AMIA Jt Summits Transl Sci Proc, 2017. CRI: p. 384-385.

Investigating the Coverage of Diseases across Biomedical Research and Clinical Ontologies

Satyajeet Raje, PhD, Olivier Bodenreider, MD, PhD
U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Introduction

Different ontologies are used to represent diseases in biomedical research and clinical settings. The Disease Ontology (DO), part of the Open Biomedical Ontologies (OBO) collection, is used in many research projects, whereas SNOMED CT is primarily used in healthcare. Interoperability between these ontologies is critical for translational applications in biomedicine. In this paper, we present the preliminary results of our study investigating the coverage of diseases between these two ontologies.

Background

A. Resources. We worked with the August 2016 OWL release of the ***Disease Ontology*** (downloaded from https://github.com/DiseaseOntology/HumanDiseaseOntology). This version has 6931 active disease concepts. Most concepts in DO have explicit cross-references ("obo:hasDbXref" relations) to concepts from SNOMED CT and other bio-ontologies. With over 300,000 concepts (including about 100,000 disease concepts), ***SNOMED CT*** is the largest clinical terminology in the world (http://www.ihtsdo.org/snomed-ct). We used the March 2016 release of SNOMED CT (US Edition), as this is the version cross-referenced by DO. Unlike DO, SNOMED CT is part of the Unified Medical Language System (***UMLS***, https://uts.nlm.nih.gov/). The UMLS Metathesaurus integrates concepts from many medical terminologies and provides a rich set of synonyms, which can be leveraged for lexical mapping.

B. Related work. Kibbe et. al. recently reported on the overall coverage of the DO and its cross-references to other terminologies [1]. In this study, we perform a deeper analysis of these cross-references to SNOMED CT specifically. Previous studies have analyzed the coverage of concepts within specific subdomains of medicine [2, 3]. For example, Fung et. al. recently assessed coverage of rare diseases in the International Classification of Diseases (ICD) and SNOMED CT [4]. The specific contribution of this work is to identify the overlap and the gaps in the coverage of disease concepts in DO and SNOMED CT. To best of our knowledge, this is the first attempt to study the coverage of DO diseases in SNOMED CT.

Methods

Our assessment of coverage between DO and SNOMED CT can be summarized as follows. We first validate existing mappings; we identify additional mappings lexically; and we characterize the unmapped concepts.

A. Validating explicit mappings. We analyze the semantic compatibility of the explicit mappings to SNOMED CT provided by DO. We first remove mappings to retired SNOMED CT concepts and resolve the mappings to "moved" (remapped) concepts in SNOMED CT. Since all DO concepts represent diseases, we consider "invalid" the mappings to concept outside the "Clinical finding" hierarchy of SNOMED CT. (No further validation is performed here, and we consider "valid" all semantically compatible mappings.) Finally, we characterize the DO concepts that have multiple explicit mappings to identify patterns and propose rules to resolve such mappings.

B. Finding additional lexical mappings. We use a lexical approach to find mappings for DO concepts with no explicit (valid) mappings to SNOMED CT. We first extract the labels for each DO concept, including preferred terms and synonyms. We take advantage of the rich set of synonyms provided by the UMLS to map these terms to UMLS concepts, using exact or normalized string matches. Finally, we identify these mappings to the SNOMED CT concepts associated with these UMLS concepts, but only restricting to concepts in the "Clinical finding" hierarchy of SNOMED CT.

C. Analyzing unmapped concepts. Even after finding additional mappings through lexical match, a number of DO concepts remain without a mapping to SNOMED CT. To characterize these unmapped concepts, i.e., to assess whether isolated concepts or entire subhierarchies are missing from SNOMED CT, we cluster them into groups of hierarchically related DO concepts ("connected components" in graph theory parlance). We further analyze the isolated concepts that do not cluster with any other concepts with their immediate parents and children to characterize these unmapped concepts.

Results

A. Validating explicit mappings. There were 12,470 explicit mappings to SNOMED CT. We removed 224 mappings to retired SNOMED CT concepts and resolved 6552 mappings to moved concepts. A total of 8352

mappings remained, covering 4195 unique DO concepts. Overall, of the 6931 disease concepts in DO, 3859 (56%) were mapped to SNOMED CT through at least one explicit valid mappings provided by DO, 336 (5%) had only invalid mappings, and 2736 (39%) were unmapped. A majority of invalid mappings were mappings of a disease concept in DO to a concept in the "Morphologic abnormality" hierarchy of SNOMED CT.

