Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies

Christopher Ochs, James Geller, Yehoshua Perl, Yan Chen, Junchuan Xu, Hua Min, James T Case, Zhi Wei

ABSTRACT

Objective Standards terminologies may be large and complex, making their quality assurance challenging. Some terminology quality assurance (TQA) methodologies are based on abstraction networks (AbNs), compact terminology summaries. We have tested AbNs and the performance of related TQA methodologies on small terminology hierarchies. However, some standards terminologies, for example, SNOMED, are composed of very large hierarchies. Scaling AbN TQA techniques to such hierarchies poses a significant challenge. We present a scalable subject-based approach for AbN TQA.

Methods An innovative technique is presented for scaling TQA by creating a new kind of subject-based AbN called a subtaxonomy for large hierarchies. New hypotheses about concentrations of erroneous concepts within the AbN are introduced to guide scalable TQA.

Results We test the TQA methodology for a subject-based subtaxonomy for the Bleeding subhierarchy in SNOMED’s large Clinical finding hierarchy. To test the error concentration hypotheses, three domain experts reviewed a sample of 300 concepts. A consensus-based evaluation identified 87 erroneous concepts. The subtaxonomy-based TQA methodology was shown to uncover statistically significantly more erroneous concepts when compared to a control sample.

Discussion The scalability of TQA methodologies is a challenge for large standards systems like SNOMED. We demonstrated innovative subject-based TQA techniques by identifying groups of concepts with a higher likelihood of having errors within the subtaxonomy. Scalability is achieved by reviewing a large hierarchy by subject.

Conclusions An innovative methodology for scaling the derivation of AbNs and a TQA methodology was shown to perform successfully for the largest hierarchy of SNOMED.

INTRODUCTION

Biomedical terminologies, such as the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT, SCT for short), are standards used to support electronic health record (EHR) encoding, meaningful use, and interdisciplinary research and many other applications. Quality assurance is an important part of standards maintenance. However, the size and complexity of modern biomedical terminologies makes terminology quality assurance (TQA) difficult, requiring manual reviews of thousands of concepts by domain experts. Resources for comprehensive content reviews are limited. Effective methodologies are needed to target portions of a terminology that are more likely to contain errors, to increase TQA yield, measured by the ratio of number of errors corrected to number of concepts reviewed.

To support effective maintenance, we have developed a theory of abstraction networks (AbNs) to summarize the content and structure of various terminologies. In particular, we have shown that the area taxonomy, partial-area taxonomy, and disjoint partial-area taxonomy kinds of AbNs, support maintenance of SCT. Taxonomies partition concepts into groups, based on similar relationship structure and semantics. Certain concept groups contain uncommonly classified or complex concepts. These groups were found more likely to contain errors when compared to control groups.

In our previous studies, taxonomies were used to support maintenance of the Specimen hierarchy of SCT, and the Biological process hierarchy of the National Cancer Institute thesaurus (NCIt). However, when we applied the same methodologies to large SCT hierarchies, for example, Procedure and Clinical finding with 53 147 and 99 440 concepts, respectively, we encountered significant issues that inhibited our taxonomy-based TQA approach. First, the Procedure and Clinical finding taxonomies contain 10 828 and 10 614 groups, respectively, too many to review individually. Secondly, thousands of concepts were categorized by too general groups. These issues make our previously developed TQA methodologies impractical for large hierarchies.

To achieve scalability of our methodologies, we have developed taxonomy subsets, called subject-based subtaxonomies, which summarize a subhierarchy that covers a particular subject. A subtaxonomy is created by selecting a concept, for example, Bleeding or Heart disease, and all of its descendants. Since quality assurance for a whole hierarchy is not practical, we observed that auditors usually concentrate on subjects of high interest. Subject-based subtaxonomies allow an auditor to focus on manageable portions covering specific subjects of a large hierarchy.

Following our study of overlapping concepts (see ‘Background’ section), we test the hypothesis that overlapping concepts in a subject-based subtaxonomy of a large SCT hierarchy are more likely to have errors than non-overlapping concepts. Furthermore, we investigate three new hypotheses that identify groups of concepts that are even more likely to have errors. A methodology based on these hypotheses is introduced, which enables an auditor to obtain a high yield of corrections with a relatively small effort.

