The readiness of SNOMED problem list concepts for meaningful use of electronic health records

Ankur Agrawal a,⁎, Zhe He a, Yehoshua Perl a, Duo Wei b, Michael Halper c, Gai Elhanan d, Yan Chen e

a Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102, USA
b Computer Science and Information Systems, School of Business, Richard Stockton College of New Jersey, Galloway, NJ 08205, USA
c Information Technology Department, New Jersey Institute of Technology, Newark, NJ 07102, USA
d Halfpenny Technologies, Inc., 725 Skippack Pike, Blue Bell, PA 19422, USA
e Computer Information Systems Department, Borough of Manhattan Community College – CUNY, New York, NY 10007, USA

A R T I C L E   I N F O

Article history:
Received 4 May 2012
Received in revised form 5 March 2013
Accepted 17 March 2013

Keywords:
SNOMED CT
Problem list
Electronic Health Record
Meaningful use
Quality assurance

A B S T R A C T

Objective: By 2015, SNOMED CT (SCT) will become the USA’s standard for encoding diagnoses and problem lists in electronic health records (EHRs). To facilitate this effort, the National Library of Medicine has published the “SCT Clinical Observations Recording and Encoding” and the “Veterans Health Administration and Kaiser Permanente” problem lists (collectively, the “PL”). The PL is studied in regard to its readiness to support meaningful use of EHRs. In particular, we wish to determine if inconsistencies appearing in SCT, in general, occur as frequently in the PL, and whether further quality-assurance (QA) efforts on the PL are required.

Methods and materials: A study is conducted where two random samples of SCT concepts are compared. The first consists of concepts strictly from the PL and the second contains general SCT concepts distributed proportionally to the PL’s in terms of their hierarchies. Each sample is analyzed for its percentage of primitive concepts and for frequency of modeling errors of various severity levels as quality measures. A simple structural indicator, namely, the number of parents, is suggested to locate high likelihood inconsistencies in hierarchical relationships. The effectiveness of this indicator is evaluated.

Results: PL concepts are found to be slightly better than other concepts in the respective SCT hierarchies with regards to the quality measure of the percentage of primitive concepts and the frequency of modeling errors. There were 58% primitive concepts in the PL sample versus 62% in the control sample. The structural indicator of number of parents is shown to be statistically significant in its ability to identify concepts having a higher likelihood of inconsistencies in their hierarchical relationships. The absolute number of errors in the group of concepts having 1–3 parents was shown to be significantly lower than that for concepts with 4–6 parents and those with 7 or more parents based on Chi-squared analyses.

Conclusion: PL concepts suffer from the same issues as general SCT concepts, although to a slightly lesser extent, and do require further QA efforts to promote meaningful use of EHRs. To support such efforts, a structural indicator is shown to effectively ferret out potentially problematic concepts where those QA efforts should be focused.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

SNOMED CT (further abbreviated as “SCT”) has been endorsed as a premier clinical terminology by many international organizations. It is slated to play a significant role in the Health Information Technology for Economic and Clinical Health (HITECH) initiative [1] to adopt electronic health records (EHRs) by providing

standardized encodings of healthcare data. By 2015, SCT will become the standard terminology for EHR encoding of diagnoses and problem lists in the USA [2].

To facilitate encoding of problem lists, the National Library of Medicine has extracted a collection of UMLS concepts dealing specifically with health problems (see the study reported in [3]). As it happens, SCT was found to be the UMLS source offering the best coverage of this so-called UMLS clinical observations recording and encoding (CORE) problem list. The SCT portion (amounting to 81% coverage) was posted on-line in July 2011 as the “SCT CORE” problem list [4], comprising 5862 active concepts. It was accompanied by the alternative “Veterans Health Administration and Kaiser
Permanent (VA/KP) problem list [5], consisting of 16,622 active SCT concepts. The two lists have 4004 concepts in common. It is not clear which of them – or both – will be used for EHR encoding.

