Contrasting lexical similarity and formal definitions in SNOMED CT: Consistency and implications

Ankur Agrawal a,⇑, Gai Elhanan b

a Manhattan College, Riverdale, NY, United States
b Halfpenny Technologies Inc., Blue Bell, PA, United States

Objective: To quantify the presence of and evaluate an approach for detection of inconsistencies in the formal definitions of SNOMED CT (SCT) concepts utilizing a lexical method.

Material and method: Utilizing SCT’s Procedure hierarchy, we algorithmically formulated similarity sets: groups of concepts with similar lexical structure of their fully specified name. We formulated five random samples, each with 50 similarity sets, based on the same parameter: number of parents, attributes, groups, all the former as well as a randomly selected control sample. All samples’ sets were reviewed for types of formal definition inconsistencies: hierarchical, attribute assignment, attribute target values, groups, and definitional.

Results: For the Procedure hierarchy, 2111 similarity sets were formulated, covering 18.1% of eligible concepts. The evaluation revealed that 38 (Control) to 70% (Different relationships) of similarity sets within the samples exhibited significant inconsistencies. The rate of inconsistencies for the sample with different relationships was highly significant compared to Control, as well as the number of attribute assignment and hierarchical inconsistencies within their respective samples.

Discussion and conclusion: While, at this time of the HITECH initiative, the formal definitions of SCT are only a minor consideration, in the grand scheme of sophisticated, meaningful use of captured clinical data, they are essential. However, significant portion of the concepts in the most semantically complex hierarchy of SCT, the Procedure hierarchy, are modeled inconsistently in a manner that affects their computability. Lexical methods can efficiently identify such inconsistencies and possibly allow for their algorithmic resolution.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

The U.S. is in the midst of a transformational change to a more efficient and coordinated healthcare delivery system. The ability to share data among providers, researchers, and consumers is essential to the success of this effort. Despite increased adoption of electronic health record (EHR) technology, the data collected within will remain in information silos unless methods are implemented to ensure its interoperability in respect to format and content. Therefore, integral to current health information technology initiatives is the adoption of applicable standards for various domains. One such standard is SNOMED CT (SCT), adopted for encoding problem lists. The lists, however, are only a starting point. Eventually, sophisticated algorithms and reasoning engines will utilize the conceptual representations within SCT for research and decision support, analytics, and various other tasks. Incomplete, inconsistent, or erroneous representations will negatively influence the performance of such engines and, directly or indirectly, negatively impact patient care. SCT is large and complex and imperfections are inevitable but quality assurance (QA) resources are scarce and manual auditing is time consuming, rendering manual QA of SCT impractical. Algorithmic QA must be developed to efficiently detect, and possibly resolve, inconsistencies and errors. While most terminological QA is retrospective, algorithmic QA can be incorporated into the authoring process.

It is reasonable to expect similarly worded strings, as represented by concepts’ preferred names, to be modeled similarly. In SCT, such sets of concepts may harbor higher rate of inconsistencies [1]. In this study we explore efficient methodologies to detect high-probability, high-yielding similarity sets. We set to better quantify the consistency of the formal logical representations in the Procedure hierarchy, the most semantically rich hierarchy of SCT with the most complex conceptual representations. We utilize a lexical methodology, and evaluate the potential yield of different set building approaches.
2. Background

SNOMED CT is a premier clinical reference terminology governed by the International Health Terminology Standards Development Organization (IHTSDO) [2]. It provides more than 311,000 active concepts (July 2012 release) organized into 19 distinct hierarchies and formalized using logic-based definitions [3]. Each of SCT’s concepts carries up to three levels of information, two of which are mandatory. The first is the lexical information conveyed by a concept’s unique descriptor and other possible synonyms. The second level of information is the hierarchical positioning of the concept within SCT’s directed acyclic graph, utilizing is-a relationship type to create a subsumption structure. The third level is the formal logic definition of each concept represented by a set of defining relationships to other concepts [4]. Not all of SCT’s concepts carry such semantic information, and the amount of information is variable by hierarchy and by concept. In fact, more than half of SCT’s concepts are defined as primitives, i.e. concepts that do not have sufficient defining relationships to computably distinguish them from more general concepts. Each defining relationship is composed of an assigned attribute and an attribute target value. For each hierarchy, specific attributes can be assigned from pre-defined sets. SCT also defines one or more ranges from which target values can be assigned. Furthermore, SCT allows multiple attributes and their values to be grouped together to create “role-groups.” These role-groups, mostly used in the Procedure and Clinical finding hierarchies, combine multiple attribute/value pairs to create specific associations between appropriately relevant target concepts, thus enhancing the precision of definitions.[4]

