MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-773

Identifying Problematic Concepts in SNOMED CT using a Lexical Approach

Ankur Agrawal[a,b], Yehoshua Perl[a], Gai Elhanan[c]

[a] *Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA*
[b] *Department of Computer Science, Manhattan College, Bronx, NY, USA*
[c] *Halfpenny Technologies Inc., Bluebell, PA, USA*

Abstract

SNOMED CT (SCT) has been endorsed as a premier clinical terminology by many organizations with a perceived use within electronic health records and clinical information systems. However, there are indications that, at the moment, SCT is not optimally structured for its intended use by healthcare practitioners. A study is conducted to investigate the extent of inconsistencies among the concepts in SCT. A group auditing technique to improve the quality of SCT is introduced that can help identify problematic concepts with a high probability. Positional similarity sets are defined, which are groups of concepts that are lexically similar and the position of the differing word in the fully specified name of the concepts of a set that correspond to each other. A manual auditing of a sample of such sets found 38% of the sets exhibiting one or more inconsistent concepts. Group auditing techniques such as this can thus be very helpful to assure the quality of SCT, which will help expedite its adoption as a reference terminology for clinical purposes.

Keywords:

SNOMED CT, Electronic Health Record, Meaningful Use, Lexical Analysis, Auditing, Quality Assurance.

Introduction

SNOMED CT (SCT) [1] is a controlled clinical reference terminology with comprehensive coverage of clinical findings, diseases, procedures, therapies and outcomes intended for recording clinical data [2, 3]. This data can be made available to computer systems for clinical decision support [4] and improved patient safety [5-7]. In the past decade, SCT gained recognition as a premier clinical terminology through endorsements from many national and international organizations. The most commonly perceived use of terminologies such as SCT is the encoding of clinical data within electronic medical systems including electronic health records (EHRs) and clinical information systems.

The usage of large standard terminologies like SCT is highly influenced by quality assurance and auditing issues. Past researches (e.g., [8, 9]) have identified instances of inconsistent modeling in SCT, which could act as a barrier for the successful use of SCT in EHRs [10]. Such inconsistencies may be perceived to have minimal implications regarding clinical coding. However, inconsistencies may significantly affect the performance of reasoners and inference generation (e.g., in the context of error detection and decision support) as these explicitly rely on the completeness and consistency of formal definitions. A number of techniques have been proposed to identify the weak spots that are likely to contain errors and present them to an auditor for manual review [8, 11, 12].

An intensive auditing effort is urgently needed to improve the quality of the suggested SCT concepts and ensure quality assurance in SCT [13]. However, an audit of all concepts of SCT requires extensive quality assurance resources and will require an extended period of time. A desired approach in coping with this urgent quality assurance need is to develop techniques for identifying subsets of SCT with expected higher concentration of errors.

This paper presents one such approach, which analyzes the textual representation of sets of concepts similar at the term-level, in an attempt to characterize the consistency of the modeling across these concepts. Sets of concepts with similar terms are gathered through standard lexical techniques for all the 19 hierarchies of SCT. Positional similarity sets are introduced and an analysis is performed on SCT's *Procedure* hierarchy to look for inconsistent modeling among these lexically similar concepts. The results show that 38% of the positional similarity sets that were reviewed have concepts with inconsistent modeling. Positional similarity sets are, thus, found to be effective in identifying inconsistent concepts with high likelihood.

Background

In [8], it was shown that if the concepts are lexically similar, they should also be modeled in a similar way at the description logic level. This would mean that the two lexically similar concepts should exhibit similar number of parents, groups and relationships. Table 1 displays two concepts from the *Procedure* hierarchy that are lexically similar.

Table 1 – An example of two similar concepts

CID	Fully Specified Name
179447000	*Prosthetic uncemented total shoulder replacement (procedure)*
179954004	*Prosthetic uncemented total elbow replacement (procedure)*

The two concepts *Prosthetic uncemented total shoulder replacement (procedure)* and *Prosthetic uncemented total elbow replacement (procedure)* are lexically similar as they only differ in the joint involved – shoulder vs. elbow. This would mean that the two concepts should also be modeled in a similar way. Figure 1 and Figure 2 displays the snapshots of the modeling of the two concepts from CliniClue Browser [14].