While problem list encoding is SCT’s primary and immediate contribution toward meaningful use of EHRs, the use of SCT, due to its inherent structure, should also support patient education and advanced clinical data repository queries that are among the other meaningful use objectives. Such intended use of clinical terminologies had been suggested previously in numerous studies [6–13]. However, there are indications that at the moment, SCT as a clinical terminology is not optimally structured for such use. For instance, it exhibits certain inconsistencies in its hierarchical and lateral relationship modeling, which forms its very foundation. As an example, the fully-defined concept Acute myocardial infarction is not hierarchically linked to the concept Ischemic heart disease, nor is there a physiological attribute linking it to the associated Myocardial ischemic process. A physician attempting to encode Acute myocardial infarction in an EHR is likely to search for it in an ischemic heart disease subsection and will not be able to locate it, as noted by Rector et al. [14]. In that same paper, several other examples show various modeling problems in SCT such as diabetes being classified as a disease of the abdomen and arteries of the foot being placed in the pelvis. It is concluded that without further quality-assurance (QA) efforts, clinicians may not realize the implications of what they are saying, researchers may not realize what their queries should retrieve, and post-ordination cannot be expected to be reliable, thus compromising the interoperability and meaningful use of SCT.

Additionally, many of SCT’s concepts are under-defined from a logical standpoint. As many as 77.5% of SCT’s concepts are primitive, i.e., lack the necessary relationships for full definition. In [15,16], issues with the part-of relationship, discrepancies in defined semantics, and definitional inconsistencies among ancestors and descendants were highlighted. A critical review of SCT’s logical and ontological issues has appeared [17]. Such problems with SCT’s structure may affect its use in decision-support and data mining. When large organizations have adopted SCT, e.g., KP’s Convergent Medical Terminology [18] and the VA’s Enterprise Reference Terminology [19], significant enhancements were required. Such enterprise-specific enhancements may not be compatible with SCT’s standards. Moreover, such efforts may not be easily replicated at smaller institutions. Also, on the HITSDO Special Interest Group Discussion Board [20] in 2010, there was a general agreement that “as is”, SCT is not suitable for use in patient-facing applications. Such issues are barriers to its successful deployment.

In this paper, we are interested in looking at the proposed problem lists with an eye toward their readiness for use in EHRs. We focus on the combination (union) of the SCT CORE and VA/KP problem lists (which we will denote as the “PL”) containing a total of 18,480 SCT concepts. We perform a study to examine the quality of the PL. In particular, we would like to know if the modeling of the PL’s concepts has reached a stable and correct state due to frequent use and the increased scrutiny on them – or whether the PL requires further QA efforts. We extract equally sized random samples of concepts from two different concept populations for analysis: the first consists of concepts strictly from the PL and the second contains general SCT concepts distributed proportionally to the PL’s in terms of their hierarchies.

The results of our analysis show that PL concepts suffer from the same issues as general SCT concepts, although to a slightly lesser extent, and thus indeed require further QA. This additional QA is especially warranted in view of the intended role of PL concepts for the meaningful use of EHRs.

Towards this end, we examine a structural indicator that can be used to ferret out concepts with higher potential for errors: one dealing with hierarchical relationship problems. The result of examining this indicator is reported.

Let us point out that specific problems were selected for inclusion in the CORE and VA/KP problem lists due to their frequency with respect to in-patient populations in the USA. However, similar problem lists may very well exist for healthcare in other countries. For example, one of the seven medical centers contributing to the CORE problem list is the Hong Kong Hospital Authority [4]. Hence, the need for a subset of frequent medical problems is typical for many countries. The particular PL being used herein is just an example.

2. Background

2.1. Systematized Nomenclature of Medicine – Clinical Terms (SCT)

SCT is a comprehensive clinical terminology that provides content and expressivity for clinical documentation and reporting [21]. Its 295,753 concepts (SCT January 2012 release) can serve as a reference point for applications doing data analysis, conducting outcomes research, designing treatment guidelines, etc. SCT has been utilized or proposed for use in a variety of settings. The American Academy of Ophthalmology, for example, has chosen SCT as its official clinical terminology [22]. See [23] for an extensive literature review regarding the use of SCT in clinical practice. In [24], another literature search sought to identify SCT applications in critical care. The findings revealed investigations of SCT or its actual use in: the representation of disorders of newborn infants [25], nursing flowsheets [26], allergic diseases and associated problems [27], the representation of common patient problems [28], anesthesia patient safety [29], and intensive care [30]. SCT has also been used in the automatic grouping of adverse drug reactions terms [31] and in the annotation of tissue microarray data [32].