The Health Information Technology for Economic and Clinical Health Act (HITECH), enacted under the American Recovery and Reinvestment Act of 2009 [5], sets goals for adoption of EHRs and other health information technologies. However, the goals go far beyond the simple adoption of technologies and emphasize the importance of meaningful use (MU) of such technologies to ensure the desired improvements in efficiency, quality, and accessibility of U.S. healthcare [6,7]. To increase adoption and ensure MU, the Centers for Medicare and Medicaid Services (CMS) and the Office of the National Coordinator (ONC) released MU regulations and sets of standards that must be adhered to by providers and certified EHR technologies (CEHRT) [8–11], respectively. The adopted standards [10,11] are designed to ensure the essential interoperability of data captured through CEHRT in respect to their structure and content. SCT and its U.S. extension were adopted for encoding of patient problem lists within CEHRT, and concepts from SCT were chosen for the encoding of other MU objectives such as familial conditions, smoking status, and others. The SCT hierarchy that is the most pertinent to the MU effort is the Clinical finding hierarchy: the largest and second-most semantically complex among SCT’s 19 hierarchies. However, MU stage 2 final specifications [11] expand the use of SCT to the realm of procedures as one of the standards for certification as well as encoding certain aspects of laboratory results [12].

The immediate planned use of SCT concepts as part of MU is limited to “lists” with no requirement that SCT be incorporated within CEHRT. However, with the impending abundance of computable data, there is no doubt that the inherent structure and semantic network of SCT and like terminologies can vastly contribute to the development of sophisticated algorithms and reasoning agents. Such systems can enhance research and disease detection, clinical decision support, and other aspects of improved healthcare [13–16]. Cimino and others [17–21] demonstrated the synergy between clinical repositories and controlled biomedical terminologies (CBTs).

For algorithms to work reliably, the validity and consistency of the conceptual representations within CBTs is crucial. Rector et al. [22] clearly demonstrated the issue utilizing the Myocardial infarction example. In SCT (January 2010 release), myocardial infarction is not classified as a type of ischemic heart disease due to incomplete formal logic definitions. As a result, a hypothetical research query that looks to gather all Ischemic heart disease patients, relying on SCT coded data, will exclude myocardial infarction patients unless the researchers had prior knowledge of the issue or run their query using an aggregate of all instances of ischemic heart disease. The example crystallizes the implications of incomplete, incorrect, and inconsistent modeling on healthcare applications down the road. Rector and his colleagues conclude that SCT cannot be used “as-is” in their applications without significant modifications and that comprehensive quality assurance effort must be undertaken.

Rector et al. are not the only ones to report on modeling issues within SCT. Issues with the part-of relationship [23], discrepancies in defined semantics, and definitional inconsistencies between ancestors and descendants have been reported [24,25], along with critical reviews of logical and ontological issues [26]. Agrawal et al. [1], in a preliminary study, found high rates of modeling inconsistencies in the Procedure hierarchy. In the study, utilizing 60 sets (204 concepts) of lexically similar concepts, 30% of sets had at least one kind of an inconsistency: hierarchical (28%), assignment of attributes (17%), attribute target values (15%), definition level (7%), or role groups (5%). Other attribute-rich hierarchies, such as the Clinical finding hierarchy, essential for the MU initiative, may exhibit similar findings. Such inconsistencies are not necessarily errors since each concept, on its own, follows SCT’s definitional requirements. These and other findings highlight that logic-based representations do not guarantee consistent end-user views.