Each of these concepts have one parent as shown by the *is-a* relationship, three attribute relationships and one role group. Also, both of these concepts have the same attribute types, viz. *method*, *direct device* and *procedure site - direct* and they only differ in their target values. So, it can be seen that the two concepts are not only similar lexically but they are also similar structurally and this is what is expected out of similar concepts.

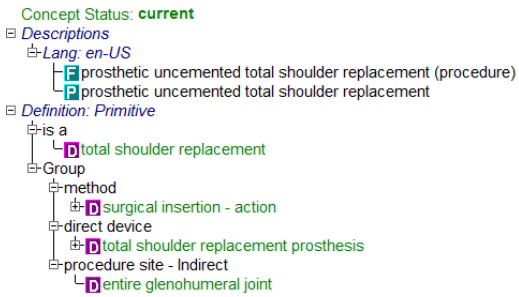

Figure 1- CliniClue snapshot of CID 179447000

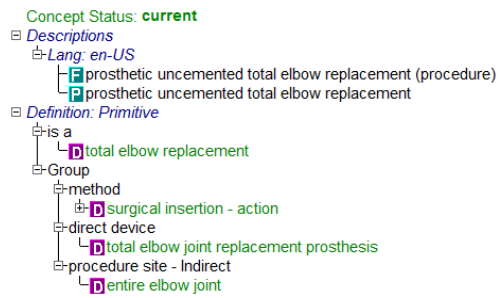

Figure 2- CliniClue snapshot of CID 179954004

A technique to group similar concepts was formulated in [8] and the groups of similar concepts were called similarity sets. All concepts whose fully specified names differed from each other by one word were grouped together as a similarity set. The position of the differing word among the concepts of a set was not considered relevant, while generating such similarity sets. Standard lexical variations as well as stop-words (like "a," "an," "the") were ignored. Table 2 displays one such example of a similarity set. The set consists of three concepts. The second concept differs from the seed (first) concept in the procedure involved – prophylactic vs. therapeutic. The third concept differs from the seed concept in the portion of the limb involved – upper vs. lower.

Table 2 –A similarity set with three concepts

CID	Fully Specified Name
305067001	*Prophylactic upper limb stretching (procedure)*
305096000	*Therapeutic upper limb stretching (procedure)*
305070002	*Prophylactic lower limb stretching (procedure)*

A random sample of 60 such sets was audited and the results were analyzed. Table 3 summarizes the results of that study. As can be seen in the table, 30% of the sets were found to be inconsistent. In terms of concepts, 13.2% of the concepts were found to be inconsistent.

The next logical step in this study would be to enhance these similarity sets by applying different techniques that can help improve the efficiency of identifying inconsistencies in these sets. This paper presents a structural indicator, which can help increase the likelihood of finding inconsistencies among concepts in a similarity set.

Table 3– Results from past study[8]

	#	%
Sets	60	
Inconsistent sets	18	30.0
Concepts	204	
Inconsistent concepts	27	13.2
Primitive concepts	159	77.9
Leaf concepts	167	81.8

In the previous study, similarity sets were generated without taking into consideration the position of the word that was different between the seed concept and the other concepts in the set. This can result in concepts with different meaning being grouped together in a set. This study makes the rule stricter by only considering those concepts to be a part of a set where the position of the differing word in the fully specified name (FSN) is the same in all the concepts of the set. Such a set is called a positional similarity set. Applying this strictness in the position of words to the similarity set shown in Table 2 will result in two positional similarity sets with two concepts each as shown in Table 4. The first set has two concepts that differ in the procedure involved – prophylactic vs. therapeutic. The second set has two concepts that differ in the portion of the limb involved – upper vs. lower. The strictness in the position of the differing word generates similarity sets with concepts that are much more similar in their lexical meaning. This technique also helps reduce the size of the sets which helps an auditor to focus on smaller sets of concepts.

