Topological-Pattern-Based Recommendation of UMLS Concepts for National Cancer Institute Thesaurus

Zhe He, PhD[1,2], Yan Chen, PhD[3], Sherri de Coronado, MS, MBA[4], Katrina Piskorski, MD[5], James Geller, PhD[6]

[1]School of Information, Florida State University, Tallahassee, FL; [2]Institute for Successful Longevity, Florida State University, Tallahassee, FL; [3]Department of Computer Information Systems, Borough of Manhattan Community College, City University of New York, New York, NY; [4]National Cancer Institute, Rockville, MD; [5]Weill Cornell Medicine, New York, NY; [6]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ

Abstract

The National Cancer Institute Thesaurus (NCIt) is a reference terminology used to support clinical, translational and basic research as well as administrative activities. As medical knowledge evolves, concepts that might be missing from a particular needed subdomain are regularly added to the NCIt. However, terminology development is known to be labor-intensive and error-prone. Therefore, cost-effective semi-automated methods for identifying potentially missing concepts would be useful to terminology curators. Previously, we have developed a structural method leveraging the native term mappings of the Unified Medical Language System to identify potential concepts in several of its source vocabularies to enrich the SNOMED CT. In this paper, we tested an analogous method for NCIt. Concepts from eight UMLS source terminologies were identified as possibilities to enrich NCIt's conceptual content.

Introduction

Biomedical ontologies and controlled terminologies provide a solid foundation in a variety of healthcare information systems [1, 2]. They have been widely used for encoding diagnoses, laboratory tests, and problem lists in Electronic Health Records [3] as well as in administrative documents such as in billing statements [4]. Moreover, with medical concepts linked by taxonomic and semantic (lateral) relationships, they also play an important role in knowledge management, data integration, and decision support [1]. Complicated and challenging natural language processing tasks also benefit from well-curated domain ontologies and controlled terminologies.

The National Cancer Institute thesaurus (NCIt), accessible through a browser at https://ncit.nci.nih.gov/ncitbrowser/ is a reference terminology developed and maintained by the National Cancer Institute (NCI). Currently, it contains over 100,000 concepts that are hierarchically organized in 19 distinct domains related to cancer research, e.g., neoplastic diseases, molecular abnormalities, and genes. It is a central reference terminology of NCI's Enterprise Vocabulary Services (EVS) [5]. As medical terminology is constantly evolving, with new concepts in healthcare entering usage, a controlled terminology needs to keep improving its conceptual content to encode these new concepts *as they are needed by users*. In Cimino's "desiderata" for controlled medical vocabularies [6], domain completeness is listed as the most desirable property. To improve the conceptual content of NCIt, NCI EVS exploits both its internal quality assurance (QA) mechanisms as well as external participation in the development and QA process of NCIt. A contributor outside of the NCI can suggest new needed terms or larger sets of concepts for NCIt that will be subsequently reviewed by EVS, validated and developed in conformance with NCIt content development and editing guidelines. The monthly update cycle of NCIt also ensures the timely incorporation of new terms.

In previous research, we have introduced a structural methodology to recommend new concepts from a UMLS source vocabulary for inclusion in another source vocabulary where they are "missing" [7, 8]. This algorithmic structural method explores *vertical density differences* between pairs of terminologies in the Unified Medical Language System (UMLS). The method consists of recognizing trapezoids in the "parent of" relationship structures of terminologies, that is, recognizing cases where concept pairs are present in two terminologies but each terminology offers different intermediate concepts along their "parent of" paths of relationships. It leverages the native term mappings of the UMLS to identify topological patterns that are indicative of a possible import of a concept from one terminology into another terminology. Examples of topological patterns will be shown in the Background and Methods sections.

A variety of vertical topological patterns ('trapezoids' of different sizes) were used in previous research to identify a list of candidate concepts for import into SNOMED CT. Human domain experts confirmed the validity of this structural method for enriching the conceptual content of SNOMED CT [7, 8]. In this paper, we apply this topological pattern-based method to recommend new concepts for inclusion in the NCIt. Just as in our previous research, it was hypothesized that a structural difference between two terminologies might suggest a different course of action besides an import, e.g., it might lead to uncovering an error in one of the two source terminologies. Different possibilities will be exhaustively enumerated in this paper.

Background

Quality assurance (QA) of the NCIt has been conducted by both NCI and external researchers [5]. Min et al. constructed an area taxonomy and a partial-area taxonomy for the NCIt that highlighted potential errors for manual review by a human expert [9]. Cohen et al. performed an automated comparative audit of the gene hierarchy of NCIt using the Entrez Gene database of the National Center for Biotechnology Information [10]. More recently, Semantic Web technologies have been leveraged to audit the NCIt. Mougin and Bodenreider stored the NCIt concepts in an RDF triple store to assess the consistency of the hierarchical and associative relationships among them [11]. Jiang et al. evaluated the data quality of cancer study common data elements by integrating the NCI Cancer Data Standards Repository, NCIt, and the UMLS Semantic Network with the use of a variety of tools of the Semantic Web [12].

