abnormality* subontology.

4.1 Validation of the suggested missing IS-A relations

Out of the 2,516 suggested missing IS-A relations, 59 were validated using the UMLS. These 59 missing relations were validated by 76 distinct atom-pairs, which indicates that one missing IS-A relation can be validated by multiple atom-pairs in the UMLS. For example, consider the suggested missing IS-A relation: *Retinal vasculitis (HP:0025188)* IS-A Vasculitis (HP:0002633). Concept *HP:0025188* was mapped to both the atom *Retinal vasculitis (A2943949)* from SNOMED CT and *Retinal vasculitis (A2370004)* from MeSH. Meanwhile, the concept *HP:0002633* was mapped to both the atom *Vasculitis (A2887916)* from SNOMED CT and atom *Vasculitis (*A0131615*)* from MeSH. According

to information provided by UMLS, in SNOMED CT, *Vasculitis (A2887916)* is the parent of *Retinal vasculitis (A2943949)*. Similarly, in MeSH, *Vasculitis (*A0131615*)* is the parent of *Retinal vasculitis (A2370004)*. Thus, the suggested missing IS-A relation has been validated independently by both SNOMED CT and MeSH.

Table 1 shows 10 examples of validated missing IS-A relations identified by our approach. For example, *Total colonic aganglionosis (HP:0011286)* IS-A *Total intestinal aganglionosis (HP:0005241)* is a validated missing IS-A relation via UMLS.

Table 1. Ten examples of validated missing IS-A relations identified by our approach.

Descendant concept	Ancestor concept
Total colonic aganglionosis (HP:0011286)	*Total intestinal aganglionosis (HP:0005241)*
Cervical clear cell adenocarcinoma (HP:0031522)	*Cervical neoplasm (HP:0032241)*
Cerebral atrophy (HP:0002059)	*Cerebral degeneration (HP:0007313)*
Chorioretinal degeneration (HP:0200065)	*Retinal degeneration (HP:0000546)*
Coronary artery dissection (HP:0006702)	*Arterial dissection (HP:0005294)*
Carotid artery occlusion (HP:0012474)	*Arterial occlusion (HP:0025324)*
Soft tissue sarcoma (HP:0030448)	*Soft tissue neoplasm (HP:0031459)*
Perifollicular fibroma (HP:0032225)	*Fibroma (HP:0010614)*
Hand muscle weakness (HP:0030237)	*Muscle weakness (HP:0001324)*
Membranous nephropathy (HP:0012578)	*Nephropathy (HP:0000112)*

Table 2 shows the 10 Derived Term Pairs (DTPs) that identified the most number of potential missing IS-A relations in HPO. For instance, the DTP ({*distal*}, {*proximal*}) has detected 321 potential missing IS-A relations.

Table 2. Ten DTPs that detected the most number of potential missing IS-A relations.

DTP (Derived Term Pairs)	Number of potential missing IS-A detected
({*distal*}, {*proximal*})	321
({*distal*}, {*middle*})	246
({*sclerosis*}, {*abnormality*})	23
({*atresia*}, {stenosis})	18
({*epiphysis*}, {})	15
({*absent*}, {*reduced*})	12
({*epiphysis, phalanx, thumb*}, {*hand, phalanges, epiphyses*})	11
({*epiphysis*, *enlarged*}, {*abnormality*})	10
({*2nd, phalanx, toe*}, {*phalanges, toes*})	10
({*phalanx, proximal, of, the*}, {})	10

For the validated missing IS-A relations, Figure 4 shows the distribution of the distances between their mapped UMLS atoms. The distances represent the numbers of edges between the concepts. For instance, there exists only 1 edge between a child concept and a parent concept (direct IS-A relation). However, there exists 2 edges between a grandchild and a grandparent.

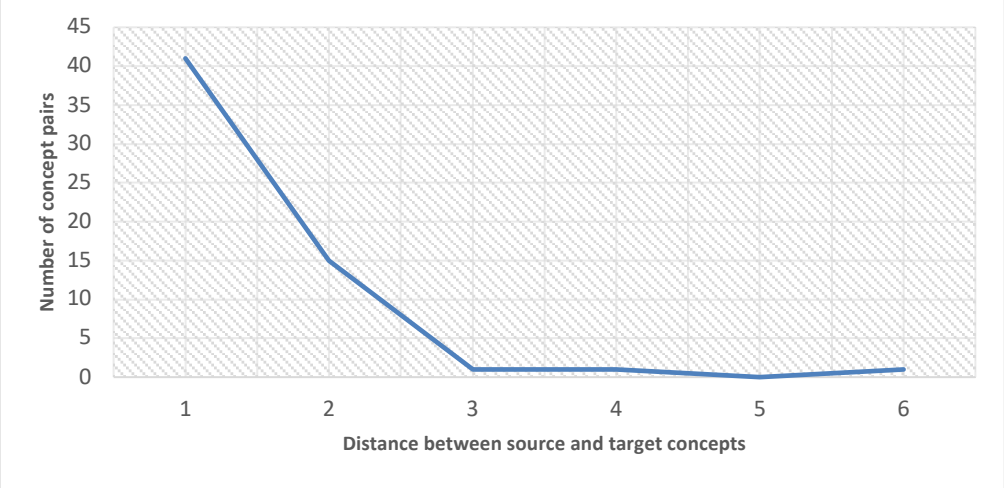

Figure 4. The distribution of the distances between the concept-pairs in validated missing IS-A relations via UMLS.