To cite: Ochs C, Geller J, Perl Y, et al. J Am Med Inform Assoc. Published Online First: [please include Day Month Year]. doi:10.1136/amiajnl-2014-003151
To test our methodology, we derive subject-based subtaxonomies for the Bleeding (and Cancer) subhierarchies in Clinical finding. We target Clinical finding due to its importance for standards encoding and widespread use. Three domain experts reviewing a sample of 300 concepts from the Bleeding subtaxonomy identified 87 erroneous concepts. The subject-based subtaxonomy quality assurance methodology is shown to uncover statistically significantly more erroneous concepts than a control experiment. Applying this methodology to many subjects, one at a time, can cover a substantial portion of a large hierarchy without overwhelming an auditor.

BACKGROUND

SNOMED CT TQA techniques

TQA is an important part of a terminology’s lifecycle. Zhu et al. provide a comprehensive survey of manual, semi-automatic, and automatic TQA methodologies. SCT is a common target for TQA studies because of its importance. Jiang and Chute audited the semantic completeness of SCT concepts. Semantic, structural, and ontological techniques are offered by Rector et al who identified major types of errors in SCT that were caused by problems in Description Logic modeling or in concept classification. Schulz et al analyzed SCT’s ‘health’ from an ontological and logical perspective, identifying major problem areas.

Ceusters et al. describe an ontology-based technique that utilized an external ontology. Mortensen et al. describe a crowd-sourcing TQA methodology for verifying the correctness of axioms in SCT. Agrawal et al. utilized a combination of lexical and structural techniques to identify inconsistently modeled concepts. Bodenreider et al. analyzed SCT’s IS-A hierarchy, found under-defined classes, and used lexical methods to evaluate consistency among SCT’s terms.

SNOMED CT taxonomy AbNs

In previous work, we have developed various kinds of AbNs to support TQA for standards terminologies, for example, NCIt and SCT. AbNs are compact, hierarchical networks composed of nodes and links. Nodes summarize groups of similar concepts and links summarize hierarchical relationships between related groups. AbNs have been shown to highlight groups of concepts that are more likely to contain errors. We utilized the following three kinds of AbNs for SCT to uncover errors. Briefly reviewed below are: area taxonomy, partial-area taxonomy, and disjoint partial-area taxonomy. For a glossary of terms see online supplementary appendix I.

Figure 1A shows an excerpt of 29 concepts from the Clinical finding hierarchy. An area summarizes all concepts with the same condition, diseases, or traits. The relationships between concepts are represented by the links. Only the types of the relationships are considered; target concepts are disregarded. An area taxonomy is a network where the nodes are areas and areas are connected by child-of links based on the underlying IS-A relationships. A root of an area is a concept that has no parents in its area. Areas are disjoint.

The partial-area taxonomy is a refinement of the area taxonomy. The (possibly multiple) root concepts of each area are used to define partial-areas. A partial-area consists of a root concept and all of its descendants in the area. Partial-areas summarize semantically similar concepts within each area, since all are descendants of this root. In a partial-area taxonomy the nodes are partial-areas, which are connected by child-of links. Figure 1B shows the area taxonomy for the excerpt of figure 1A.

The partial-area taxonomy is a refinement of the area taxonomy. The (possibly multiple) root concepts of each area are used to define partial-areas. A partial-area consists of a root concept and all of its descendants in the area. Partial-areas summarize semantically similar concepts within each area, since all are descendants of this root. In a partial-area taxonomy the nodes are partial-areas, which are connected by child-of links. Figure 1B shows the area taxonomy for the excerpt of figure 1A.

Partial-areas sometimes overlap, that means, a concept can be contained in multiple partial-areas. This happens whenever a concept is a descendant of multiple roots. In figure 1A the concepts Hemorrhage of kidney and Bladder hemorrhage, with two parents each, are descendants of two roots: Genitourinary tract hemorrhage and Hemorrhage of abdominal cavity structure. In figure 1C, the partial-areas defined by these roots both contain Hemorrhage of kidney and Bladder hemorrhage, which explains why the number of concepts in the (Associated morphology, Finding site) area (5) is smaller than the sum of the concepts summarized by each partial-area (3+4=7). A concept that is summarized by more than one partial-area is called an overlapping concept. Summarization with overlapping concepts is not desirable.

To make sure that concepts appear only in one partial-area, we developed the disjoint partial-area taxonomy, which partitions the concepts of an area into disjoint partial-areas. ‘Disjoint’ means ‘free of overlap.’ For a formal definition see Wang et al. Intuitively, disjoint partial-areas are ‘carved out’ from partial-areas with overlapping concepts, thus eliminating the overlap. The roots of disjoint partial-areas are called ‘overlapping roots.’