The UMLS is a medical terminological system. Its Metathesaurus integrates over 12 million terms from more than 190 source vocabularies into about 3.25 million concepts, such that terms with the same meaning are assigned the same Concept Unique Identifier (CUI). Due to its large scale, source integration and term mapping are challenging tasks. To aid UMLS source integration, Huang et al. developed an extrinsic method that uses WordNet synonym substitution and showed promising results [13]. Moreover, the syntactic patterns and the semantics of the UMLS have been exploited, supporting ontology alignment for OBO Foundry ontologies [14].

The use of topological patterns was introduced in our previous research [7, 8] and will now be explicated based on "*k*:1" trapezoids. Figure 1 shows the topological pattern of a *k*:1 trapezoid with *k*=2. The instances of the concept A have the same UMLS CUI (Concept Unique Identifier) in both terminologies, which means that the UMLS curators regarded them as the same concept. The same is true for the concept B. The identity of the "two" concepts A in the two terminologies is hinted at by the double line connecting them that reminds of the = symbol. However, Terminology 1 has an additional concept X located on a path from B to A. In our previous work, SNOMED CT was the terminology of interest. Encouraged by the success with SNOMED CT, we are taking steps towards establishing the generality of the method by using a different target terminology, namely the NCIt. Therefore, we will formulate our explanation in terms of the NCIt.

Looking at Figure 1, one can argue that in the limited scope of paths from B to A, Terminology 1 contains more information than the NCIt, by including the additional concept X. The concept X is not just missing on the path from B to A in NCIt, but we ascertain that X does not appear anywhere in NCIt before we consider importing it. The topological pattern illustrated in Figure 1 is referred to as 2:1 trapezoid, because there are two parent-child links on the Terminology 1 side and one parent-child link on the NCIt side with links of equal length connecting A with A and B with B.

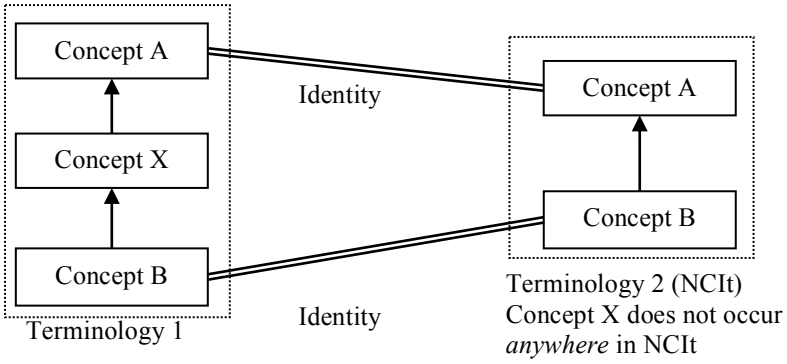

Figure 1. Abstraction of a 2:1 trapezoid between Terminology 1 and Terminology 2 (NCIt). Should X be imported into the NCIt?

Methods

In this work, we use the term "candidate terminology" to refer to any selected UMLS source terminology that could potentially contribute concepts to NCIt (version 2014_03E). Candidate terminologies were determined as follows. We first identified all English source terminologies with "PAR" (Parent) relationships annotated with "INVERSE_IS_A" labels that overlap the content of NCIt. The rationale for excluding "RB" (Broader) relationships in this study is based on the fact that "PAR" represents an explicitly defined parent-child relationship in the source, whereas "RB" represents an implicit one, inserted by the UMLS editorial team [15]. In the UMLS 2015AA release, we identified CPM, CPT, FMA, GO, MEDCIN, SNOMEDCT_VET, ATC, and UMD as the candidate terminologies. Table 1 lists the versions and full names of these terminologies. In previous research [7, 8], we have used this method to identify potential concepts in NCIt to enrich SNOMED CT, and vice versa. Thus, in this work, we excluded SNOMED CT from the list of candidate terminologies even though it also has "PAR" relationships annotated with "INVERSE_IS_A" labels.

Table 1. The list of the candidate terminologies, their versions, and their abbreviated names.

Terminology	Version	Abbreviation
MEDCIN	December 17, 2014	MEDCIN
Foundational Model of Anatomy Ontology	3.1	FMA
Gene ontology	May 19, 2014	GO
Medical Entity Dictionary	2003	CPM
Universal Medical Device Nomenclature System	2015	UMD
Current Procedural Terminology	2015	CPT
Anatomical Therapeutic Chemical classification system	March 2, 2015	ATC
SNOMED CT Veteran Extension	October 20, 2014	SNOMEDCT_VET

The topological patterns considered in this paper are k:1 (k = 2, 3, 4, …) trapezoids, 3:2 and 2:3 trapezoids as well as 2-rectangles. The definitions of these topological patterns are given below.