Of the 3859 DO concepts with at least one valid mapping, 3334 had only valid mappings, while 525 were mapped to both a valid and an invalid concept. Of the 3334 DO concepts with only valid mappings, 1950 had a single mapping (e.g., "Cycloplegia" [DOID:10033] mapped to "Cycloplegia (disorder)" [SCTID:68158006]), and 1384 had multiple valid mappings. Of these, 110 had mapping to concepts in both the "Disease" and the "Clinical Findings" hierarchies, usually to a disease its associated finding. For example, "Mevalonic aciduria" [DOID:0050452] was mapped to "Mevalonic aciduria (disorder)" [SCTID:124327008] and "Hyperimmunoglobulin D with periodic fever (finding)" [SCTID:234538002]. In such cases, we suggest to keep only the mapping to the "Disease" concept. The 525 DO concepts mapped to both a valid and an invalid concept usually correspond to mappings to a disease and its associated morphology. For example, "Acute promyelocytic leukemia (APL)" [DOID: 0060318] was mapped to "APL, FAB M3 (disorder)" [SCTID:110004001] and "APL, t(15;17)(q22;q11-12) (morphologic abnormality)" [SCTID:28950004]. In such cases also, we suggest to keep only the mapping to the "Disease" concept.

B. Finding additional lexical mappings. A total of 3072 DO concepts had no (valid) explicit mapping to SNOMED CT (336 with only invalid mappings and 2736 with no mapping at all). Of these, we mapped 2583 lexically to some UMLS concept, of which 619 mapped to a concept in the "Clinical finding" hierarchy of SNOMED CT. In other words, lexical mapping through the UMLS helped us uncover a (valid) mapping to SNOMED CT for 619 (20%) of the 3072 DO concepts with no (valid) explicit mapping.

C. Analyzing unmapped concepts. Finally, 2453 (36%) DO concepts remained unmapped to SNOMED CT. We grouped these concepts into 1469 clusters of hierarchically related concepts (connected components). On average, each cluster had 1.67 concepts. The largest cluster contained 263 concepts, while 1262 concepts were isolated. Examples of large clusters include those rooted by the DO concepts "organ system benign neoplasm" [DOID:0060085] (263 concepts), "monogenic disease" [DOID:0050177] (35 concepts), and "chromosomal deletion syndrome" [DOID:0060388] (35 concepts).

Of the 1262 isolated concepts, 1184 were leaf concepts (i.e., fine-grained concepts in DO), while 78 were non-leaf concepts (i.e., intermediary grouper concepts in DO). An example of such unmapped leaf concept is "multiple mucosal neuroma" [DOID:5155]. Of note, this concept is the only unmapped child of "neuroma" [DOID:2001], which means that its sibling, including "Neurilemmoma" [DOID:3192], are all mapped to SNOMED CT. Finally, examples of unmapped intermediary grouper concepts include "multifocal dystonia" [DOID:0050837]. This concept is unmapped to SNOMED CT, while its parent, "dystonia" [DOID:543], and child, "hemidystonia" [DOID:0050846], are both mapped to SNOMED CT.

Conclusion

This preliminary analysis reveals important gaps in the coverage of diseases from the Disease Ontology in SNOMED CT. In future work we will further analyze the DO concepts with multiple valid mappings to SNOMED CT and investigate the differences in the hierarchical organization of disease concepts between the two terminologies. We expect this analysis to provide insights for bridging the gaps between the two ontologies.

Acknowledgements

This work is supported by the intramural finding and in part by the Intramural Research Program of the U.S. National Library of Medicine (NLM) and an appointment to the NLM Research Participation Program administered by ORISE through an interagency agreement between the U.S Dept. of Energy and the NLM.

References

1. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res 2015;43(Database issue):D1071-8.
2. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. J Am Med Inform Assoc 1996;3(3):224-33.
3. Dhombres F, Winnenburg R, Case JT, Bodenreider O. Extending the coverage of phenotypes in SNOMED CT through post-coordination. Stud Health Technol Inform 2015;216:795-9.
4. Fung KW, Richesson R, Bodenreider O. Coverage of rare disease names in standard terminologies and implications for patients, providers, and research. AMIA Annu Symp Proc 2014;2014:564-72.