Each SCT concept has a collection of descriptions (terms), including one fully specified name (FSN), along with a preferred term (PT) and possibly a set of synonyms. Concepts are organized in 19 broad hierarchies, such as Clinical Finding and Body Structure. Within a given hierarchy, concepts are linked by IS-A relationships. Aside from the IS-As, concepts also have attribute relationships, sometimes called “lateral relationships” (or simply “relationships”). For example, Ear problem has the relationship finding site to Ear structure. A concept is further classified as being either fully-defined or primitive. In the latter case, the definition is underspecified and does not allow for the automated detection of sub-concepts.

SCT and its precursors, with 50–70% coverage of concepts of interest [33,34], have consistently outperformed other sources. In [35], SCT was found to cover 88.4% of the diagnosis/problem list terms used by clinicians within a computerized physician order entry (CPOE) system. In 2004, the VA concluded that SCT has promise as a coding system for clinical problems [36]. A survey in 2010 indicated that 68% of users perceived coverage as satisfactory or better [37,38]. In [39], SCT was deemed suitable to provide standardized representations of information created by two interface terminologies, noting that enriching SCT semantics would improve representation of the external terms.

2.2. SCT as part of the HITECH initiative

Reaffirming convictions that electronic information systems are essential to improving healthcare [40], the HITECH component of the American Recovery and Reinvestment Act [1] was designed to jumpstart the transition of medical providers to electronic health information systems (EHRs) [41]. In the proposal for the initial HIT standards [2], SCT is to be used to “enable a user to electronically record, modify, and retrieve a patient’s problem list for longitudinal care (i.e., over multiple office visits).” To accelerate the adoption and
meaningful use of EHRs by providers, incentives and penalties were defined [2,42]. The encoding of problem lists of current and active diagnoses for at least 80% of all unique patients was proposed as one indication of meaningful use. Moreover, SCT is slated to become the exclusive coding system for problem lists by 2015 [2].

Defining meaningful use of EHRs will involve problem lists encoded with SCT [2,42]. This lead to the creation of the problem lists noted above. Let us note that in the e-prescribing domain, the FDA has adopted the VA/KP problem list to represent indications in electronic labels [5]. In an evaluation of medication indication phrases, SCT, as a whole, was found to provide 90.3% coverage, while its Clinical Finding hierarchy offered 79.5% coverage.

2.3. Adverse effects of SCT deficiencies

HITECH’s meaningful-use regulations include provisions for decision support [43] to improve performance on high priority health conditions and cover both clinical and patient related aspects. These regulations actually call for context-sensitive decision support. This could include, for example, a documentation form to be presented for patients with diabetes that would include a required section for the diabetic foot exam. When the same form is presented for patients without diabetes, the diabetic foot exam section would be omitted. Similarly, certified electronic health record technology (CEHRT) can suggest that a patient with diabetes should be referred to a diabetic foot screener. While such decision support can be achieved by simply hard-coding the linkage between a diagnosis to a specific form or activity, it is easy to see how hierarchical (IS-A) and especially lateral relationships in a controlled terminology would offer better support. The employment of controlled terminologies to enable such uses of decision support has been described in the past [6–13]. However, accurate and consistent modeling of IS-As and relationships is critical for dynamic, context-sensitive, secondary use of clinical controlled terminologies.

Much of the decision support proposed in meaningful-use regulations is diagnosis related. Thus, SCT’s major role in encoding problem lists stands to directly affect the ability of CEHRT to provide appropriate decision support. The fact that a concept is marked primitive indicates a potential deficiency in its relationship structure, which in turn may lead to its incorrect positioning in a hierarchy. Missing or incorrect relationships influence the inheritance of properties. CEHRT, trying to take advantage of SCT’s inherent structure, could thus be negatively impacted.

3. Methods

A study is conducted to assess various qualities and properties of the PL and determine how well its concepts are currently modeled in comparison with the rest of SCT. Two random samples of 50 concepts each are examined. The first sample is made up of concepts strictly from the PL. The second is composed of non-PL SCT concepts, but these are chosen directly in proportion to the distribution of concepts in the PL sample across hierarchies. That is, if 5% of the PL sample is from the Procedure hierarchy, then 5% of this second, so-called “proportional,” sample is randomly chosen from that hierarchy.