K:1 Trapezoids

In this study, Terminology 1 is always one of the candidate terminologies listed in Table 1. Discovery of a 2:1 trapezoid suggests that the concept X could be imported into the NCIt. It has been our experience that the owners of several different terminologies have expressed reasons why not to include such new concepts. The main justifications were that "*the concepts describe intermediate, non-coding terms or alternative classifications that would not add substantial value*" and "*we cannot include everything, if we start thinking about what we* could *include, it never ends*." Thus the final decision whether "X" should be included in NCIt always has to be made by its curators. An *alternative classification* recognizes the fact that not every concept in one terminology is valid for another, because the two terminologies might have different purposes, or because the concept is part of a different classification schema in the two ontologies. A technical explanation of alternative classification is given below.

We implemented a program that generates all possible k:1 (k = 2, 3, 4, …) trapezoids for all terminologies from Table 1 with the following termination condition for the size of k. For a given terminology T from Table 1, if no k_0:1 trapezoid is found, then the algorithm checks whether a (k_0+1):1 trapezoid exists. If the answer is "no," the algorithm terminates for T and continues to the next terminology until all terminologies have been processed. We note that determining the largest value of k for which k:1 trapezoids exist is of practical and "academic" interest. For a k:1 trapezoid, the k – 1 concepts on the path from the concept B to the concept A are then proposed by the algorithm as possible imports into the NCIt to the human expert. The algorithms were published previously [7,8]. We stress that for a k:1 trapezoid, when k >2, i.e., there are two or more intermediate concepts in Terminology 1, the expert can make an independent decision for each intermediate concept.

Experiment 1: For a sample of 2:1 trapezoids we gave our oncology expert (KP) the following choices:

1) The intermediate concept in Terminology 1 (e.g., Concept X in Figure 1) should be imported into NCIt.
2) The intermediate concept in Terminology 1 should not be imported into NCIt because it is not relevant to cancer.
3) The concept structure is incorrect to begin with (e.g., Concept X should not be a child of Concept A).
4) Other (please fill in).

The distribution of concepts between these four choices will be shown in the Results section.

M:N Trapezoids and M-Rectangles

M:N trapezoids are a generalization of *k:1* trapezoids. Whenever *M = N,* it is geometrically more appropriate to refer to an *M*-rectangle. We formally define 2-rectangles and *M*:*N* trapezoids as follows:

Definition 1: The concepts A, B, and X (from Terminology 1) and A, B, and Y (of NCIt) form a 2-rectangle if and only if:

- The concepts X and Y have identical parents in Terminology 1 and in the NCIt (in this case A).
- The concepts X and Y have identical children in Terminology 1 and in the NCIt (in this case B).
- The concept X does not appear anywhere in the NCIt.
- The concept Y does not appear anywhere in Terminology 1.
- There is no synonymy relationship and no hierarchical relationship between X and Y known (in the UMLS).

Definition 2: The concepts A, B, and X_i (from Terminology 1) and A, B, and Y_j (of NCIt) form an *M*:*N* trapezoid if and only if:

- The concepts X_1 and Y_1 have identical parents in Terminology 1 and in the NCIt (in this case A).
- The concepts X_{M-1} and Y_{N-1} have identical children in Terminology 1 and in the NCIt (in this case B).
- The concepts X_i do not appear anywhere in the NCIt (i = 1..M-1)
- The concepts Y_j do not appear anywhere in Terminology 1 (j=1..N-1).
- There is no synonymy relationship and no hierarchical relationship between the X_i and Y_j known (in the UMLS)
- Every concept X_{i+1} is a child of the concept X_i. Every concept Y_{j+1} is a child of the concept Y_j.

Figure 2 shows an abstract layout of a 2-rectangle to elucidate the above Definition 1.

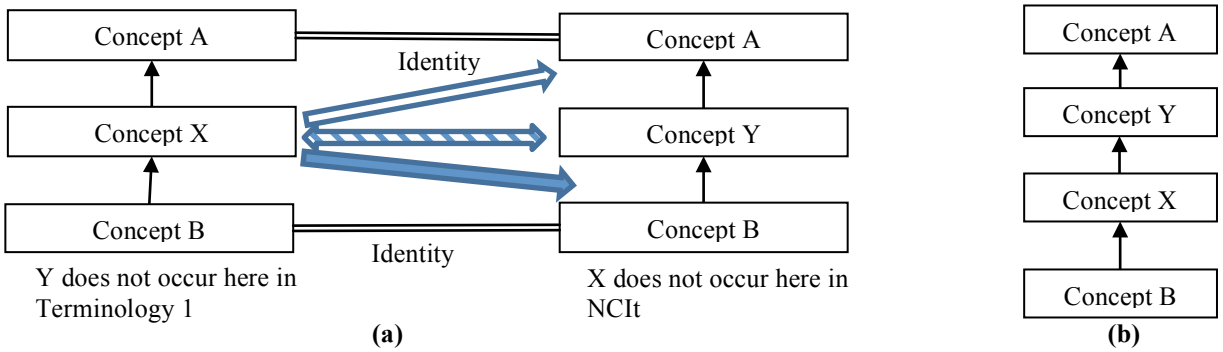

Figure 2. (a) An abstract layout of a 2-rectangle topological pattern. The double-headed arrow indicates the case that X is a synonym of Y. The no-fill, one-headed arrow indicates that X could be imported into the NCIt as child of A and as a parent of Y. The solid-fill one-headed arrow indicates that X could be imported into the NCIt as child of Y and as parent of B. This scenario is shown in (b). Only one of these three cases is possible. In addition it is possible that there is an error, e.g., Y is not really a child of A.