In figure 2B, Genitourinary tract hemorrhage, Hemorrhage of abdominal cavity structure, Gastrointestinal hemorrhage, and Hemorrhage of anastomosis are the roots of the (Associated morphology, Finding site) hierarchy. Hemorrhage of kidney, Lower gastrointestinal hemorrhage, and Gastric hemorrhage are overlapping (since they are overlapping concepts) roots, because all of their parents are non-overlapping concepts. Hematoma of kidney and Gastrjejunal ulcer with hemorrhage are also overlapping roots, because they are descendants of different roots than their parents.

Level 1 disjoint partial-areas (single colored) are referred to as non-overlapping disjoint partial-areas, since their concepts are non-overlapping in the partial-area taxonomy. Disjoint partial-areas at higher indexed levels (multi-colored) are referred to as overlapping disjoint partial-areas, since their concepts overlap between multiple partial-areas in the partial-area taxonomy.

Figure 2C shows the disjoint partial-area taxonomy for the concepts in figure 2A. When compared to the partial-area taxonomy, the disjoint partial-area taxonomy provides a more accurate summary of the subhierarchies of concepts within an area when these subhierarchies overlap.

These three taxonomies were shown to support SCT maintenance for example, small partial-areas were found to contain more errors than large partial-areas and overlapping concepts were found to contain more errors than non-overlapping concepts. However, the taxonomy for the large Clinical finding hierarchy has 10 614 partial-areas in 357 areas, which is too large for effective quality assurance.

Ochs et al. described the relationship subtaxonomy, which displays subsets of structurally similar groups from large partial-area taxonomies. This method was successfully used in support of TQA for the large Procedure hierarchy. However, anecdotal evidence reveals that terminology curators do not consider groups of concepts with structural similarity but groups of concepts from a specific subject (topic), for example, Bleeding or Lung cancer.

To support TQA, we developed the Biomedical Layout Utility for SCT (BLUSNO), a software tool for creating, visualizing, and exploring taxonomies. BLUSNO enables an auditor to view
a concept in the context of a local neighborhood or a partial-area taxonomy. BLUSNO has been tested on many SCT hierarchies, creating taxonomies for 1000 to 100 000 concepts.

METHODS

Subject-based subtaxonomy derivation
The complete taxonomy for the Clinical finding hierarchy (zoomed-out excerpt in online supplementary appendix II) contains 10 614 partial-areas, smaller than the underlying hierarchy (99 440 concepts). However, it is too large to support TQA. It does not allow an auditor to focus on subject concepts, which are ‘hidden’ inside large partial-areas like Finding by site, with 9602 concepts. Thus, we now describe an innovative subject-based TQA approach to select a concept of current interest, for example, for bleeding-related disorders the concept Bleeding is selected.

A subtaxonomy is a taxonomy for a subhierarchy. Given an arbitrary concept $c$, a subtaxonomy is derived using the methodology described in the ‘Background’ section, but it is applied to the SCT subhierarchy rooted at $c$. The root area and unique root partial-area consist of $c$ and all of its descendents with the same relationships. BLUSNO derives $c$’s subtaxonomy.

The definition of the subtaxonomy is applicable to any SCT concept in a hierarchy with attribute relationships. This is important, because our previous approach, which allowed a taxonomy to be derived for an entire hierarchy only, led to many concepts being ‘hidden’ inside large partial-areas, for example, Finding by site contains several of the major causes of death listed in table 4.

By enabling the choice of an arbitrary concept as a root of a subtaxonomy, we enable an SCT editor to view a summary for a subhierarchy of concepts that meet some criteria. If an editor wants to concentrate on a subject area, she can choose a concept that best represents the subject area to be the root of a subtaxonomy, for example, Cancer can be selected as the root of a subtaxonomy, as done in figure 5, making its subhierarchy of concepts more accessible for TQA. Hence, we introduce the subject-based subtaxonomy, a subtaxonomy that provides a compact view of the SCT subhierarchy rooted at a chosen concept that best represents the subject area.

Subtaxonomies are not necessarily disjoint, because concepts may belong to multiple subtaxonomies. Additionally, subtaxonomy partial-areas are not always a subset of those in the complete taxonomy (see Cancer subtaxonomy in the ‘Results’ section).
Within a subtaxonomy, the disjoint partial-area taxonomy derivation methodology is altered to account for concepts in a subtaxonomy overlapping with partial-areas that are outside of the subtaxonomy. For example, the concept *Intra-abdominal hematoma* has two parents in its area in the complete taxonomy: *Hemorrhage of abdominal cavity structure* (in Bleeding’s subtaxonomy) and *Mass of abdominal cavity structure* (in the partial-area *Mass of body structure*, outside the subtaxonomy).