SCT is large and complex. In one study, SCT covered more than 88% of diagnosis and problem list terms in a large institution [27] but coverage of other domains may not be as extensive [28–32]. Therefore, SCT is expected to continually grow in coverage, but requires significant auditing efforts to improve and complete the formal logic definitions of its concepts. Rector and his colleagues [22] conclude that SCT’s description logic (DL) foundation is not only part of the problem, but also part of the solution as it can be used more effectively to detect and correct root errors.

Typical DL classifiers cannot detect that which was not explicitly stated or inferred, as Wei and Bodenreider found [33]. Methods other than those purely based on formal definitions must be employed in order to facilitate more complete QA of large CBTs. Perl and colleagues [34–36] applied structural methodologies to SCT to detect concepts with higher likelihood of errors. Bodenreider, Campbell, Pacheco, Rector, and their colleagues [37–40] applied lexical methodologies to SCT and its predecessors to detect or measure inconsistencies. Zhu et al. [41] provide a methodical review of possible auditing methodologies.

The Procedure hierarchy of SCT is the most semantically complex of the 19 hierarchies of SCT, currently with 23 potential defining attributes [4]. It is slated for a more significant role with the MU Stage 2 regulations [11]. Concepts in the hierarchy have an average of 2.4 unique attributes and 1.9 parents per concept (compared with 1.8 and 1.7, respectively, for Clinical finding). This makes the Procedure hierarchy a prime target to examine methods to explore and detect issues with SCT’s formal definitions.

3. Methods

Our core assumption is that within the realm of SCT, non-synonymous lexical representations are likely to have similar, but non-identical logical representations. In SCT, FSNs are
non-arbitrary terms that are designed to clarify the meaning of the concept [4]. Our methodology was based on the formation of “similarity sets.” We defined a similarity set as a collection of concepts where SCT fully specified names (FSNs) have lexical similarity. To further define the lexical similarity, each set had a base lexical description and all other lexical descriptors in the set differed from the base description by one word. Thus, for the base description Prophylactic upper limb stretching (procedure), Therapeutic upper limb stretching (procedure) and Prophylactic lower limb stretching (procedure) combine to form a similarity set with their associated logical representations. For the creation of similarity sets in this study we required each lexical string to be of at least five words in length (FSN) after removing certain stopwords [42]. As we were dealing with concept descriptions, we used a customized stopwords list based on the one published by the National Library of Medicine [43]. However, stopwords from the list were not ignored during the analysis phase.

We defined “inconsistency” in a set as any instance where at least one of its concepts could unequivocally incorporate conceptual modeling elements from any other concept in the set. Fig. 1 depicts a similarity set of two concepts: Conversion from uncemented total knee replacement and Conversion to uncemented total knee replacement. Both concepts are somewhat ambiguous (arguably the “from” more than the “to” one) since they do not indicate to, or from (respectively), what the conversions occur. Both concepts involve a total knee replacement (TKR) procedure and both are revisions since their FSNs indicate a transition between different types of TKRs. As both concepts are primitives, we cannot assume that all the defining information is present. Nevertheless, significant modeling discrepancies are evident. Although both concepts have a single parent, its type is different. The “from” concept is only linked hierarchically to Revision of knee arthroplasty even though logically, it must be some form of TKR. The “to” concept, although a revision, is not linked hierarchically to any revision-type parent, not even through an attribute. The “to” concept lacks the Revision status attribute but has the Procedure site – Indirect and the Direct device attributes with their assigned values. As for the assigned attribute values, although both concepts have the attribute Method, their respective assigned values differ: Surgical action for the “from” concept and Surgical insertion – action and Repair – action for the “to” concept. Surgical action is an ancestor of both Surgical insertion – action and Repair – action. The two possible Method values for the “to” concept also highlight that it has two attribute groups whereas the “from” concept has only one group. Thus, utilizing a similarity set of minimum size (two concepts), we demonstrate four different types of possible inconsistencies: hierarchical, attribute assignment, attribute values, and groups. Our findings are only minimally affected by the vagueness of the concepts or the auditor’s subjectivity.

We formulated five hypotheses:

Hypothesis 1. Similarity sets whose concepts exhibit different number of parents are more likely to harbor inconsistencies than randomly selected similarity sets.