In a 2-rectangle there are six possible cases for how X and Y may relate to each other.

1) It holds that NCIt could contain the path of concepts connected by links B → Y → X → A. This import case is hinted at by the no-fill, one-headed arrow in Figure 2(a).
2) It holds that NCIt could contain the path of concepts B → X → Y → A. This import case is hinted at by the solid-fill, one-headed arrow in Figure 2(a). The resulting path of this import is shown in Figure 2(b).
3) Concept X is a real world synonym of concept Y, which was previously not recognized by the UMLS editors. In other words, it holds that X=Y. More precisely, it holds that CUI(X) ≡ CUI(Y), using the symbol for identity.
4) There might be a structural error in Terminology 1, e.g., X is not really a child of A.
5) There might be a structural error in NCIt.
6) The concepts X and Y are *alternative classifications*. This is the most difficult case from a conceptual standpoint. It indicates two different ways of how to conceptualize a domain that are both valid but not immediately compatible. This is best understood with an artificial example. Assume that A = *extremities* and B = *left leg*. Then X may validly be *upper extremities* while Y may be *left extremities*. These are two different, valid ways how to organize the immediate subclasses of A, however, a deeper analysis is required to make the two approaches compatible. This case is of limited interest in the current paper.

We are primarily interested in cases 1) and 2), while case 3) may also provide an enrichment of NCIt. We note, parenthetically, that Y might also be imported into Terminology 1 (Figure 2).

Experiment 2a: A sample of 2-rectangles was reviewed by YC. YC was given the six choices above.

Experiment 2b: SDC reviewed the results of Experiment 2a and commented on them.

We will now advance from 2-rectangles to 3:2 trapezoids. Again, Concept A and Concept B are identical in both Terminology 1 and NCIt. There are two intermediate concepts (Z and Y) on a path from B to A in Terminology 1, while there is one intermediate concept (X) between Concept B and Concept A in NCIt.

Experiment 3: A sample of 3:2 trapezoids was reviewed by YC. YC was given eight choices.

1) X and Y are alternative classifications of A. This is analogous to alternative classifications for 2-rectangles.
2) Y and Z are both imported into NCIt and it holds that B → Z → Y → X → A.
3) Y and Z are both imported into NCIt and it holds that B → Z → X → Y → A.
4) Y and Z are both imported into NCIt and it holds that B → X → Z → Y → A.
5) Y = X (Z could also be a child of X)
6) Z = X (Y could also be a parent of X)
7) There might be a structural error in Terminology 1, e.g., Z is a synonym of Y.
8) There might be a structural error in NCIt, e.g., X is not a child of A at all.

We limited the burden for the reviewer by excluding the consideration whether the concepts Y and Z are desirable for import into a cancer terminology, otherwise additional cases would arise, such as:

9) Only Y is a desirable import, and it holds that B → Y → X → A.
10) Only Y is a desirable import, and it holds that B → X → Y → A.
11) Only Z is a desirable import, and it holds that B → Z → X → A.
12) Only Z is a desirable import, and it holds that B → X → Z → A.
13) Neither Y nor Z is desirable to be imported into the NCIt cancer terminology.

For a 2:3 trapezoid, analogous cases can be defined, which was done in this research, even though the primary interest is in importing into NCIt. A 2:3 trapezoid would indicate an opportunity of exporting concepts from NCIt.

Experiment 4: A sample of 2:3 trapezoids was reviewed by YC. YC was given eight choices as in Experiment 3.

The UMLS contains cycles of hierarchical relationships [15]. Furthermore, multiple parents may lead to overlapping trapezoids which could lead to counting the same intermediate concept multiple times. We eliminated cycles (by detecting the repetition of a CUI along a PAR path) as well as duplicate intermediate concepts in the results.

Results

Table 2 shows the numbers of potential import concepts identified in *k*:1 trapezoids. When calculating the numbers in Column 3, we made sure that these are unique concepts (and therefore there might be fewer concepts than trapezoids). Column 4 lists the total numbers of *k*:1 trapezoids found by the algorithm. Among the eight candidate terminologies, MEDCIN could contribute the largest number of concepts to the NCIt, followed by FMA and GO.

Table 2. Potential concepts in *k*:1 trapezoids.