For both samples, we count the number of primitive concepts and parents. Also considered is the number of words in the concepts’ preferred terms. The samples are also used to evaluate the correctness of the modeling of the concepts. The proportional sample serves as a control. Every concept in the samples is visually reviewed within the CliNiClue browser [44] graphical user interface. The CliNiClue browser provides a hierarchical and semantic neighborhood view of a desired focal concept. Utilizing CliNiClue, the reviewer is able to traverse the hierarchical tree in both directions as well as evaluate each of the available relationships, groups of relationships, and their respective target values. For each and every sample’s concept, all these relationships are analyzed. Findings are recorded according to a four-point scale of presumed significance: none, mild, moderate, and severe. Specifically, “severe” is assigned in cases of obvious errors in the hierarchy or relationship targets, i.e., the parent or relationship target is absolutely wrong. For example, the concept Occlusion and stenosis of posterior cerebral artery is not a Basilar artery occlusion.

“Moderate” is used for correct but overly broad or redundant parents or relationship targets that could be removed or replaced by more specific values. An example of a “moderate” finding is exhibited by the concept Benign neoplasm of skin of umbilicus (disorder), from the PL sample, having three parents. Two of the parents are Benign neoplasm of skin of trunk, excluding scrotum and Benign neoplasm of skin of abdomen. These two parents are siblings related to their shared parent Benign neoplasm of skin of trunk. Considering that the focus of the concept is the umbilicus, which is located in the abdomen, and that the trunk is composed of the abdomen and the thorax, the parent Benign neoplasm of skin of trunk, excluding scrotum seems redundant and too broad. This is supported by the fact than in SCT, Umbilical structure is considered a Structure of central region of abdomen and not that of the trunk.

An assignment of “mild” denotes cases of missing knowledge elements such as missing parents or missing children. Such cases are classified as “mild” since no information is incorrect but some desirable information is not given in the current modeling. We note that these levels are not based on clinical significance but rather the degree of deviation from appropriate modeling compared to other SCT concepts. The concept on examination – nipple normal (finding) is an example of a “mild” finding. It does not, but possibly should, have a has interpretation relationship with the target value of normal and an interprets relationship with a target of nipple, as we find for the concept on examination – reflexes normal (finding) whose respective values for those relationships are normal and reflex. While these relationships and related targets are desirable for consistent modeling within SCT, their absence does not fundamentally constitute an error.

A theme we discovered in our SCT auditing research is that “complex” concepts typically have higher likelihood of errors than “regular” concepts. One example of a type of complex concept for which studies confirmed such a tendency is a concept residing in a region of strict inheritance (of relationships) [45], defined in the context of an abstraction network for an SCT hierarchy [46]. Another example is overlapping concepts, defined with respect to elements of another kind of SCT abstraction network [47,48].

The number of parents of a concept can also be taken as a parameter of complexity in the sense that a concept with multiple parents is a specialization of each and inherits their properties, making it a build-up of knowledge coming from multiple paths. Such a concept reflects multiple identities, being “a kind of this and a kind of that.” Such an observation led us to wonder if multiple parentage is an indicator of problems. That is, do concepts with multiple parents in the PL tend to have more problems than those with one parent? We analyze any errors found in the abovementioned review for the PL sample and the proportional sample with respect to the number of parents to see if multiple parentage is indeed an indicator. In fact, we consider the following two hypotheses:

Hypothesis 1. PL concepts with multiple parents are more likely to be in error than concepts with one parent.

Hypothesis 2. The likelihood of errors in PL concepts increases with the number of the parents.
In Table 1, we list the distribution of the PL concepts according to their numbers of parents. For example, 7823 concepts (42.3%) in the PL have exactly one parent. Column 4 gives the cumulative amount of concepts with "n or more parents." For example, 100 concepts have at least seven parents. There is one PL concept, *Granaloma inguinale (disorder)*, with 15 parents, and another, *Menkes kinky-hair syndrome (disorder)*, with 12. The average number of parents for a PL concept is 1.90.

To test our hypotheses, we further study random samples of PL concepts generated according to their numbers of parents. In particular, we audit six samples of 50 random concepts of the PL. For a given n (1 ≤ n ≤ 6), Sample n has 50 concepts, each having n parents. Note that there are only 47 concepts with seven parents in the PL, so all those are audited in a seventh sample. We also audit in an eighth sample all concepts with eight or more parents, the number of which is 53 (Table 1, Column 4). In order to assess Hypothesis 1, we additionally study a random sample of 250 PL concepts with only one parent each. In this audit, we limit ourselves to errors in the IS-A relationships, that is, either incorrect or missing parents or children of a concept. The importance of such hypotheses, if proven correct, is that they can lead to efficient indicators of hierarchical problems in PL concepts.

