MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-863

Perceiving the Usefulness of the National Cancer Institute Metathesaurus for Enriching NCIt with Topological Patterns

Zhe He[a], Yan Chen[b], James Geller[c]

[a] *School of Information, Florida State University, Tallahassee, FL, USA*
[b] *Department of Computer Information Systems, BMCC, City University of New York, New York, USA*
[c] *Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA*

Abstract

The National Cancer Institute Thesaurus (NCIt), developed and maintained by the National Cancer Institute, is an important reference terminology in the cancer domain. As a controlled terminology needs to continuously incorporate new concepts to enrich its conceptual content, automated and semi-automated methods for identifying potential new concepts are in high demand. We have previously developed a topological-pattern-based method for identifying new concepts in a controlled terminology to enrich another terminology, using the UMLS Metathesaurus. In this work, we utilize this method with the National Cancer Institute Metathesaurus to identify new concepts for NCIt. While previous work was only oriented towards identifying candidate import concepts for human review, we are now also adding an algorithmic method to evaluate candidate concepts and reject a well defined group of them.

Keywords:
Biomedical Ontologies; Vocabulary; Quality Assurance

Introduction

Biomedical ontologies, terminologies and controlled vocabularies are widely used in a variety of healthcare information systems for encoding healthcare data such as diagnoses, laboratory test results, patient-reported problem lists, and billing statements. They are also used to faciliate knowledge management, data integration, decision support, and biomedical natural language processing. The development and curation of biomedical ontologies are mostly driven by ontology engineers and subject matter experts. As domain completeness is one of the key properties of a biomedical ontology [1], new concepts need to be included as they are needed by the users. As manual curation is costly and time consuming, automated or semi-automated methods for identifying new concepts that are relevant to the domain of an ontology are therefore in high demand.

The National Cancer Institute (NCI) Thesaurus (NCIt) is an important reference terminology in the cancer domain. It currently contains over 100,000 concepts that are hierarchically organized into 19 distinct hierarchies relevant to cancer research, such as *neoplastic diseases* and *molecular abnormalities*. NCIt is a central reference terminology of NCI's Enterprise Vocabulary Services (EVS) [2]. The EVS leverages both an internal quality assurance (QA) team and external participation in the development and QA of NCIt. Outside contributors can suggest new concepts or terms for NCIt, which will be reviewed, validated, and incorporated into it, based on NCI's content development and editing guidelines.

In previous work [3; 4], we have developed a topological-pattern-based method to demonstrate the vertical density differences across pairs of source terminologies in the Unified Medical Language System (UMLS), the most comprehensive biomedical terminological system in existence, developed by the U.S. National Library of Medicine [5]. Leveraging the topological patterns in the UMLS, we have identified potential new concepts for SNOMED CT [3; 4]. Figure 1 illustrates the simplest case of a topological pattern, called a *k*:1 trapezoid where *k*=2. In this case, both Terminology 1 (T1) and Terminology 2 (T2) contain Concept A and Concept B. In other words, Concept A and Concept B in T1 and T2 have the same UMLS Concept Unique Identifier (CUI). There is one intermediate concept X between Concept A and Concept B in T1 but no intermediate concept in T2. Upward pointing arrows indicate IS-A links.

One can argue that Concept X may be missing in T2. However, the intermediate concept(s) from T1 may not be needed in T2 according to human judgment. If $k > 2$, there is more than one intermediate concept in T1. In this paper, T2 is always NCIt. The final decision whether the intermediate concepts should be included in NCIt or not always has to be made by its curators.