Candidate terminology	Size of candidate terminology (# of CUIs)	Potential concepts in candidate terminology	Number of *k*:1 trapezoids
MEDCIN	318,647	288	413
FMA	82,043	156	228
GO	57,226	154	179
CPT	38,975	22	19
SNOMEDCT_VET	36,032	7	6
CPM	3,077	7	4
UMD	21,794	1	1
ATC	5,204	0	0

Table 3 shows the number of observed *k*:1 trapezoids ordered by increasing values of *k*. The table shows that *k*:1 trapezoids were found with *k* up to 9. Both the number of trapezoids and the number of potential import concepts decrease with an increasing value of *k*, which is consistent with our previous work for enriching SNOMED CT [8]. For 2:1 trapezoids, we chose a random sample of 30 trapezoids.

Table 3. Number of *k*:1 trapezoids of each kind and corresponding number of potential import concepts.

Kinds of trapezoids	Number of trapezoids	Potential concepts in candidate terminologies
2:1	520	314
3:1	153	160
4:1	75	106
5:1	63	102
6:1	26	50
7:1	9	25
8:1	2	8
9:1	2	9
10:1	0	0
11:1	0	0

Of the 30 randomly chosen 2:1 trapezoids reviewed by the domain expert (KP), the intermediate concepts in 22 (73.3%) trapezoids were recommended for potential import into NCIt, whereas the intermediate concepts in 8 (26.7%) trapezoids were not recommended for import into NCIt. No trapezoids were assigned the choices 3) and 4) in Experiment 1. Among the 22 2:1 trapezoids, KP recommended that the intermediate concepts in three of them should be imported into NCIt with a variation of their original term. For example, "*disorders of peripheral nerve, neuromuscular junction and muscle*" (C2102996) was recommended to be imported between "*nervous system disorder*" (C0027765) and "*peripheral neuropathy*" (C0031117), but with a simplified string, which SDC changed to "*peripheral nervous system disorder.*" Table 4 shows three example trapezoids in which intermediate concepts were recommended for potential import into NCIt.

Among the eight 2:1 trapezoids from which the intermediate concept was not recommended for import into NCIt, six are from MEDCINE, one is from CPM, and the last one is from FMA. Four out of eight intermediate concepts are assigned the semantic type "Pharmacologic Substance." One example where import was not recommended consists of A = "*Muscle relaxants*," B = "*Baclofen*" and X = "*Skeletal muscle relaxants*." Investigating the definition of *Muscle relaxants* as it is given in the NCIt, it turns out that it is defined as "Any agent that relaxes skeletal muscles and reduces muscle contraction," thus its semantics is equal to the semantics of X itself. Thus, importing X would be redundant from the point of view of the NCIt.

Table 4. Example *k*:1 trapezoids in which the intermediate concept in the candidate terminology was recommended.

Candidate terminology	NCIt
2:1	
MEDCIN	*NCIt*
Lymphoid Tissue (C0024296)	Lymphoid Tissue (C0024296)
Epithelium-associated lymphoid tissue (C1179414)	
mucosa-associated lymphoid tissue (C0599921)	mucosa-associated lymphoid tissue (C0599921)
2:1	
GO	*NCIt*
Cell Cycle Checkpoints (C1155874)	Cell Cycle Checkpoints (C1155874)
mitotic cell cycle checkpoint (C2263179)	
Mitotic Spindle Checkpoints (C1155750)	Mitotic Spindle Checkpoints (C1155750)
2:1	
FMA	*NCIt*
Epithelial Cells (C0014597)	Epithelial Cells (C0014597)
Endo-epithelial cell (C1181294)	
Thymic epithelial cell (C0229951)	Thymic epithelial cell (C0229951)

Table 5 shows the numbers of 2-rectangles, 2:3 trapezoids and 3:2 trapezoids found by our algorithm. In order to analyze the relationships of intermediate concepts in the trapezoids, random samples of two times 30 trapezoids and 30 rectangles were chosen from MEDCIN. That yielded a total of 90 topological patterns for analysis. For all the other candidate terminologies, the domain expert (YC) reviewed all of the trapezoids that the algorithm identified. In total, 100 2-rectangles, 69 2:3 trapezoids, and 71 3:2 trapezoids were reviewed. Subsequently, SDC of the National

Cancer Institute further reviewed the marked up sample of 100 2-rectangles. We measured the inter-rater agreement between YC and SDC using Cohen's Kappa. The observed Kappa is 0.6175, indicating substantial agreement [16].

Table 6 shows the human review results of the sample of 2-rectangles. All cases except "Error in NCIt" were observed. The results show that 55% of the rectangles in the sample are alternative classifications. Another 20% + 17% = 37% fall into two categories where the intermediate concept in the candidate terminologies could be imported into the NCIt as a parent or child of its intermediate concept. For comparison, in our previous study on SNOMED CT [7], 23.6% of the 2-rectangles in the sample fell into these two categories. Thus, the NCIt is a better target terminology for concept import than SNOMED CT, which may be due to the fact that SNOMED CT is more comprehensive than the NCIt. The percentage of cases of synonymy is 6.0% in this study, which is lower than the 14.5% in our previous study on SNOMED CT [7].