Table 4 – Two positional similarity sets

CID	Fully Specified Name
305067001	*Prophylactic upper limb stretching (procedure)*
305096000	*Therapeutic upper limb stretching (procedure)*
305067001	*Prophylactic upper limb stretching (procedure)*
305070002	*Prophylactic lower limb stretching (procedure)*

An example of inconsistencies that can be identified in concept modeling using similarity sets was explained in detail in [8]. It involved a set containing five concepts, three of which are *Primary cemented total ankle replacement*, *Primary cemented total hip replacement*, and *Primary cemented total knee replacement*. Although, the concepts looked alike lexically, discrepancies were identified in their hierarchy, role-groups, attributes, attribute values and definitional status.

The test-bed for this study is the *Procedure* hierarchy of January 2011 release of SCT. *Procedure* hierarchy is one of the main hierarchies of SCT where retrieval of clinical data is most useful and relevant. There are a set of 23 potential defining attributes for the *Procedure* hierarchy [15]. Table 5 lists these defining attributes. For most attribute domains, SCT defines one or more ranges from which target values can be assigned. For example, the attribute *Direct substance* can be assigned target values from two ranges: *Substance* and *Pharmaceutical/biologic product*. Furthermore, SCT allows multiple attributes and their values to be grouped together to create

what are called "role-groups". These role-groups combine multiple attribute/value pairs to create specific associations between appropriately relevant target concepts, thus enhancing the precision of definitions.

Table 5– Defining attributes for Procedure hierarchy

Access	*Procedure site*
Direct device	*Procedure site - Direct*
Direct morphology	*Procedure site - Indirect*
Direct substance	*Recipient category*
Has focus	*Revision status*
Has intent	*Route of administration*
Indirect device	*Surgical Approach*
Indirect morphology	*Using device*
Method	*Using access device*
Priority	*Using energy*
Procedure device	*Using substance*
Procedure morphology	

Materials and Methods

An algorithm is designed to create positional similarity sets. A positional similarity set is a group of concepts where the FSNs of the concepts have a lexical similarity, i.e., the concept names are similar in their word structure. The focus, here, is on FSNs that differ from each other by one word. The sets are positional in the sense that the comparison between the two concepts is one-to-one, i.e., the matching words and the differing word should correspond in their positions in the two concepts in a set. For example, let f1 and f2 be two five-word FSNs with f1 = "w1 w2 w3 w4 w5" and f2 = "w1 w6 w3 w4 w5", where each wi is an individual word in the concept FSN. Then the concepts f1 and f2 are in a set together as there FSNs differ only by one word, w2 versus w6, and at corresponding positions, i.e., second word in each of the concept.

The sets are created by randomly choosing a concept from the hierarchy as the seed concept of a set. The seed concept is then the first concept of the set. All the concepts in the hierarchy, where FSN differ from that of the seed concept by one word and at the corresponding position, are placed in the same set as this seed concept. The process of randomly selecting a seed concept for a set and finding other concepts that form the set continues until all the concepts in the hierarchy, which can be classified as a part of a set, have been classified. There must be a minimum of two concepts in order to form a set.

The algorithm takes care of the stop-words in the concept FSNs in order to improve the efficiency of the result. Stop words are words that are filtered out prior to the processing of the concept FSNs. Standard lexical variations as well as stop-words (like "a," "an," "the") are ignored in order to improve performance. Only the single word prepositions, coordinating conjunctions, and definite and indefinite articles were used as stop-words.

Positional similarity sets are generated for all the 19 hierarchies of SCT using the January 2011 release. A randomly selected sample of 50 sets from Procedure hierarchy is then evaluated by one of the authors (GE) who is an MD with extensive terminology training and experience.