Hypothesis 1.1. The inconsistency type is more likely to be hierarchical.

Hypothesis 2. Similarity sets whose concepts exhibit different number of attributes are more likely to harbor inconsistencies than randomly selected similarity sets.

Hypothesis 2.1. The inconsistency type is more likely to be attribute related.

Hypothesis 3. Similarity sets whose concepts exhibit different number of role groups are more likely to harbor inconsistencies than randomly selected similarity sets.

Hypothesis 3.1. The inconsistency type is more likely to be role-group related.

Hypothesis 4. Similarity sets whose concepts exhibit different number of parents, relationships, and groups are more likely to harbor inconsistencies than randomly selected similarity sets.

Hypothesis 5. Similarity sets with sibling association between any of their member concepts are likely to exhibit higher rate of inconsistencies than similarity sets without any sibling association between any member concepts.

Accordingly, and based on the inferred view [44] of the January 2011 release of SCT, we exhaustively formulated all possible similarity set combinations based on the Procedure hierarchy FSNs, subsumed and irrespective of the internal sequence of the lexical strings within a set. We used these similarity sets to randomly create five samples. Four corresponded with hypotheses one through four: samples 1 (Diff-Par), sample 2 (Diff-Rel), sample 3 (Diff-Grp), sample 4 (Diff-All), while the fifth sample (Control) served as a control sample, composed of concepts that differed from the base concept by one word without respect to the number of parents, relationships, or groups. We consider the Control sample a close equivalent to the sample used in [1]. Each sample consisted of randomly selected, 50 mutually exclusive similarity sets, controlled only for their respective parameter. In each similarity set (except Control), at least two concepts differed in the number occurrences of the sample’s main criteria.

The samples were presented (non-blinded, single spreadsheet) to, and evaluated by, a single auditor (GE), a physician with

---

**Fig. 1.** A two-concept similarity set from the Procedure hierarchy of SCT (using partial screenshots from the CliniClue Xplore browser).
extensive background in CBTs. The auditor did not look for errors but rather for clear inconsistencies between the inferred views of the concepts in a similarity set: hierarchical, definitional, attribute assignment, attribute target values, and groups. Within each set, the auditor looked for all types of inconsistencies. The CliniClue Xplore browser [45] was used to review the formal definitions.

4. Results

SCT’s January 2011 Procedure hierarchy contains 52,011 concepts. After removing stopwords and selecting FSNs of five remaining words or more, our algorithm utilized 26,980 unique concepts from the hierarchy (51.9%) for similarity sets. Overall, 4886 unique concepts were included in the 2111 similarity sets generated for the Procedure hierarchy, representing 9.4% of all concepts in the hierarchy, and 18.1% of all eligible concepts in the hierarchy. These sets formed the base for the selection of our five samples that included 250 sets containing 797 unique concepts. Table 1 provides general set information for the Procedure hierarchy while Table 2 summarizes the characteristics of each sample. None of the samples’ similarity sets was excluded due to irrelevant association between the concepts.

Table 3 summarizes our findings across the five samples. Control exhibited inconsistencies in 38% of the similarity sets. The non-Control samples exhibited inconsistency rates of 52–70%. For Diff-Rel, with 70% of inconsistent sets, this was a statistically significant difference compared to Control (Fisher’s exact test, two-tailed, \( p = 0.002 \)). Thus, our findings strongly confirm our second hypothesis: Concepts in a similarity set with different number of relationships have a higher likelihood of inconsistency. In Fisher’s exact test, we chose to use a strict statistical test. In fact, under the Chi-square test, the findings in all samples are statistically significant compared to Control (Fisher’s exact test, two-tailed, \( p < 0.05 \)). Therefore, it is essential to develop and implement a variety of procedures as none of the sample sets had to be excluded due to irrelevant association between the concepts.