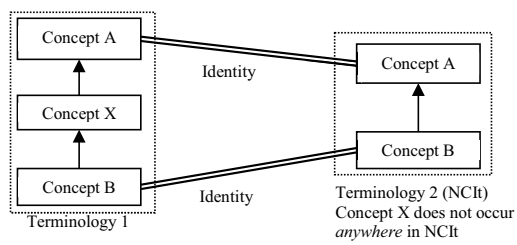

Figure 1 – A hypothetical 2:1 trapezoid between Terminology 1 and Terminology 2 (NCIt)

Recently, we applied this method to identify concepts in the UMLS that could enrich NCIt [6]. The National Cancer Institute Metathesaurus (NCIm) is a terminological system with the same structure as the UMLS, but with more cancer related content [7]. Therefore, a natural question arises – "*Is NCIm a better source for utilizing topological patterns to identify new concepts for NCIt than the UMLS Metathesaurus?*" Because NCIm and UMLS have much overlap, we were especially interested in concepts suggested for import by topological patterns that exist only in the UMLS

but not in the NCIm and vice versa. The latter case would indicate a new source for candidate concepts compared to previous research.

One difficulty with our previous approach is that all evaluations of candidate concepts have to be performed by a human expert. In this paper, we were investigating ideas for algorithmically rejecting some of the candidates, reducing the work load for the human expert. We formulated the following hypotheses about reasons for missing intermediate concepts in *k*:1 trapezoids in NCIt:

- The parent concept (A in Figure 1) has only one child (B in Figure 1) in NCIt. Therfore, an intermediate concept is not needed to organize the conceptual content.
- The intermediate concept is a synonym of the parent concept (A in Figure 1) or the child (B in Figure 1) in NCIt.
- The curators of T1 and NCIt made different modeling decisions.

The contributions of this work are twofold: 1) We compare the usefulness of the NCIm and UMLS Metathesaurus as data sources for using topological-pattern-based methods to identify new concepts for NCIt; 2) We propose a practical method to algorithmically reject some concepts for import, thereby reducing the work load of the human expert.

Background

UMLS Metathesaurus (Meta)

The UMLS Metathesaurus (Meta), which integrates over 190 biomedical terminologies, is the most comprehensive biomedical terminological system in existence [5]. It maps over 12 million terms into over 3.2 million concepts.

NCI Metathesaurus (NCIm)

The NCI Metathesaurus (NCIm) is a wide-ranging biomedical terminological system that covers most terminologies used by NCI for clinical care, translational and basic research, public information and administrative activities [7]. It maps about four million terms from more than 75 source vocabularies into two million concepts that represent their meaning. Importantly, it has a monthly update cycle, ensuring the timely update of its core terminology, NCIt. It covers most of the public domain terminologies of the NLM's UMLS Metathesaurus as well as many other biomedical terminologies created by or of interest to NCI and its partners such as RadLex. Some vocabularies such as SNOMED CT were also shrunk in the NCIm.

Methods

Identifying Candiate Terminologies

In this work, we compared the effectiveness of using NCIm versus UMLS Meta to identify potentially useful concepts for enriching the conceptual content of NCIt. We used the August 2015 version of the NCIm and the 2015AA version of the UMLS Meta. Both NCIm and UMLS 2015AA contain SNOMED CT US (March 31, 2015 version). The NCIt version in NCIm is 2015 08E, whereas the NCIt version in the UMLS is 2014 03E. The main criteria for selecting a candidate terminology for this research include: 1) the terminology must be in English; 2) the terminology must be organized with an IS-A hierarchy backbone; 3) the terminology must exhibit sufficient overlap in content with NCIt; and 4) the terminology must exist in both UMLS Meta and NCIm. We first identified seven English source terminologies with "PAR" (parent-child) relationships and "INVERSE_IS_A" relationship attributes, including SNOMED CT, Foundational Model of Anatomy Ontology (FMA), Universal Medical Device Nomenclature System (UMD), RadLex (a radiology lexicon), University of Washington Digital Anatomist (UWDA), MGED Ontology (MGED), and Gene Ontology (GO). Out of these seven terminologies, RadLex and MGED are not in the UMLS Meta. UWDA is part of FMA and was therefore excluded.