4. Results

4.1. Modeling errors

To test the accuracy of the modeling of the PL concepts, a review of the two samples was carried out by one of the authors (GE), an MD with extensive experience auditing terminologies. The SCT’s July 2011 release was used for our study. Table 2 shows the results of our study of the modeling correctness of the PL and proportional samples. In total, 17 problems were found for each. No severe problems were found for the PL. Eleven concepts displayed moderate issues, and six exhibited mild ones. From the proportional sample, four concepts displayed severe problems, six exhibited moderate issues, and seven displayed mild issues.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL concept distribution according to the number of parents.</td>
</tr>
<tr>
<td># Parents (n)</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>Total:</td>
</tr>
</tbody>
</table>

Let us note that the number of primitive concepts (which are not fully defined with all their necessary lateral relationships) in the PL sample was 29 (amounting to 58% of the concepts). For the proportional sample, it was 31 (62%). So, overall, the two samples are similar with regard to the percentages of primitive concepts. These percentages may be an indicator to the accuracy of the modeling.

A PL sample concept with moderate finding is *Lumbosacral spondylosis without myelopathy*. The concept has four parents: *Lumbosacral spondylosis, Disorder of trunk, Degenerative disorder, and Spondylosis without myelopathy*. Some of these are clearly related, while other seems to be defined at the wrong level. For example, why is Disorder of trunk a parent at this refined level? Should it not be defined at a higher level as parent of Lumbosacral spondylosis? The same can be said of Degenerative disorder as a parent of the grandparent Spondylosis. See Fig. 1 for the modeling of the parents of Lumbosacral spondylosis without myelopathy, before and after QA.

An example of severe problems from the proportional sample can be found with *Cystic adventitial disease of popliteal artery (disorder)*, which has the following parents: *Vascular disease of abdomen and Systemic arterial finding*. However, the popliteal artery is not an abdominal artery, and this is not a systemic finding. Thus, these two parents were considered inappropriate. Additionally, the concept has the relationship *associated morphology* with the target *Cystic medial necrosis*. But cystic adventitial disease, a rare disorder, is not characterized by cystic medial necrosis but rather by mucinous cysts in the outer media or adventitia that progressively compromises the arterial lumen. Thus, this target value was also considered inappropriate. Altogether, these findings resulted in a “severe” rating due to both inappropriate parents and relationship targets. We also note that in SCT such concepts are not generally associated with a laterality attribute. However, from a clinical perspective (and aside from the clinical knowledge used in our review or that of a presumed clinical user), the SCT content does not provide any clues to the fact that this concept should be associated with “left,” “right,” or “bilateral” modifiers.

As we see, the severity of the modeling problems in the proportional sample is slightly higher than for the PL sample, as it is with regards to percentage of primitive concepts. Thus, this preliminary study indicates that probably due to their frequent use, the PL concepts are to some extent more accurate in their modeling than other concepts in their respective hierarchies. They are still far from satisfying the expected needs of coding diagnoses and problem lists for longitudinal care in EHRs due to their high percentage of primitive concepts.

<table>
<thead>
<tr>
<th>4.2. Number of parents as an indicator of errors</th>
</tr>
</thead>
</table>
| To see if the number of parents is a possible indicator of higher error rates, Tables 3 and 4 show the distribution of errors in the PL sample and the proportional sample, respectively, with regard to number of parents. We combined the moderate and severe errors into one category; the mild errors are in a separate category to facilitate a comparison with the proportional sample. For example, in the PL sample, there are 24 concepts with exactly two parents. Five of them had mild errors, while another five had moderate or severe errors. In total, 41.7% of such concepts exhibited errors. For the 35 concepts with multiple parents, there are a total of 15 (42.9%) in error versus just two (13.3%) among the 15 concepts with one parent. There are five erroneous concepts (45.4%) for those with at least three parents. Interestingly, all these errors were moderate; none were mild. These results are in line with Hypothesis 1. Another interesting observation is that three out of four concepts with at least four parents have moderate errors (75% error rate). However, this sample is too small for the evaluation of Hypothesis 2. For the proportional sample, the findings are, in general, similar to those for the PL sample (with the exception of some errors being
severe) with a 21.7% error-rate for concepts with one parent, 44.4% for concepts with two or more parents, and 70% for concepts with at least three parents. We also see that the error-rate for concepts with four or more parents is 50%, with one of the two concepts exhibiting a severe error level and the other exhibiting a moderate error level. This distribution also supports Hypothesis 1. But, again, there is not enough data for the evaluation of Hypothesis 2.