The auditing process strictly looked for the five inconsistency types reported in [1] within each similarity set. Table 4 breaks down the inconsistency types found within concepts for the different samples. Set concepts from Diff-Par predominantly exhibited hierarchical inconsistencies (95.8%, \( p < 0.001 \)), whereas set concepts from Diff-Rel predominantly exhibited inconsistencies involving attribute assignments (98.1%, \( p < 0.001 \)) thus confirming Hypotheses 1.1 and 2.1, respectively. The results also demonstrate a meaningful correlation in the Diff-Par, Diff-Rel, and Diff-All sample concepts between hierarchical and attribute assignment issues.

5. Discussion

We start with the premise that lexically similar concepts are expected to exhibit similar modeling. The current study validates preliminary findings [1] that many of the similarly worded concepts in SCT’s Procedure hierarchy, are not modeled in a consistent manner (30–38% overall). Furthermore, the study indicates that concepts in similarity sets with difference in attributes are much more likely to be inconsistently modeled (70%). Additionally, our findings disprove a seemingly logical assumption that similarity sets with sibling association between their member concepts are more prone to inconsistencies (Hypothesis 5). Moreover, the findings suggest that algorithmic detection and resolution of inconsistencies is feasible, as will be discussed later.

While the inconsistency rate may seem high, it is essential to consider that we formulated highly specialized sets: concepts of at least five words to their FSNs that differ from the base concept by only a single word. These sets cover 9.4% of the concepts in the Procedure hierarchy and 18.1% of all length-eligible concepts. Our lexical algorithm proved itself successful in creating relevant similarity sets as none of the sample sets had to be excluded due to irrelevant association between the member concepts.

The results indicate that in the authoring process of SCT, very little attention is given to identify similar concepts, with little consideration to the importance of modeling them in a consistent manner. While the vast majority of the inconsistencies cannot be considered errors, as each individual concept conforms to SCT’s guidelines, they may pose significant obstacles to reasoning engines based on SCT’s modeling structure. This is not a trivial manner as Rector et al. [22] so amply demonstrated. Revisiting their Myocardial infarction example clearly illustrates how such deficiencies can interfere with meaningful utilization of data collected in clinical repositories: research queries may not return all relevant cases, decision support opportunities may be missed, analytics may be skewed, and clinical care can be affected. Campbell et al. discuss similar issues [38].

MU Stage 2 expands SCT’s role beyond the realm of clinical findings to procedures as well and it is reasonable to assume that its role will be expanded even further as MU progresses. Although in the current context of HITECH and MU, SCT serves mostly as a source for subsets and lists, it is hard to imagine that it was chosen only due to its lexical comprehensiveness. Naturally, the next step beyond using SCT’s concept descriptions in lists is taking advantage of SCT’s hierarchical structure and formal definitions. The true potential of any CBT is embodied in the knowledge captured within its semantic network [16–20]. SCT faces expectations to serve as an interface terminology and not only as a reference terminology [31–33,41,46]. In its current state, SCT cannot serve “as-is” in clinical applications even as a reference for limited sets [47,48]. It is expected that for use within clinical applications vendors will use well-curated subsets and that dedicated extensions will be developed. However, not all CEHRT vendors can purchase or invest resources to develop such subsets and extensions may diverge from each other in a manner that will be counter-productive for data interoperability. The IHTSDO invests significant effort in formulating SCT with DL for computational purposes. However, incomplete and inconsistent application results in a structure that is questionable for use except for the generation of SCT’s inferred view from the stated one.

Rector et al. [22] suggest that a comprehensive auditing effort is urgently needed, estimated at up to two years for the CORE subset. However, the CORE subset is just a small portion of SCT’s Clinical finding hierarchy. A broader auditing effort will require a much larger coordinated effort that may be beyond the reach of the IHTSDO. As SCT continues to grow, delays will complicate matters further. Therefore, it is essential to develop and implement a variety
of auditing methodologies that can be incorporated into the authoring process or routinely executed after the fact with high yield. As Wei and Bodenreider [33] concluded, DL classifiers cannot detect that which is not defined. Other methods are needed to complement the classifiers. Our analysis is independent of SCT’s DL-based infrastructure as it inspects, holistically, modeling elements of one concept and compares them to those of similar concepts. Inconsistencies of the types described in this study must be evaluated outside the realm of DL since, ultimately, SCT’s usefulness from an algorithmic and individual perspective will be judged by the consistency and sufficiency of its conceptual definitions [31–33,41,46].