Identifying and Analyzing *K*:1 Trapezoids

We first identified all the *k*:1 trapezoids in NCIm and UMLS. The NCIm, which is based on the UMLS, may contain cycles of IS-A relationships [8]. We eliminated the cycles in the trapezoid by detecting repeating Concept Unique Identifier (CUIs) in the IS-A paths.

After we identified all the *k*:1 trapezoids in the NCIm and the UMLS, we calculated the number of trapezoids for each kind and the number of intermediate concepts in T1. Note that multiple parents may lead to overlapping trapezoids with the same intermediate concept. We eliminated duplicate intermediate concepts in the results.

Manual Review of the 2:1 Trapezoid Samples

To compare the utility of the topological patterns in the NCIm and the UMLS Meta for identifying new concepts for enriching NCIt, we took a random sample of 50 2:1 trapezoids between SNOMED CT and NCIt that can be found in the UMLS but not in the NCIm (Sample 1), as well as a random sample of 50 2:1 trapezoids between SNOMED CT and NCIt that can be found in the NCIm but not in the UMLS (Sample 2). We combined the two samples and randomized the order. The terminology expert (YC) investigated the content of both SNOMED CT and NCIt using the Neighborhood Auditing Tool (NAT) for the UMLS [9], and assessed whether the intermediate concepts in SNOMED CT should be suggested for inclusion in NCIt or not. The terminology expert chose one of the following three options: 1) the intermediate concept (Concept X in Figure 1) in T1 should be imported into NCIt; 2) the intermediate concept should not be imported into NCIt; and 3) the intermediate concept *may* be imported to NCIt. For the options 2) and 3), the terminology expert was also asked to give rationales for making such a choice. The NAT tool (http://nat.njit.edu/) allows an auditor to concentrate on a single focus concept and its neighborhood (i.e., parents, children, siblings), thereby well meeting the need of this study. We will report the manual review results and the reasons for options 2) and 3) in the Results section.

Identiying More Complex Topological Patterns

There could also be one or more intermediate concepts in T2 in a topological pattern. We therefore defined *M*:*N trapezoids* as topological patterns in which both T1 and T2 have both Concept A and Concept B, but there are *M*-1 intermediate concepts between A and B in T1 and *N*-1 intermediate concepts between A and B in T2. The intermediate concepts in T1 do not appear anywhere in T2 and vice versa. *M:N* trapezoids are a generalization of *k:1* trapezoids. The relationships among intermediate concepts in *M*:*N* can be categorized into the following three types: 1) an intermediate concept in T1 can be a parent/child of an intermediate concept

in T2; 2) an intermediate concept in T1 can be a synonym of an intermediate concept in T2; 3) T1 and T2 have alternative classifications, indicating two different ways of conceptualizing a domain that are both valid but not immediately compatible [3]. A trapezoid may also indicate an error in one of the two terminologies. In a recent publication [10], we provided an estimate of the difficulty faced by a domain expert in a concept import task. In this paper, we merely identified the M:N trapezoids (M >= 2, N >= 2) in both the NCIm and the UMLS. The analysis of the relationships among intermediate concepts in *M:N* trapezoids is beyond the scope of this work.

Results

2:1 Trapezoids in the NCIm and the UMLS Meta

Table 1 shows the number of 2:1 trapezoids identified in the NCIm and the UMLS Meta. As shown in Table 1, notably fewer 2:1 trapezoids between SNOMED CT and NCIt were identified in NCIm than in the UMLS. In these 2:1 trapezoids, 1,019 distinct intermediate concepts can be found in both NCIm and the UMLS; 890 distinct intermediate concepts were found in the UMLS but not in the NCIm; and 174 intermediate concepts in the NCIm were not in the UMLS. This may be due to the fact that the NCIm contains a newer version of NCIt than the UMLS Meta. Only a small number of trapezoids could be found between terminologies other than SNOMED CT and NCIt. In the subsequent analysis, we will focus on the trapezoids between SNOMED CT and NCIt.