Results

Table 6 displays some data regarding the positional similarity sets generated for each of the 19 hierarchies of SCT. The first column displays the hierarchy name, the second column displays the number of positional similarity sets generated, the third column displays the number of concepts in all these sets and the fourth column displays the average number of concepts per set. For instance, 7,641 positional similarity sets were generated for the *Procedure* hierarchy. These sets had a total of 20,384 concepts with an average of 2.7 concepts per set. The positional similarity sets are not disjoint and concepts may be repeated between the sets of a hierarchy. This phenomenon can also be seen in the positional similarity sets in Table 4 where the concept *Prophylactic upper limb stretching (procedure) appears in both the sets.*

Table 6– Set data for all 19 hierarchies

Hierarchy	#Sets	#Cpts	Avg #cpts/set
Body Structure	16629	46349	2.8
Clinical Finding	14882	42516	2.8
Environment	75	262	3.5
Event	802	1961	2.4
Linkage	7	15	2.1
Observable	1793	5018	2.8
Organism	674	2846	4.2
Pharmaceutical	2722	7576	2.8
Physical Force	13	37	2.8
Physical Object	824	2331	2.8
Procedure	7641	20384	2.7
Qualifier	710	2647	3.7
Record	20	58	2.9
Situation	395	1145	2.9
Social	151	444	2.9
Special	107	553	5.2
Specimen	171	460	2.7
Staging	135	358	2.6
Substance	1274	4938	3.9

The 50 sample sets from the *Procedure* hierarchy that were evaluated for inconsistencies consisted of 116 concepts with an average of 2.3 concepts per set. Table 7 displays one such sample set having two concepts, one being the aortography of abdomen and the other being the aortography of thorax.

Table 7 – One of the 50 sample sets

CID	Fully Specified Name
43145003	*Abdominal aortography, positive contrast (procedure)*
48735005	*Thoracic aortography, positive contrast (procedure)*

Snapshots of the modeling of the two concepts from Table 7 are shown in Figure 3 and Figure 4. These snapshots have been captured from the CliniClue Xplore browser.

The two procedures, by their nature, are identical except for the part of the aorta that is meant to be imaged; the thoracic vs. the abdominal aorta. Thus, it is expected that their modeling will be similar except where attribute values matter. However, on close examination, it is clear that the hierarchical tree of the two concepts is quite different. *Thoracic aortography, positive contrast* has two parents, while it's seemingly sibling concept *Abdominal aortography, positive contrast* has six parents, none shared. The two concepts only share one grandparent *Aortography*. As a result, only the thoracic procedure conveys the information that this is a procedure by injection. On the

other hand the abdominal procedure has the parent *Radiographic imaging of soft tissue*. This parent may seem redundant since both concepts, through numerous levels of ancestors, are linked to *Procedure on soft tissue*. Similarly, *Arteriography using contrast* is redundant and has been eliminated in future releases of SCT, whereas the parent *Contrast radiography of abdominal cavity* lacks a parallel concept for the thoracic cavity. As for the attributes, although the abdominal concept has only one group and the thoracic one has three, overall they display the same set of attributes. However, the thoracic concept carries an additional attribute target for the *procedure site - direct* attribute - *Intrathoracic vascular structure*, which is not directly related to the thoracic and abdominal aortic structure attribute targets in the respective concepts.