In Table 5, we present the results of auditing the six random samples of 50 concepts of the PL derived based on the number of parents for 1 ≤ n ≤ 6, the collection of 47 concepts having seven parents, and the collection of 53 concepts with eight or more parents. For example, among the 50 concepts with three parents examined, four were found to be in error. Moreover, a total of 10 errors were discovered, yielding a rate of 2.5 errors per erroneous concept.

In addition, we audited a random sample of 250 PL concepts with only one parent. Both samples were reviewed by one of the authors (YC), who has training in both medicine and terminologies and extensive experience in auditing terminologies. In Table 6, we compare the results of this auditing to those for the sample of 250 PL concepts obtained by aggregating the results of Samples 2–6 from Table 5. The purpose of this comparison is to further test Hypothesis 1. The results in Table 6 support this hypothesis in two ways. First, we see a much higher percentage of erroneous concepts in the case of multiple parents versus one parent: 27% and 8%, respectively. Second, the multi-parent concepts display about 40% more errors per erroneous concept on average than do single-parent concepts. The difference is statistically significant according to the Chi-squared test (with p < 0.0001).

Even though, as seen in Table 5, there is a trend of increasing errors with the number of parents, it is not strictly monotonic. For example, the sample of two-parent concepts shows three erroneous concepts as compared to four for the sample of single-parent concepts. Similarly, the collection of 47 concepts with seven parents exhibits 15 erroneous concepts as compared to the 28 for the sample of 50 concepts with six parents.

To capture this general trend, we aggregate the results of Table 5 as shown in Table 7. The samples of concepts with 1–3 parents are combined in the first row, those with 4–6 parents are in the second row, and 7–15 are at the bottom. The absolute number of errors in the 1–3 group, as compared to either the 4–6 or the 7–15 group, is significantly lower (Chi-squared test with p < 0.0001 and p < 0.001, respectively). Note that such p values for Chi-squared tests indicate statistical significance. Accordingly, the percentages of erroneous concepts for the second group (4–6 parents) and the third group (7–15 parents) are much higher than for first group (1–3 parents). Furthermore, the average number of errors per erroneous concept in the 4–6 sample is larger, with a ratio of two to the 1.26 (≈2.17/1.72). The average number of errors per erroneous concept in the 7–15 sample is also larger, with a ratio of 1.22 (≈2.10/1.72). Hence, Hypothesis 2 is confirmed. Note that the high percentage of errors does not continue to grow for the extremely high number of parents (7–15).

Next, we consider the distribution of these errors according to their severity. We found a total of 129 erroneous concepts in the 600 concepts (250 concepts with one parent and 350 concepts with more than one parent) that were audited. Note that a concept could

---

**Table 3**

Errors found in the PL sample of 50 concepts.

<table>
<thead>
<tr>
<th># Parents</th>
<th># Concepts</th>
<th>Concepts with mild errors</th>
<th>Concepts with moderate + severe errors</th>
<th>Total concepts with errors</th>
<th>% Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>13.3</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>5</td>
<td>2</td>
<td>10</td>
<td>41.7</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>–</td>
<td>2</td>
<td>2</td>
<td>28.6</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>–</td>
<td>2</td>
<td>2</td>
<td>66.7</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**Table 4**

Errors found in the proportional sample of 50 concepts.

<table>
<thead>
<tr>
<th># Parents</th>
<th># Concepts</th>
<th>Concepts with mild errors</th>
<th>Concepts with moderate + severe errors</th>
<th>Total concepts with errors</th>
<th>% Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>21.7</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>29.4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>83.3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>33.3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>100.0</td>
</tr>
</tbody>
</table>
5. Discussion

In this study, we set out to investigate whether and to what extent the concepts of the PL, the problem lists derived from SCT, suffer from the deficiencies that SCT in general is known to suffer from and its potential effect, regarding SCT’s ability to contribute towards advanced decision support that relies on its conceptual representations. Of particular interest was whether due to frequent use and increased scrutiny, the concepts of PL are of better quality. According to the results of our study, the PL concepts show less severe modeling errors than those concepts in the proportional sample, derived with an eye toward the major hierarchies covered in the PL. This may be due to more attention paid in their initial modeling or improvement resulting from users’ feedback.