Table 1 – Number of 2:1 Trapezoids Identified in NCIm and UMLS Meta

Candidate Terminology	NCIm		UMLS Meta	
	# of Trapezoids	# of Intermediate Concepts	# of 2:1 Trapezoids	# of Intermediate Concepts
SNOMED CT	2,308	1,193	3,894	1,909
FMA	115	55	112	55
GO	57	38	54	37
UMD	2	2	1	1

Figure 2 shows the histogram of semantic types of 890 intermediate SNOMED CT concepts that were identified in the 2:1 trapezoids in the UMLS Meta but not in the NCIm. Semantic types with fewer than 10 concepts are not shown in the figure.

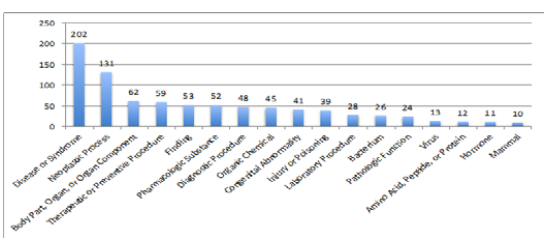

Figure 2 – The Semantic Types of 890 Intermediate SNOMED CT Concepts in the 2:1 Trapezoids in the UMLS, but not in the NCIm.

Figure 3 shows the histogram of semantic types of 174 intermediate SNOMED CT concepts that were identified in the 2:1 trapezoids in the NCIm but not in the UMLS. All the semantic types except for ENZYME also appear in Figure 2.

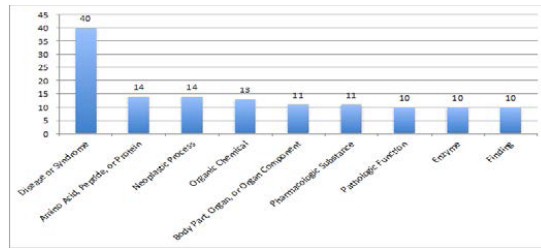

Figure 3 – The Semantic Types of 174 Intermediate SNOMED CT Concepts in the 2:1 Trapezoids in the NCIm, but not in the UMLS.

K:1 Trapezoids in NCIm and the UMLS Meta

Table 2 shows the number of trapezoids and the number of unique intermediate concepts identified in *k*:1 (*k* > 2) trapezoids between SNOMED CT and NCIt in both NCIm and the UMLS Meta. As the value of *k* increases, the difference between NCIm and UMLS becomes larger, indicating that NCIm has significantly downsized SNOMED CT and removed concepts that are unnecessarily granular for the cancer domain.

Table 2 – Number of Trapezoids and Unique Intermediate Concepts from SNOMED CT Identified in k:1 Trapezoids (k >2) in NCIm and the UMLS

Kind	NCIm		UMLS Meta	
	# of Trapezoids	# of unique intermediate concepts	# of Trapezoids	# of unique intermediate concepts
3:1	967	876	2,386	1945
4:1	356	532	1,686	1495
5:1	256	462	1,509	1143
6:1	209	373	1,357	1001
7:1	82	172	1,032	687
8:1	15	44	700	424
9:1	3	20	461	261

Manual Review of the Samples of 2:1 Trapezoids

According to the review results of YC, in Sample 1 (2:1 trapezoids found in the UMLS but not in the NCIm), 20 intermediate concepts (40%) should be imported into NCIt, 27 intermediate concepts (54%) should not be imported into NCIt, and 3 intermediate concepts (6%) may be imported into NCIt. In Sample 2 (2:1 trapezoids found in the NCIm but not in the UMLS), 20 intermediate concepts (40%) should be imported into NCIt, 23 intermediate concepts (46%) should not be imported into NCIt, and 7 intermediate concepts (14%) may be imported into NCIt. For the cases in which the intermediate concepts should not be imported, the major reasons for rejecting the import are: 1) The term of the intermediate concept is a synonym of the parent/child concept in NCIt; 2) the parent concept has a single child in NCIt; it was felt that creating a structure of one child with one grandchild contradicts the idea of hierarchically organizing concepts in an ontology into groups of similar concepts; and 3) NCIt and SNOMED CT are using two different categorizations. For example, SNOMED CT models the concepts by sites but NCIt does not.