Our study demonstrates that a simple lexical algorithm can very effectively detect similar concepts that are inconsistent in their hierarchical modeling utilizing differences in attributes as an indicator. Moreover, we believe that our methodology can be applied to other semantically rich SCT roots such as the Clinical finding (16 attributes) hierarchy with similar effectiveness. It is reasonable to expect that other such hierarchies harbor similar inconsistencies but that the effectiveness and yield of our method will decline with declining semantic complexity. Other algorithms, utilizing different and more sophisticated lexical methods and word length selection may improve on our results. However, additional methodologies could introduce noise and reduce specificity as discussed by Campbell et al. [38] and, with the current yield described throughout this study, we do not see an immediate need to employ such methodologies.

The present study opens the possibility for algorithmic enhancement of SCT’s formal definitions utilizing an indicator that was used to identify sets, i.e. different attribute assignments in similarly worded concepts. Although more than half of SCT’s concepts are not fully defined, we can reliably assume that the vast majority of them are not erroneous. Thus we posit that most of the additional attributes and attribute target values (when the attribute target value is not directly associated with the specific word that differentiates between the similar concepts) can be reasonably assigned to the other similarity set member concepts that lack them.

Consider the example in Fig. 2: for the purpose of this discussion we can ignore the differences in hierarchical modeling. The concept on the right lacks the Has specimen attribute. Adding this attribute with its target value to create the hypothetical concept depicted in Fig. 3 will be correct, improve the consistency of the modeling, and potentially contribute toward qualifying the concept as a fully specified concept. Other algorithmic approaches to identify possible missing attributes can be employed. For example, a method can detect that certain FSN words are not represented as an attribute target in the formal definition. In this case, “serum” is not present as an attribute target value. However, such a method may be less effective in proposing a possible resolution.

Our study was limited due to the use of a non-blinded, single auditor (GE). However, we consider the nature of the evaluation for inconsistencies as only minimally subjective, if at all, due to our definition of an inconsistency. For example, the consideration of a missing attribute, a yes/no type of decision, is algorithmically detectable. Furthermore, it is not likely that we identified missing attributes as false positives. It is more likely that our review process included a certain degree of false negatives as missed findings. Any bias towards a specific inconsistency type in its respective sample would have affected each sample in a similar manner while the auditor was instructed to exhaustively document all types of inconsistencies in each and every similarity set. For practical reasons our sample sets were only controlled for their main characteristic (between at least two concepts in the set) and did not exclude any bias towards a specific inconsistency type in its respective sample.

### Table 2
Sample characteristics.

<table>
<thead>
<tr>
<th>Set type</th>
<th>#Sets</th>
<th>#Cpts</th>
<th>Max #cpts</th>
<th>%Non-prim</th>
<th>%Leaf</th>
<th>Avg #par/cpt</th>
<th>Avg #rel/cpt</th>
<th>Avg #grp/cpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1: Diff-Par</td>
<td>50</td>
<td>149</td>
<td>9</td>
<td>29.5</td>
<td>71.8</td>
<td>1.9</td>
<td>2.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Sample 2: Diff-Rel</td>
<td>50</td>
<td>222</td>
<td>50</td>
<td>40.0</td>
<td>64.8</td>
<td>1.6</td>
<td>2.7</td>
<td>0.4</td>
</tr>
<tr>
<td>Sample 3: Diff-Grp</td>
<td>50</td>
<td>148</td>
<td>6</td>
<td>39.2</td>
<td>71.6</td>
<td>1.6</td>
<td>3.1</td>
<td>1.5</td>
</tr>
<tr>
<td>Sample 4: Diff-All</td>
<td>50</td>
<td>150</td>
<td>7</td>
<td>38.0</td>
<td>67.3</td>
<td>1.8</td>
<td>3.1</td>
<td>1.2</td>
</tr>
<tr>
<td>Sample 5: Control</td>
<td>50</td>
<td>128</td>
<td>5</td>
<td>22.6</td>
<td>79.7</td>
<td>1.3</td>
<td>2.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