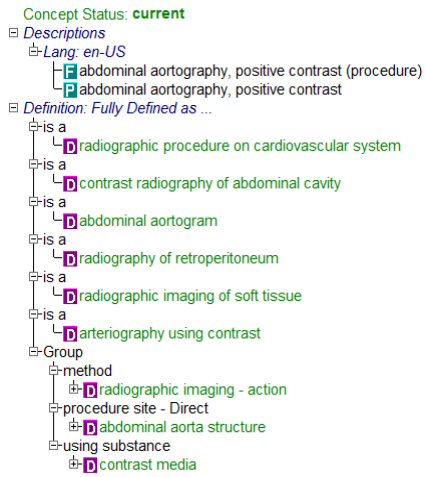

Figure 3- CliniClue snapshot of CID 43145003

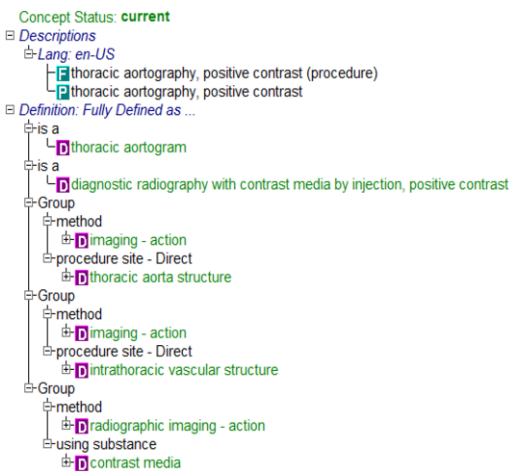

Figure 4- CliniClue snapshot of CID 48735005

Table 8 – Sample set data

	#	%
Sets	50	
Inconsistent sets	19	38.0
Concepts	116	
Inconsistent concepts	26	22.4
Primitive concepts	94	81.0
Leaf concepts	96	82.7

Table 8 summarizes the findings of the auditing of the 50 sample sets from *Procedure* hierarchy. 38% (19 out of 50) of the sample sets were found to be inconsistent in their modeling. In terms of the percentage of concepts, 22.4% (26 out of 116) of the concepts were found to be inconsistent. 81% (94 out of 116) of the concepts in the sample set were primitive and 82.7% (96 out of 116) of the concepts were leaf nodes.

Discussion

SCT is built upon an underlying description logic model. The ability of description logic classifiers to operate is directly related to the robustness of the underlying logical formulations. Inconsistencies, as described in this work, combined with the fairly inexpressive logic underlying SCT, are bound to escape detection [16]. This creates a need to have auditing techniques that can help identify such problems.

This study emphasizes the importance of group auditing as an effective way to identify inconsistencies in the concepts of SCT as it gives the auditors an opportunity to compare the modeling of concepts that have been grouped together based on certain criteria. The study presents one such group auditing technique in the form of positional similarity sets and the results show that the method can be an effective way to identify inconsistencies in SCT and help in improving its quality.

The premise of reviewing positional similarity sets for errors is based on having a contrast between lexical similarities and structural differences. This study groups concepts together into a set if the fully specified names differ from one another by a single word and at the same position. If we allow a difference of more than one word, many concepts which are not semantically similar will enter such a similarity set. Hence the lexical similarities will be weak and the contrast with structural differences will not be strong. This will result in a lower likelihood of finding errors as compared to the current method. Our purpose is to devise a technique which increases the likelihood of finding errors for more efficient use of limited QA resources.

The study used *Procedure* hierarchy as a test bed to analyze the consistency in textual modeling as compared to the structural modeling of the concepts. There were five different kinds of inconsistencies discovered during the analysis of the concepts, namely, hierarchical discrepancies, attribute assignment, attribute values, role groups and concept definition. The description of the modeling differences between the two concepts of Figure 3 and Figure 4 as shown in the results section demonstrates the efficacy of finding inconsistencies among concepts using positional similarity sets. Since the method uses terms in concept's FSN to group similar concepts, it will work for all hierarchies in SCT to identify inconsistent concepts irrespective of the hierarchy being rich in hierarchical relationships or attribute relationships.

A comparative analysis of the data presented in Table 8 with that of the past study (Table 3) shows an improvement in the efficiency of positional similarity sets in identifying inconsistent sets (38.0% with positional similarity sets vs. 30.0% with similarity sets). Similarly, the use of the position of words as a structural indicator has also improved the efficiency in identifying the inconsistent concepts in these sets (22.4% with positional similarity sets vs. 13.2% with similarity sets). Future

work will involve deriving other structural indicators that can be used to enhance the efficiency of positional similarity sets.

The sample analyzed in this study contained 81% primitive concepts and 82.7% leaf concepts as seen in Table 8 which is comparable to the data from the past study as seen in Table 3. Thus, it can be seen that a large number of concepts are leaf nodes and under defined. This could be a reason for the sample generating large percentage of inconsistent concepts, since under-defined concepts do not guarantee that the modeling of similar concepts should be the same. However, the issues identified are general enough to assume that such inconsistencies may be ingrained throughout the *Procedure* hierarchy.