For the cases of possible import, the rationale is that SNOMED CT is more granular than NCIt, i.e., the parent concept has more children/descendants in SNOMED CT than in NCIt. We list the review results of the mixed sample in Table 3. Even though both samples have the same percentage of trapezoids that can contribute concepts to NCIt, Sample 2 has a higher percentage of intermediate concepts that *may* be imported into NCIt than Sample 1 (14% vs. 6%).

Table 3 – Results of Sample Review by the Terminology Expert

Class	Reason	Sample 1 (%)	Sample 2 (%)
Should be imported	--	20 (40%)	20 (40%)
Should not be imported	1) Synonyms	5 (10%)	4 (8%)
	2) Single child	8 (16%)	12 (24%)
	3) Different categorizations	14 (28%)	7 (14%)
May be imported	--	3 (6%)	7 (14%)

Figure 4 illustrates a case of reason 1). In this case, the term of the intermediate concept *Blood Cell Count* is a synonym of the concept *Complete Blood Count* (C0009555) in NCIt. Therefore, there is no need to add the intermediate concept in SNOMED CT to NCIt.

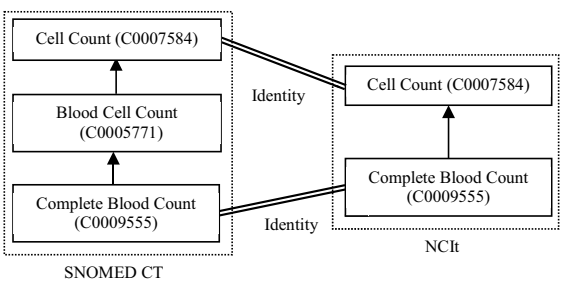

Figure 4 – A 2:1 trapezoid between SNOMED CT and NCIt in the UMLS in which the term of the intermediate concept in SNOMED CT is a synonym of the child concept in NCIt.

Figure 5 illustrates a case of reason 2) in which the parent concept *Retinitis* (C0035333) has only one child - *Cytomegaloviral Retinitis* (C0206178) in NCIt. It is therefore not recommended to add an intermediate concept in NCIt.

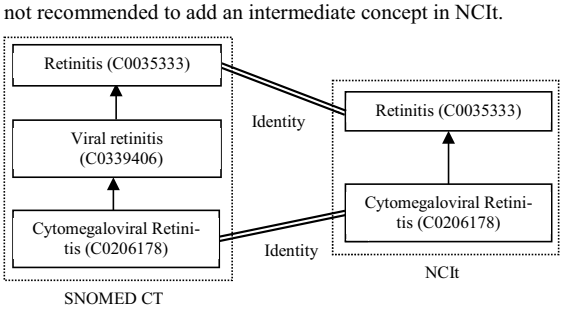

Figure 5 – A 2:1 trapezoid between SNOMED CT and NCIt in the NCIm, in which the parent concept Retinitis has only one child in NCIt.

Figure 6 illustrates a case of reason 3) for rejecting a concept import. In this 2:1 trapezoid, the intermediate concept in SNOMED CT is modeled by sites but NCIt does not follow the same design. Thus, the intermediate concept should not be imported.

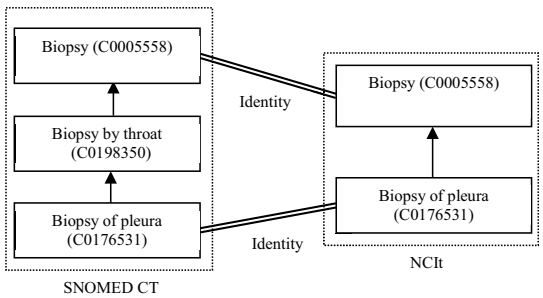

Figure 6 – A 2:1 trapezoid between SNOMED CT and NCIt in the NCIm in which the intermdiate concept is modeled by site.