### Table 3
Summary of findings per sample. (S-) denotes similarity sets without sibling association between any member concepts of a set. (S+) denotes similarity sets with at least one sibling association between any member concepts of a set.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Consistent sets</th>
<th>Inconsistent sets</th>
<th>Concepts</th>
<th>Inconsistent concepts</th>
<th>P-value (two-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># (S-/S+)</td>
<td># (%)</td>
<td></td>
<td># (%)</td>
<td></td>
</tr>
<tr>
<td>Diff-Par</td>
<td>21 (10/11)</td>
<td>58</td>
<td>149</td>
<td>48</td>
<td>32.2</td>
</tr>
<tr>
<td>Diff-Rel</td>
<td>15 (7/8)</td>
<td>70</td>
<td>222</td>
<td>54</td>
<td>24.3</td>
</tr>
<tr>
<td>Diff-Grp</td>
<td>24 (14/10)</td>
<td>52</td>
<td>148</td>
<td>38</td>
<td>25.7</td>
</tr>
<tr>
<td>Diff-All</td>
<td>22 (16/6)</td>
<td>56</td>
<td>150</td>
<td>49</td>
<td>32.6</td>
</tr>
<tr>
<td>Control</td>
<td>31 (16/15)</td>
<td>38</td>
<td>128</td>
<td>27</td>
<td>21.1</td>
</tr>
</tbody>
</table>

### Table 4
Breakdown of inconsistency types within concepts of inconsistent sets.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Inconsist cpts</th>
<th>Hierarchical</th>
<th>Definitional</th>
<th>Attrb assign</th>
<th>Attrb value</th>
<th>Groups</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tbl 5, Col 5</td>
<td>#</td>
<td>%</td>
<td>P-val</td>
<td>#</td>
<td>%</td>
<td>P-val</td>
<td>#</td>
</tr>
<tr>
<td>Diff-Par</td>
<td>48</td>
<td>46</td>
<td>95.8</td>
<td>&lt;0.001</td>
<td>5</td>
<td>10.4</td>
<td>17</td>
</tr>
<tr>
<td>Diff-Rel</td>
<td>54</td>
<td>24</td>
<td>44.4</td>
<td>0</td>
<td>0</td>
<td>53</td>
<td>98.1</td>
</tr>
<tr>
<td>Diff-Grp</td>
<td>38</td>
<td>24</td>
<td>63.2</td>
<td>4</td>
<td>10.5</td>
<td>6</td>
<td>15.8</td>
</tr>
<tr>
<td>Diff-All</td>
<td>49</td>
<td>20</td>
<td>40.8</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>49.0</td>
</tr>
<tr>
<td>Control</td>
<td>27</td>
<td>10</td>
<td>37.0</td>
<td>5</td>
<td>18.5</td>
<td>14</td>
<td>51.8</td>
</tr>
</tbody>
</table>
the main characteristics of other samples. We do not believe this significantly affected our analysis, as the results in Table 4 indicate.

SCT does not usually progress in giant leaps between releases. All of our 250 sample similarity sets continue to exist in the July 2013 release of SCT. Although certain changes occurred to the conceptual representations of concepts in 15 similarity sets, they did not produce any significant changes necessitating an update to our results; this indicates the applicability of our methodology to current and future versions of SCT.

In light of scarce auditing resources, we believe this methodology is suitable for use by a single reviewer and can be easily utilized during the authoring process. We propose that this and other complementary lexical and non-classifier methodologies be adopted by the IHTSDO as part of the editing process in conjunction with current methodologies as well as for the routine maintenance of the inferred view of SCT.

6. Conclusion

Lexically similar SNOMED CT concepts exhibit significant degree of inconsistent modeling. Attribute assignment inconsistencies may present an opportunity for algorithmic detection and enhancement. In light of SNOMED CT’s increasing stature in the U.S. and the significance of inconsistent formal definitions for decision making in healthcare, this and similar methods should be employed for routine maintenance of this prominent biomedical terminology. We plan to explore the applicability of our method in other semantically rich SCT hierarchies and the prospect of their algorithmic enhancement.

References