*Intra-abdominal hematoma* inherits the semantics of both partial-area roots and belongs in the disjoint taxonomy. Thus, for the disjoint partial-area subtaxonomy we (1) ensure that all of the concepts in the disjoint partial-area taxonomy are semantically related to the subject $c$ by considering only concepts that are descendants of $c$, and (2) consider overlapping concepts that overlap with partial-areas outside of the subtaxonomy, since such concepts are complex (figure 4).

Based on our experience with the *Specimen* hierarchy, a subject-based subtaxonomy containing 500–1500 concepts is of reasonable size to avoid overwhelming an auditor.

**Subject-based subtaxonomy TQA**

Our previous SCT TQA studies have focused on complex concepts, for example, overlapping concepts, which were shown to have more errors with high statistical significance for the small *Specimen* hierarchy due to the difficulty in modeling complex concepts. Overlapping concepts are more complex than non-overlapping concepts, since they are specializations of all the roots of the partial-areas they are contained in.

In this scalability study, we repeat our analysis of overlapping concepts (hypothesis H1) and test three new refined hypotheses (H2–H4) for a subject-based subtaxonomy of a large hierarchy.

**Hypothesis H1:** Overlapping concepts are more likely to have errors than non-overlapping concepts.

Another group of concepts, which was also shown to have more errors with high statistical significance, are uncommonly classified concepts, for example, those in small partial-areas. A possible reason for their uncommon classification may be a modeling error. Once the error is corrected (eg, by adding a parent or relationship), a concept may join another common classification according to its revised modeling.

However, to account for concepts that overlap between a small partial-area and a large partial-area we introduce H2:

**Hypothesis H2:** Concepts in small disjoint partial-areas are more likely to have errors than concepts in large disjoint partial-areas.

H1 and H2 can be compounded into H3:

**Hypothesis H3:** Concepts in small overlapping disjoint partial-areas are more likely to have errors than concepts in large overlapping disjoint partial-areas.

H3 expresses that concepts that are both complex and uncommonly classified tend to have more errors than concepts that are just complex.

We call the number of partial-areas a concept belongs to the ‘degree of overlap.’

**Hypothesis H4:** Concepts with a higher degree of overlap exhibit a higher error rate.
Concepts that overlap between more partial-areas inherit the semantics of more roots, and thus, are more complex than concepts that overlap between fewer partial-areas.

Even the number of overlapping concepts may be overwhelming when only limited resources may be available to audit them. The above hypotheses can guide a TQA methodology by prioritizing which overlapping concepts should be reviewed first to maximize yield (see ‘Results’ section).

To test the hypotheses, a sample of 300 concepts was reviewed for errors by three of the authors, (YC, JX, and HM), who are trained in medicine and have extensive terminology auditing experience. The review process consists of two phases. First, each auditor is given the complete sample as a list of concepts in alphabetical order and works independently. Each auditor then reports all errors found. As shown in Gu et al.,40 TQA reports from different auditors show substantial differences and a report from one auditor is not reliable. However, a consensus among several auditors’ reports was shown to result in a reliable TQA report. Thus, we are using the second phase for consensus building. Each auditor is given a complete list of errors from all auditors. Each auditor then marks ‘agree’ or ‘disagree’ for each error. A concept is considered erroneous if all auditors agree on the error. A similar consensus TQA protocol was used when auditing overlapping concepts in the Specimen hierarchy.21 If H2–H4 are confirmed, they could guide a TQA methodology that prioritizes the review of concepts summarized by a subtaxonomy according to the error rates for H2–H4.

RESULTS

Bleeding subject-based subtaxonomy

We derived a subject-based subtaxonomy for the concept Bleeding from the January 2013 release of SCT (shown in figure 3). Compared to the Clinical finding partial-area taxonomy with 10,614 partial-areas, this subject-based subtaxonomy is significantly smaller.

The largest area in the Bleeding subtaxonomy, {Associated morphology, Finding site}, has 290 overlapping concepts (55.5%). Figure 4 shows an excerpt of 23 disjoint partial-areas from the disjoint partial-area subtaxonomy for {Associated morphology, Finding site}.

The disjoint partial-area taxonomy for {Associated morphology, Finding site} contains 236 disjoint partial-areas. Most disjoint partial-areas are small: 176 (78.8%) of them are singletons (one concept). The disjoint partial-area taxonomy more accurately summarizes the concepts in this area than the partial-area taxonomy (figure 3).