Figure 7 illustrates a 2:1 trapezoid in which the intermediate concept *Digestive System Disorder* in SNOMED CT should be imported into NCIt.

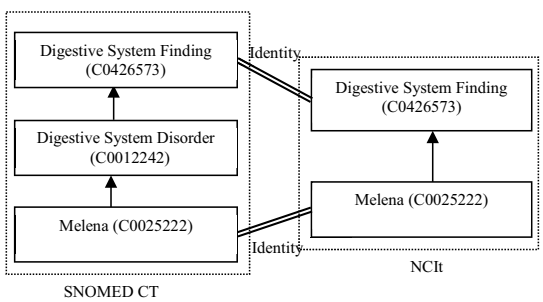

Figure 7 – A 2:1 trapezoid between SNOMED CT and NCIt in the NCIm in which the intermdiate concept should be imported into NCIt.

Automatic Rejection of Import from K:1 Trapezoids

As discussed in the manual review section, one of the major reasons why an intermediate concept in a 2:1 trapezoid should not be imported into NCIt is that the parent concept in the trapezoid has only one child in NCIt. To assess the scale of this problem, we investigated all the *k*:1 trapezoids between SNOMED CT and NCIt in both the UMLS and NCIm to find the percentage of parent concepts in the *k*:1 trapezoids with only one child in NCIt. As can be seen in Table 4, 8.7% of parent concepts in 2:1 trapezoids in the UMLS have only one child in NCIt, whereas 7.5% of parent concepts in 2:1 trapezoids in the NCIm have only one child in NCIt. As *k* increases, the percentage drops fast. About 5% of intermediate concepts identified in 2:1 trapezoids with such a structure can be automatically rejected for import. Another reason for rejecting an import is that the term of the intermediate concept is a synonym of the parent/child concept in NCIt. However, it is not easy to fully automate the detection of such cases because certain judgments of synonymy need to be made by human experts.

***M*:*N* Trapezoids in NCIm and the UMLS Meta**

Table 5 shows the number of *M*:*N* trapezoids (*M*, *N* >= 2) between SNOMED CT and NCIt in the NCIm and the UMLS, respectively. Due to the space limit, we only show the

numbers of 2:*N* and 3:*N* trapezoids. A smaller number of *M*:*N* trapezoids was found in the NCIm than in the UMLS. Whether a higher percentage of trapezoids is useful for concept enrichment or not warrants further investigation.

Table 4 – Automatic Rejection of Concept Import When the Parent Concept has a Single Child in NCIt

Kind	# of Parent Concepts in UMLS with Single Child in NCIt	# of intermed-iate concepts that can be rejected	# of Parent Concepts in NCIm with Single Child in NCIt	# of intermediat e concepts that can be rejected
2:1	84 (8.7%)	89 (4.7%)	55 (7.5%)	57 (4.8%)
3:1	24 (5.6%)	50 (2.6%)	13 (5.2%)	30 (3.4%)
4:1	6 (3.1%)	21 (1.4%)	2 (0.2%)	6 (1.1%)
5:1	0 (0%)	0 (0%)	1 (0.2%)	8 (1.7%)
6:1	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Table 5 – Number of M:N trapezoids (M, N >=2) identified between SNOMED CT and NCIt in NCIm and the UMLS

Kind	# *of* Trapezoids		Kind	# of Trapezoids	
	NCIm	UMLS		NCIm	UMLS
2:2	688	1097	3:2	492	841
2:3	296	464	3:3	333	825
2:4	77	170	3:4	170	364
2:5	61	65	3:5	82	189
2:6	32	30	3:6	44	98
2:7	3	3	3:7	12	28
			3:8	1	5
			3:9	2	2
			3:10	1	1

Discussion and Conclusions

In this work, we compared the usefulness of the NCIm and the UMLS Metathesaurus for finding new concepts for NCIt using topological patterns. We found that even when limiting ourselves to 2:1 trapezoids the NCIm can possibly contribute 1,193 concepts, on the same order of magnitude as the UMLS Metathesaurus, 1,909. We further developed a method to automatically reject *k*:1 trapezoids for concept imports in which the parent concept has only one child in NCIt. For the UMLS Metathesaurus 4.7% intermediate concepts can be automatically rejected, while for the NCIm it is 4.8% based on 2:1 trapezoids alone.