Table 5. Number of 2-rectangles, 2:3, and 3:2 trapezoids and sample sizes for review.

Candidate terminologies	Number of 2:3 trapezoids	Sample size of 2:3	Number of 3:2 trapezoids	Sample size of 3:2	Number of 2-rectangles	Sample size of 2-rectanges
MEDCIN	119	30	90	30	183	30
GO	3	3	25	25	12	12
ATC	1	1	0	0	2	2
CPM	11	11	3	3	9	9
CPT	12	12	1	1	7	7
SNOMEDCT_VET	1	1	1	1	4	4
FMA	11	11	11	11	35	35
UMD	0	0	0	0	1	1
Total	--	69	--	71	--	100

Table 6. Human review results of 2-rectangles.

Candidate terminology	Sample Size	Alternative classification	Y IS_A X	X IS_A Y	Error in Term. 1	Error in NCIt	Synonym
MEDCIN	30	17	6	3	2	0	2
GO	12	9	0	1	0	0	2
ATC	2	2	0	0	0	0	0
CPM	9	9	0	0	0	0	0
CPT	7	6	0	1	0	0	0
SNOMEDCT_ VET	4	3	0	1	0	0	0
FMA	35	9	14	11	0	0	1
UMD	1	0	0	0	0	0	1
Total (%)	100 (100%)	55 (55%)	20 (20%)	17 (17%)	2 (2%)	0	6 (6%)

Figure 3 shows an example 2-rectangle in which the intermediate concept in MEDCIN could be a child of the intermediate concept in NCIt. Neutropenia, a condition in which there is a lower-than-normal number of neutrophils (neutrophilic white blood cell), is a kind of non-neoplastic hematologic and lymphocytic disorder.

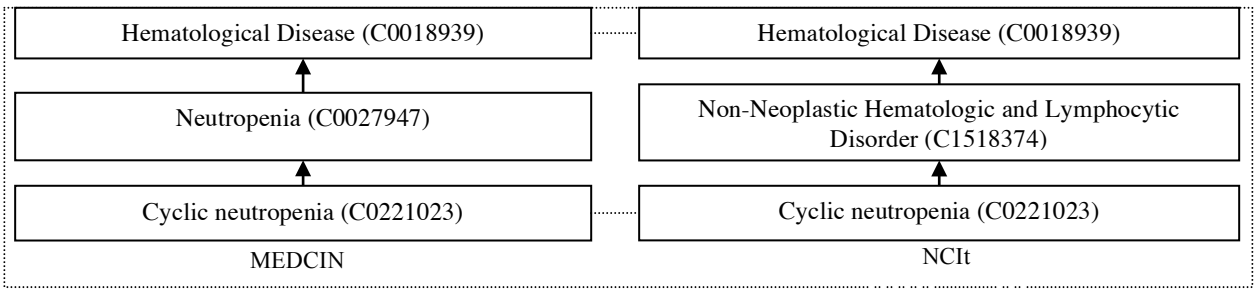

Figure 3. An example 2-rectangle in which the intermediate concept in MEDCIN could be a child of the intermediate concept in NCIt.

Figure 4 shows another example 2-rectangle where the intermediate concept in FMA "*White matter of telencephalon*" (C2327688) was deemed to be an alternative classification of the intermediate concept in the NCIt "*Brain White Matter*" (C1706995). In another 2-rectangle, two intermediate concepts "*Implantable prosthesis*" (C1961790) and "*Implants*" (C0021102) were deemed to be synonymous by the human evaluator.

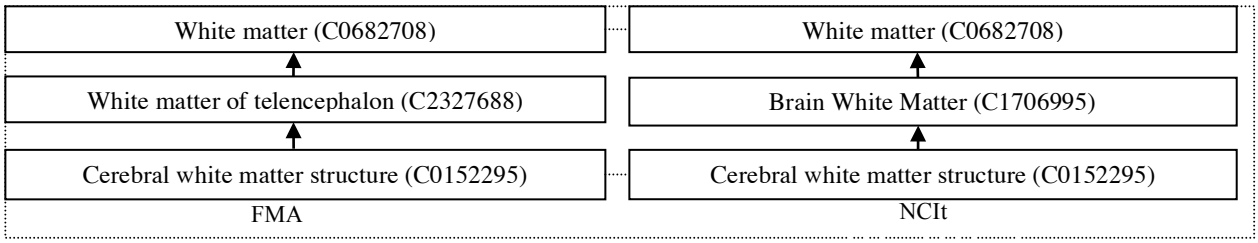

Figure 4. An example 2-rectangle in which the intermediate concepts were deemed to be alternative classifications.