The study used concepts that had a minimum of five words in their FSN to generate the positional similarity sets. Increasing this threshold would likely increase specificity, while reducing it would likely increase sensitivity. Future work will involve identifying the effect of the number of words in finding inconsistencies using positional similarity sets. The seed selection is random as we have no reason to prefer one concept over another. Future research will consider identifying properties for selecting seeds to increase the ratio of expected error to the number of concepts reviewed.

The study was conducted on a small sample of 50 sets comprising of 116 concepts which were audited by a single auditor. The aim of the study was to formulate a method and examine its results. The results are promising with 38% of the sets exhibiting erroneous concepts. Future work will involve evaluating larger sample sets, as well as sets from other hierarchies of SCT. Future work will also involve implementing a system based on this technique, which will be presented to the curators of SNOMED CT for their review of the inconsistent concepts.

Conclusion

SCT is slated to become an integral component of standardization in health information technology in the United States and play a significant role in adopting EHRs by providing standardized encoding of health-care data. However, past studies have identified instances of inconsistent modeling in SCT, which could act as a barrier for the successful use of SCT in EHRs. This study presented a group based auditing technique in the form of positional similarity sets that can be used to identify inconsistencies in the modeling of the concepts in SCT by grouping concepts with similar terms together. 38% of the sample sets that were analyzed by an auditor were found to contain inconsistent concepts. Future studies will involve identifying other structural indicators that can be compounded together to increase the likelihood of finding inconsistencies among concepts using the positional similarity sets.

References

[1] SNOMED CT. Available at: http://www.ihtsdo.org/snomed-ct [accessed 23 October 2012]

[2] Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc. 2006;81:741-8.

[3] Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc. 2003:699-703.

[4] Ciolko E, Lu F, Joshi A. Intelligent clinical decision support systems based on SNOMED CT. Conf Proc IEEE Eng Med Biol Soc. 2010;2010:6781-4.

[5] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform. 2006;121:279-90.

[6] Elevitch FR. SNOMED CT: electronic health record enhances anesthesia patient safety. AANA J. 2005;73:361-6.

[7] Farfán Sedano FJ, Terrón Cuadrado M, García Rebolledo EM, Castellanos Clemente Y, Serrano Balazote P, Gómez Delgado A. Implementation of SNOMED CT to the medicines database of a general hospital. Stud Health Technol Inform. 2009;148:123-30.

[8] Agrawal A, Elhanan G, Halper M. Dissimilarities in the Logical Modeling of Apparently Similar Concepts in SNOMED CT. AMIA Annu Symp Proc. 2010;2010:212-6.

[9] Wei D, Halper M, Elhanan G, Chen Y, Perl Y, Geller J, et al. Auditing SNOMED relationships using a converse abstraction network. AMIA Annu Symp Proc. 2009:685-9.

[10] Giannangelo K, Fenton SH. SNOMED CT survey: an assessment of implementation in EMR/EHR applications. Perspect Health Inf Manag. 2008;5:7.

[11] Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. J Biomed Inform. 2007;40:561-81.

[12] Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. J Biomed Inform. 2009;42:413-25.

[13] Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. J Am Med Inform Assoc. 2011;18:432-40.

[14] CliniClue Xplore. Available at: http://www.cliniclue.com [accessed 15 September 2012]

[15] SNOMED CT User Guide. Available at: http://www.ihtsdo.org/fileadmin/user_upload/doc [accessed 17 September 2012]

[16] Wei D, Bodenreider O. Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies - a case study in SNOMED CT. Stud Health Technol Inform. 2010;160:1070-4.

Address for correspondence

Ankur Agrawal
Department of Computer Science, GITC Room 4400
New Jersey Institute of Technology
Newark, NJ 07102-1982 USA
E-mail: agrawal@njit.edu
Phone: (201) 214-2960