The automatic rejection scenario may be justified in the following manner. When looking at an ontology from the bottom up, as opposed to a top down view, the fundamental idea of the IS-A hierarchy in an ontology is to organize the most specific concepts into groups so that group members are more closely related to each other than to concepts from different groups. Thus, the ontology functions as a way to organize concepts in a way that reflects the real world. This process is then repeated at higher levels, so that groups are themselves grouped together into larger groups of concepts for which the same rule holds, group members are more similar to each other than to members of other groups. While nature might force us to accept groups of size "1" at the bottom level of the ontology, there is no reason to create artificial groups containing only one concept that are themselves alone in their groups.

For the *k*:1 trapezoid cases where the intermediate concepts were deemed to be possible imports into NCIt, the final decision will depend on the NCIt curators at two levels. First the curators will have to decide whether this topic area in NCIt is sparse on purpose, or whether it is only sparse because of lack of time and budget. Then the curators will have to decide about every concept individually whether it is desirable for import into NCIt.

A few limitations need to be noted. The version of NCIt in NCIm (2015 08E) is different from the one in the UMLS. Because the NCIt version in the UMLS always falls behind the NCIt version in the NCIm, we were not able to find NCIm and UMLS Metathesaurus releases with the same version of NCIt and SNOMED CT. Nevertheless, both the 2015AA release of the UMLS and the August 2015 version of NCIt contain the SNOMED CT US March 31, 2015 version. In future work, we plan to develop a more robust method that leverages more sophisticated topological patterns to recommend new concepts for NCIt and reject inapprorpriate ones.

Acknowledgements

This work was partially supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

References

[1] J.J. Cimino, Desiderata for controlled medical vocabularies in the twenty-first century, *Methods Inf Med* 37 (1998), 394-403.

[2] S. de Coronado, L.W. Wright, G. Fragoso, M.W. Haber, E.A. Hahn-Dantona, F.W. Hartel, S.L. Quan, T. Safran, N. Thomas, and L. Whiteman, The NCI Thesaurus quality assurance life cycle, *J Biomed Inform* 42 (2009), 530-539.

[3] Z. He, J. Geller, and Y. Chen, A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization, *Artif Intell Med* 64 (2015), 29-40.

[4] Z. He, J. Geller, and G. Elhanan, Categorizing the Relationships between Structurally Congruent Concepts from Pairs of Terminologies for Semantic Harmonization, *AMIA Jt Summits Transl Sci Proc* 2014 (2014), 48-53.

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

[6] Z. He, Y. Chen, S. de Coronado, et al., Topological-Pattern-based Recommendation of UMLS Concepts for National Cancer Institute Thesaurus, *AMIA Annu Symp Proc* 2016 (2016), 618-627.

[7] NCI, Homepage of NCI Metathesaurus, in. https://ncim.nci.nih.gov/ncimbrowser/

[8] O. Bodenreider, Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention, *Proc AMIA Symp* (2001), 57-61.

[9] C.P. Morrey, J. Geller, et al., The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS, *J Biomed Inform* 42 (2009), 468-489.

[10] Z. He and J. Geller, Preliminary Analysis of Difficulty of Important Pattern-Based Concepts into the National Cancer Institute Thesaurus, *Stud Health Technol Inform* 288 (2016), 389-393.

Address for correspondence:

Zhe He, PhD.
School of Information, Florida State University
142 Collegiate Loop
Tallahassee FL, 32306

Email: zhe.he@cci.fsu.edu