Table 7 shows four examples of 2:3 and 3:2 trapezoids. In the first example, the intermediate concept in MEDCIN "*Ovarian Carcinoma*" (C0029925) was deemed to be a synonym of "*Epithelial ovarian cancer* (C0677886)." In the second example, two intermediate concepts "*oncologic disorders*" (C3853943) and "*cancer-related problem/condition*" (C0280950) were deemed to be alternative classifications. The reason that these are not synonyms is that the hierarchy of oncologic disorders in MEDCIN is used to represent the cancer diagnoses, which is not the same as the meaning of cancer-related conditions (the NCIt term). In the third example, the second intermediate concept "*inherited genetic conditions*" (C3648695) was deemed to be synonymous with "*Hereditary Diseases*" (C0019247). The fourth example is another case that is best understood as an alternative classification. *Subclavian artery* and *Axillary Artery* are *Systemic arteries*, but NCIt does not classify arteries as systemic.

Table 7. Four examples of 2:3 and 3:2 trapezoids.

	Candidate terminology	**NCIt**
2:3	*MEDCIN*	*NCIt*
	Malignant neoplasm of ovary (C1140680)	Malignant neoplasm of ovary (C1140680)
	X: Ovarian Carcinoma (C0029925)	**Y:** Malignant Ovarian Surface Epithelial-Stromal Tumor (C1518236)
		Z: Epithelial ovarian cancer (C0677886)
	small cell carcinoma of ovary (C0029925)	small cell carcinoma of ovary (C0029925)
2:3	*MEDCIN*	*NCIt*
	Disease (C0012634)	Disease (C0012634)
	X: oncologic disorders (C3853943)	**Y:** cancer-related problem/condition (C0280950)
		Z: Precancerous Conditions (C0032927)
	Carcinoma in Situ (C0007099)	Carcinoma in Situ (C0007099)
3:2	*MEDCIN*	*NCIt*
	Disease (C0012634)	Disease (C0012634)
	X: Pediatric Disorder (C0679381)	**Z:** Hereditary Diseases (C0019247)
	Y: inherited genetic conditions (C3648695)	
	Congenital chromosomal disease (C0008626)	Congenital chromosomal disease (C0008626)
3:2	*FMA*	*NCIt*
	Arteries (C0003842)	Arteries (C0003842)
	X: Systemic artery (C0933549)	**Z**: Branch of axillary artery (C0815974)
	Y: Branch of subclavian artery (C1283356)	
	Structure of superior thoracic artery (C0226419)	Structure of superior thoracic artery (C0226419)

Table 8 and Table 9 show the review results of the 2:3 and 3:2 trapezoids. No errors in the sample were observed. The results show that "alternative classification" is the most prevalent case, followed by "possible import as parent or child," and then "synonyms." These findings are also consistent with our previous study on SNOMED CT [8].

Table 8. Human review results of 2:3 trapezoids.

Candidate terminology	Sample size	Alter. classi-fication	Z → Y → X	Z → X → Y	X → Z → Y	X is a synonym of Y	X is a synonym of Z	Error in Term. 1 or NCIt
MEDCIN	30	18	0	0	5	1	6	0
GO	3	2	0	0	0	1	0	0
ATC	1	0	1	0	0	0	0	0
CPM	11	0	11	0	0	0	0	0
CPT	12	0	0	0	11	0	1	0
SNOMED CT VET	1	1	0	0	0	0	0	0
FMA	11	2	3	0	2	3	1	0
UMD	0	0	0	0	0	0	0	0
Total	69	23	15	0	18	5	8	0
Percentage	100%	33.3%	21.7%	0%	26.1%	7.2%	11.6%	0%

Table 9. Human review results of 3:2 trapezoids.

Candidate terminology	Sample size	Alter. classi-fication	Y → X → Z	Y → Z → X	Z → Y → X	Z is a synonym of X	Z is a synonym of Y	Error in Term. 1 or NCIt
MEDCIN	30	17	2	1	1	3	6	0
GO	25	16	3	0	2	4	0	0
ATC	0	0	0	0	0	0	0	0
CPM	3	1	0	0	0	0	2	0
CPT	1	1	0	0	0	0	0	0
SNOMED CT VET	1	0	0	0	0	0	1	0
FMA	11	3	0	7	1	0	0	0
UMD	0	0	0	0	0	0	0	0
Total	71	38	5	8	4	7	9	0
Percentage	100%	53.5%	7.0%	11.3%	5.6%	9.9%	12.7%	0%

Discussion and Conclusions

In this work, we applied a topological-pattern-based method to recommend concepts from eight UMLS source terminologies that could potentially enrich the NCIt's conceptual content. A variety of topological patterns between pairs of UMLS source terminologies were identified by our algorithm with potential import concepts for expert review. Three domain experts reviewed the samples and determined whether the potential concepts in the samples should be imported and suggested how they should be inserted into existing structures of the NCIt. The results demonstrated the effectiveness of the topological-pattern-based method for enriching the NCIt. For 2-rectangles, 2:3 trapezoids, and 3:2 trapezoids, the prevalence of possible relationships between intermediate concepts in pairs of terminologies was consistent with our previous studies on SNOMED CT [8]. The most prevalent cases were alternative classification, followed by various forms of import.