5 Discussion

In this paper, we explored an automated method to identify potential missing IS-A relations in HPO. Almost all the missing IS-A relations (except one) identified by our approach was in the *Phenotypic Abnormality* subontology of HPO, which is the main subontology of HPO with the most number of concepts. We performed automated evaluation through UMLS with relations in external ontologies and validated 59 missing IS-A relations in HPO. From Figure 4, it can be seen that the most number of validated missing IS-A relations were direct IS-A relations (i.e., distance=1) in their respective ontologies. Direct missing IS-A relations are generally easier to fix than indirect ones as for the latter, intermediate relations that are missing need to be identified. In any case, it should be noted that the fix may not always be the direct addition of the missing relation to HPO but may involve modifying the logical definitions of concepts so that the missing relation can be inferred through reasoning.

5.1 Comparison with related work

In previous work, we leveraged a similar lexical-based inference approach to identify hierarchical relation inconsistencies in the Gene Ontology, SNOMED CT, and National Cancer Institute thesaurus. In that approach, we identified Partial Matching Concept Pairs having the same number of words, at least one word in common and *n* different words (where we investigated *n=1,2,3,4,5*)[18,19]. However, in the current approach, the concept-pair does not have to restrict to the same number of words and can have any number of different words. In addition, in the previous work, the unlinked concept-pairs were not picked from the same subontology, which may lead to suggestions of invalid missing IS-A relations among concept-pairs belonging to different subontologies. The current approach avoids such issue by generating unlinked-pairs within the same subontology. Moreover, in this work we apply our auditing approach on HPO, while in the previous work we experimented with three different ontologies as mentioned above. It should be noted that external investigations into the quality of HPO is really scarce.

5.2 Limitations and future work

In this work, we leveraged UMLS to validate potential missing IS-A relations identified by the approach. The concepts in a missing IS-A relation were mapped to UMLS concepts and the existence of an IS-A relation was verified through the external ontology information provided by UMLS. However, only 59 out of 2,516 missing IS-A suggestions by the method were automatically validated this way. It would be interesting to investigate whether we can leverage external knowledge in biomedical literature to automatically validate additional potential missing IS-A relations. However, both the UMLS-based evaluation and biomedical literature-based evaluation will not be able to identify

false positives generated by our method. Note that false positives generated by the method may actually lead to the identification of erroneous IS-A relations among the linked concept-pairs. In addition, such automatic evaluation can only help validate part of the potential missing IS-A relations identified by the approach. Therefore, to comprehensively assess the effectiveness of the approach, we would like to invite domain experts to manually evaluate a random sample of our results, so that all the potential missing IS-A relations in the sample could be reviewed and evaluated, and false positives and corresponding erroneous IS-A relations (if valid) in the sample could be identified. More specifically, we plan to submit a random sample of potential missing IS-A relations to the HPO issue tracking system[30], so that the HPO curators could further evaluate these missing IS-A relations and make necessary changes to the ontology where appropriate.

In the current work, we only leveraged lexical features of the names of the concepts to identify potential missing IS-A relations. We would like to explore whether incorporating additional features such as ancestor lexical features would improve the method's effectiveness to uncover missing IS-A relations.

Additionally, missing IS-A relations in HPO may affect the quality of its downstream applications and analyses (e.g., making them less accurate). For instance, in an HPO-based cohort search application, missing IS-A relations may reduce the recall of the search result. An interesting future direction is to analyze the impact of the missing IS-A relations on downstream applications. Such an impact analysis will help in assessing the practical significance of the missing IS-A relations identified in this work.

6 Conclusion

In this paper, we presented an automated lexical approach to identify missing IS-A relations in the Human Phenotype Ontology (HPO). We first identified linked and unlinked concept-pairs of HPO. Then we generated Term Pairs (DTPs) based on concept-pairs. If a linked concept-pair and an unlinked concept-pair generate the same DTP, then we suggested a missing IS-A relation between the unlinked concept-pair. Applying this approach on the 2022-02-14 release of HPO, we extracted 2,516 potential missing IS-A relations. Leveraging external ontology information in the UMLS, we were able to validate 59 missing IS-A relations. The results of the automated validation encourage us to further work on a manual evaluation of the results. This work is a first step towards automated methods for identifying and fixing quality issues in HPO.