In this study, we used the proportional sample rather than a general SCT sample as a control sample. The reason is that the accuracy of modeling may differ from one hierarchy of SCT to another. Thus, if we obtain a difference between the percentage of modeling errors of the PL sample and a general SCT sample, we will not know if the difference is due to different modeling quality of PL concepts or different modeling qualities for various hierarchies.

This difference among hierarchies may be illustrated by the percentage of primitive concepts in various samples and hierarchies. As mentioned earlier, the percentage of primitive concepts may be an indicator of the quality of modeling since these concepts lack some lateral relationships. For example, there are 52.5% primitive concepts in the PL but as many as 77.5% in the whole SCT. As it happens, more than 90% of PL concepts are from the Clinical Finding hierarchy, with 53.8% of these being primitive. Among the entire Clinical Finding hierarchy, we find that 59.3% of the concepts are primitive. As we see the percentage of primitive concepts in the PL (52.5%) differs largely from its value in SCT (77.5%), while for the Clinical Finding hierarchy, the respective percentage values are much closer (53.8% versus 59.3%). A similar situation exists for other hierarchies contributing to the PL such as Situation and Procedure. Hence, using the proportional sample gives a better estimate of the relative difference of PL concepts from the concepts of SCT at large.

However, even with better modeling, the PL concepts still suffer from the same problems as those in the general SCT population, just to a slightly lesser extent. For effective secondary meaningful use, extensive efforts to improve the relationship modeling need to be made. Such efforts are complex and are expected to demand extensive editorial resources. Note that progress in accurate relationship modeling is also expected to manifest itself in a decrease in the number of primitive PL concepts, currently amounting to approximately 60% of its overall content.

As partial remedies for the findings of this study, we presented a structural indicator that can help to optimize the effectiveness of QA work on the PL concepts. This indicator, namely, the simple measure of the number of parents, deals with general modeling problems. Let us note that according to a recent study of SCT users [49], about 85% found “severe” errors (i.e., obvious errors in the hierarchy or relationship targets) to be somewhat bothersome. Out of these, 60% were very much bothered about incorrect parents. This indicator can guide the ordering of the QA efforts starting with concepts having extremely high numbers of parents and working downward from there.

In our audit of a sample of 400 concepts with various numbers of parents, summarized in Table 7, we found 102 erroneous concepts out of the 250 concepts (40.8%) with at least four parents. There were just 11 erroneous concepts out of the 150 concepts (7.3%) with less than four parents. Furthermore, they were severe errors, and the ratio of errors per erroneous concept for those concepts with at least four parents was 2.13 (=218/102) versus a ratio of 1.72 (=19/11) for those with less than four parents.

Based on the percentages of erroneous concepts and error ratios reported in Table 5, one can calculate a weighted estimate of the expected findings for the entire set of 1302 PL concepts with at least

Table 5
Error concentration in samples of concepts with different numbers of parents.

<table>
<thead>
<tr>
<th># Parents</th>
<th># Concepts</th>
<th># Erroneous concepts</th>
<th>% Erroneous concepts</th>
<th># Errors</th>
<th>Avg # errors per erroneous concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>4</td>
<td>8</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>1.0</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>2.5</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>8</td>
<td>16</td>
<td>11</td>
<td>1.4</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>24</td>
<td>48</td>
<td>46</td>
<td>1.9</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>28</td>
<td>56</td>
<td>73</td>
<td>2.6</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>15</td>
<td>32</td>
<td>23</td>
<td>1.5</td>
</tr>
<tr>
<td>≥8</td>
<td>53</td>
<td>27</td>
<td>51</td>
<td>65</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 6
Error concentrations in samples of concepts with one parent vs. concepts with 2–6 parents.

<table>
<thead>
<tr>
<th># Parents</th>
<th># Concepts</th>
<th># Erroneous concepts</th>
<th>% Erroneous concepts</th>
<th># Errors</th>
<th>Avg # errors per erroneous concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>250</td>
<td>20</td>
<td>8</td>
<td>30</td>
<td>1.5</td>
</tr>
<tr>
<td>2–6</td>
<td>250</td>
<td>67</td>
<td>27</td>
<td>143</td>
<td>2.1</td>
</tr>
</tbody>
</table>

Table 7
Error concentrations for aggregate samples of concepts.