A few limitations need to be noted. In this study, only "PAR" links with "INVERSE_IS_A" annotations were used. We did not look for rectangles larger than 2-rectangles, and we did not attempt to identify 4:2, 4:3, etc. trapezoids. As shown above, even a 3:2 trapezoid leads to a remarkable number of possibilities as to how intermediate concepts could relate to each other. For *M:N* trapezoids with larger values of *M* and *N* there will be a combinatorial explosion of possible cases; that will make it difficult for a human expert to consider all of them even for a small number of trapezoids. Human review is important, however. In [17], we have conducted a preliminary analysis of the difficulty of importing pattern-based concepts into the NCIt. The contexts and definitions of potential new concepts originating from a source need to be evaluated to validate the intended meaning. In a few cases, SDC "overrode" the decisions of YC. We note that the viewpoints of an outside auditor and a curator of a terminology are necessarily different. A curator has a deeper knowledge and better understanding of how individual representational decisions

were reached. We also note that not all the potential concepts identified would be imported, because of the different uses and alternative schemas of different terminologies. However, our method provides a way to look at possibly needed areas for improving domain coverage. In future, it could play a role in regular QA of terminologies at NCI.

In future work, we will investigate the use of the NCI Metathesaurus instead of the UMLS Metathesaurus. It is a subset of UMLS with other sources added, and is well curated in areas of interest to NCI [18]. Moreover, NCIt is more frequently updated in the NCI Metathesaurus than in the UMLS. Another useful source for future work is the NCIt GO mapping of the NCI [19]. We plan to extensively use feedback of the NCI EVS. After getting feedback and arguments for inclusion and exclusion of our suggested concepts, we will refine our methods to produce more accurate recommended concepts accordingly. We also plan to investigate horizontal density differences with varying numbers of sibling concepts between pairs of terminologies to identify more missing concepts for import. Lastly we will work on 3:1, 4:1, 5:1 and 6:1, and 4:2 samples.

Acknowledgments

Research reported in this publication 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 necessarily represent the official views of the National Institutes of Health.

References

1. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008:67-79.
2. Cimino JJ. High-quality, standard, controlled healthcare terminologies come of age. Methods Inf Med. 2011;50(2):101-4.
3. Rector A, Qamar R, Marley T. Binding ontologies and coding systems to electronic health records and messages. Applied Ontology. 2009;4(1):51-69.
4. Finnegan R. ICD-9-CM coding for physician billing. J Am Med Rec Assoc. 1989;60(2):22-3.
5. de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, et al. The NCI Thesaurus quality assurance life cycle. J Biomed Inform. 2009;42(3):530-9.
6. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4-5):394-403.
7. He Z, Geller J, Elhanan G. Categorizing the Relationships between Structurally Congruent Concepts from Pairs of Terminologies for Semantic Harmonization. AMIA Jt Summits Transl Sci Proc. 2014;2014:48-53.
8. He Z, Geller J, Chen Y. A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization. Artif Intell Med. 2015;64(1):29-40.
9. Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. J Am Med Inform Assoc. 2006;13(6):676-90.
10. Cohen B, Oren M, Min H, Perl Y, Halper M. Automated comparative auditing of NCIT genomic roles using NCBI. J Biomed Inform. 2008;41(6):904-13.
11. Mougin F, Bodenreider O. Auditing the NCI thesaurus with semantic web technologies. AMIA Annu Symp Proc. 2008:500-4.
12. Jiang G, Solbrig HR, Chute CG. Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. J Am Med Inform Assoc. 2012;19(e1):e129-36.
13. Huang KC, Geller J, Halper M, Perl Y, Xu J. Using WordNet synonym substitution to enhance UMLS source integration. Artif Intell Med. 2009;46(2):97-109.
14. Marquet G, Mosser J, Burgun A. A method exploiting syntactic patterns and the UMLS semantics for aligning biomedical ontologies: the case of OBO disease ontologies. Int J Med Inform. 2007;76 Suppl 3:S353-61.
15. Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp. 2001:57-61.
16. Viera A.J., Garrett JM. Understanding Interobserver Agreement: The Kappa Statistic. Fam Med. 2005;37(5):360-3.
17. He Z, Geller J. Preliminary Analysis of Difficulty of Importing Pattern-Based Concepts into the National Cancer Institute Thesaurus. Studies in Health Technology and Informatics. 2016; In press.
18. The National Cancer Institute NCI Metathesaurus [2/25/2016]. Available from: https://ncimeta.nci.nih.gov/ncimbrowser/.
19. GO to NCIt Mapping. Version April 2014 [2/25/2016]. Available from: https://ncit.nci.nih.gov/ncitbrowser/pages/vocabulary.jsf?dictionary=GO_to_NCIt_Mapping&version=1.1