Acknowledgment
This work was supported in part by the National Science Foundation (NSF) through grant 2047001, National Institutes of Health (NIH) National Library of Medicine through grant R01LM013335, and National Institute on Aging through grant R01AG073435. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

References

1. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Briefings in bioinformatics. 2008;9(1):75-90.
2. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics. 2008;83(5):610-615.
3. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM, Callahan TJ. The human phenotype ontology in 2021. Nucleic acids research. 2021;49(D1):D1207-D1217
4. Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. Journal of biomedical informatics. 2009;42(3):413-425.
5. Halper M, Gu H, Perl Y, Ochs C. Abstraction networks for terminologies: supporting management of "big knowledge". Artificial intelligence in medicine. 2015;64(1):1-6.
6. Campbell, K. E., Tuttle, M. S., & Spackman, K. A. (1998). A "lexically-suggested logical closure" metric for medical terminology maturity. In: Proceedings of the AMIA Symposium 1998. pp. 785-789. American Medical Informatics Association.

7. Agrawal A, Perl Y, Ochs C, Elhanan G. Algorithmic detection of inconsistent modeling among SNOMED CT concepts by combining lexical and structural indicators. In: 2015 IEEE international conference on bioinformatics and biomedicine (BIBM) 2015. pp. 476-483. IEEE.
8. Agrawal A, Revelo P. Analysis of the consistency in the structural modeling of SNOMED CT and CORE problem list concepts. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2017. pp. 292-296. IEEE.
9. Agrawal A, Qazi K. Detecting modeling inconsistencies in SNOMED CT using a machine learning technique. Methods. 2020;179:111-118.
10. Agrawal A. Evaluating lexical similarity and modeling discrepancies in the procedure hierarchy of SNOMED CT. BMC Medical Informatics and Decision Making. 2018;18(4):27-33.
11. Bodenreider O. Identifying Missing Hierarchical Relations in SNOMED CT from Logical Definitions Based on the Lexical Features of Concept Names. ICBO/BioCreative. 2016;2016.
12. Zhang GQ, Xing G, Cui L. An efficient, large-scale, non-lattice-detection algorithm for exhaustive structural auditing of biomedical ontologies. Journal of biomedical informatics. 2018;80:106-119.
13. Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. Journal of the American Medical Informatics Association. 2017;24(4):788-798.
14. Abeysinghe R, Brooks MA, Cui L. Leveraging non-lattice subgraphs to audit hierarchical relations in NCI Thesaurus. In: AMIA annual symposium proceedings 2019. pp. 982-991. American Medical Informatics Association.
15. Hao X, Abeysinghe R, Zheng F, Cui L. Leveraging non-lattice subgraphs for suggestion of new concepts for SNOMED CT. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2021. pp. 1805-1812. IEEE.
16. Cui L, Bodenreider O, Shi J, Zhang GQ. Auditing SNOMED CT hierarchical relations based on lexical features of concepts in non-lattice subgraphs. Journal of biomedical informatics. 2018;78:177-184.
17. Zhang GQ, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. In: AMIA annual symposium proceedings 2010. pp. 922-926. American Medical Informatics Association.
18. Abeysinghe R, Hinderer EW, Moseley HN, Cui L. Auditing subtype inconsistencies among gene ontology concepts. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2017. pp. 1242-1245. IEEE.
19. Abeysinghe R, Zheng F, Hinderer EW, Moseley HN, Cui L. A lexical approach to identifying subtype inconsistencies in biomedical terminologies. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2018. pp. 1982-1989. IEEE.
20. Jean-Baptiste Lamy. Welcome to Owlready2's documentation! [Internet]. 2022 [accessed 22 July 2022]. Available from: https://owlready2.readthedocs.io/en/v0.36/
21. Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic acids research. 2019;47(D1):D330-D338.
22. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004;32(suppl_1):D258-D261.
23. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic acids research. 2007;36(suppl_1):D344-D350.
24. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research. 2016;44(D1):D1214-D1219.
25. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl_1):D267-D270.
26. National Library of Medicine. UMLS 2021AA release available [Internet]. 2022 [accessed 22 July 2022]. Available from: https://www.nlm.nih.gov/pubs/techbull/mj21/mj21_umls_2021aa_release.html
27. UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-. 2, Metathesaurus. [cited 14 July 2022]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9684/
28. Bird, S., Klein, E., & Loper, E. Natural language processing with Python: analyzing text with the natural language toolkit. 2009. O'Reilly Media, Inc.

29. National Library of Medicine (US). Metathesaurus Representation [Internet]. 2022 [accessed 14 July 2022]. Available from: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CPT/metarepresentation.html#relationships
30. Human Phenotype Ontology issue tracker [Internet]. 2022 [accessed 22 July 2022]. Available from: https://github.com/obophenotype/human-phenotype-ontology/issues/new/choose