<table>
<thead>
<tr>
<th># Parents</th>
<th># Concepts</th>
<th># Erroneous concepts</th>
<th>% Erroneous concepts</th>
<th># Errors</th>
<th>Avg # errors per erroneous concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–3</td>
<td>150</td>
<td>11</td>
<td>7</td>
<td>19</td>
<td>1.72</td>
</tr>
<tr>
<td>4–6</td>
<td>150</td>
<td>60</td>
<td>40</td>
<td>130</td>
<td>2.17</td>
</tr>
<tr>
<td>7–15</td>
<td>100</td>
<td>42</td>
<td>42</td>
<td>88</td>
<td>2.10</td>
</tr>
</tbody>
</table>
four parents (see Table 1). Using the data from Table 1 (Column 2) and Table 5 (Column 4), we can calculate the estimate as follows:

\[
809 \times 0.16 + 291 \times 0.48 + 102 \times 0.56 + 47 \times 0.32 + 53 \times 0.51 = 129 + 140 + 57 + 15 + 27 = 368
\]

Therefore, we can expect 28.2\% (368/1302) of the concepts with four or more parents to be in error. The weighted estimate of the number of errors for these expected 368 erroneous concepts can be calculated from the above formula and the data in the last column of Table 5 as:

\[
129 \times 1.4 + 140 \times 1.9 + 57 \times 2.6 + 15 \times 1.5 + 27 \times 2.4 = 682
\]

Hence, their expected error ratio is 1.85 (682/368). This estimate emphasizes the power of this indicator to derive a high yield of severe errors for a modest amount of auditing effort.

In contrast, if one were to audit all 17,178 PL concepts with less than four parents, the weighted estimate of the number of erroneous concepts is just 1237 (7.2\%), with only 1855 total errors and an error ratio of 1.5. Considering the extensive efforts required for such a large audit, this expected yield would likely not warrant it.

On the other hand, one can limit the audit to the 493 concepts with at least five parents. The weighted expected number of erroneous concepts in this case is 293 (48.5\%). The corresponding number of errors is 502, with an error ratio of 2.1.

As can be seen, the “number of parents” indicator represents a trade-off between the extent of the QA efforts and the results. By starting with concepts with the highest number of parents, a more limited effort can still provide a relatively high yield in terms of the relative number of errors, even though the total number of errors is smaller. Considering the typical situation with QA resources, such a trade-off is quite reasonable.

6. Limitations and future work

This paper’s results indicate problems that are very important to the current and future success of the meaningful use of EHRs in the HITECH initiative. These results suggest the need for broader and more refined follow-up studies. The future studies should include larger samples of PL concepts and control concepts from the SCT hierarchies contributing to the PL.

In addition, deeper efforts are needed in assessing and improving the PL concepts’ modeling. In the current study, a PL concept was reviewed together with its neighboring knowledge. However, our experience in group-based auditing (e.g., [50]) says that more problems are exposed when a concept is reviewed in a context of a group of similar concepts.

The number of parents is a simple indicator of a degree of complexity. In previous work, we have utilized much more sophisticated complexity indicators (e.g., as described in Section 3 and in [45,51]). We plan to study the effectiveness of combining multiple indicators in the context of the PL to guide QA efforts. It would be interesting to determine whether Hypothesis 1 or 2 holds for SCT concepts in general and not just for the PL concepts.

7. Conclusion

SNOMED CT is slated to become the standard for encoding problem lists in EHRs by 2015. Toward that end, two problem lists comprising SNOMED CT concepts have been published. We have performed a study to examine the readiness of the concepts in these problem lists for their intended meaningful use in EHRs. It was found that these concepts tend to suffer from the same problems as the concepts found throughout the general SNOMED CT content, just to a slightly lesser degree. Such problems include a high percentage of primitive concepts (likely to be missing relationships), and deficient and inconsistent modeling of relationships. Our conclusion is that further QA efforts are needed for the problem lists’ concepts. Leaving the problems unaddressed will have deleterious effects on secondary meaningful use of EHRs, a hallmark of the HITECH initiative. To help guide such QA efforts, we examined a straightforward structural indicator that can be used to ferret out concepts with potential errors. It was shown to be good in dealing with hierarchical relationship problems. This indicator was shown to be effective in targeting concepts with higher likelihood of errors for QA efforts.

References