Figure 3 Top five (out of six) levels of the Bleeding subject-based subtaxonomy. Each level is color coded according to the number of relationships. Levels have been organized into multiple rows due to space limitations. Partial-areas in each area are listed in decreasing order, from left to right, according to their size. Child-of links are not shown for readability. A total of 932 bleeding-related concepts are summarized by 199 partial-areas in 42 areas. Over half (56% = 522/932) of the concepts summarized by this subtaxonomy are in {Associated morphology, Finding site}. The first row of larger partial-areas in this area indicates the major types of bleeding-related findings in SNOMED CT, such as Hemorrhage of abdominal cavity structure (186 concepts), Gastrointestinal hemorrhage (117), and Genitourinary tract hemorrhage (88), demonstrating the summary effect provided by the subject-based subtaxonomy.


Downloaded from http://jamia.oxfordjournals.org/ by guest on January 15, 2016
partial-area Hemorrhage of body cavity structure, but only 10 are descendants of just this root. The other 176 concepts also belong to other partial-areas. The overlapping disjoint partial-areas are made explicit in figure 4.

TQA results
To test H1–H4, three auditors reviewed a sample of 300 concepts from the (Associated morphology, Finding site) area in the Bleeding subtaxonomy for errors: 200 randomly selected overlapping concepts (70%=200/290) and 100 randomly selected non-overlapping concepts (43%=100/232). The latter were taken from partial-areas that had overlapping concept.

The auditors reviewed the January 2013 inferred version of SCT. Together, the auditors first found 131 erroneous concepts. Next all auditors agreed that 87 (66%) of these concepts had at least one same error (table 1). Among the erroneous concepts, 36 were primitives and 51 were fully defined. The auditors all agreed on 123 errors in these 87 concepts (1.41 errors per erroneous concept). For a breakdown by error type, see online supplementary appendix V.

All erroneous concepts and proposed corrections were reported to JTC, head of the US Extension of SCT and a co-author. JTC noted that several significant modeling errors and inconsistency patterns were uncovered by this study (see online supplementary appendix IV).

For H1 we found 39% (=78/200) of overlapping concepts erroneous, versus 9% (=9/100) of non-overlapping concepts. Thus, overlapping concepts are 4.33 times more likely to be erroneous, compared to the previous level. From level 1 to level 7 the error rate is increasing, as expected. We found this hypothesis statistically significant when comparing level 1 to level 2 (p = 0.0322) and level 2 to level 3 (p = 0.0336). Other comparisons were not significant due to the smaller sample sizes of level 4 and above. However, when we compared level 3 to levels 4–8 combined (error rate of 24/39=61.5%), we found significance (p = 0.0116). Table 3 shows five examples of errors and their proposed solutions.

Cancer subject-based subtaxonomy
Table 4 documents the metrics of several other subject-based subtaxonomies. We derived the Cancer (Malignant neoplasm disease) subject-based subtaxonomy for the January 2014 SCT release (figure 5). The majority of the Cancer subtaxonomy’s concepts (3124, 88.5%) are in (Associated morphology, Finding site in a gray box), shown in a gray box, are not part of the Bleeding subject-based subtaxonomy, but many Bleeding concepts overlap with them. Partial-areas outside of the subtaxonomy, such as Mass of body structure, which overlap with partial-areas in the subtaxonomy, for example, Hemorrhage of abdominal cavity structure, are not part of the subtaxonomy and can be hidden, but are important for terminology quality assurance (TQA) to capture the complexity of the overlapping concepts. For example, the disjoint partial-area Pelvic hematoma (3) would not exist if such overlap was not considered.
site) (like Bleeding). While this area has several large partial-areas, for example, Malignant neoplasm of soft tissue (804), it has 2398 overlapping concepts (76.8%). Thus, the disjoint partial-area taxonomy will better summarize its content.

The Cancer subject-based subtaxonomy includes 64 partial-areas (highlighted in yellow in figure 5) that are not in the complete Clinical finding taxonomy. These concepts are typically inside large partial-areas in the complete taxonomy, for example, all of the concepts in the Associated morphology, Finding site yellow partial-areas in figure 5 are inside the large Mass of body structure (7010 concepts) partial-area. This occurs because the relationships are introduced in the subtaxonomy at a lower descendant than in the complete taxonomy. Thus, the Cancer subject-based subtaxonomy summarizes SCT Cancer disorders in a view that is more useful for TQA.

**DISCUSSION**

We demonstrated scalability of taxonomy-based terminology maintenance to large SCT hierarchies using subject-based sub-taxonomies. This represents a significant improvement over our previous approach of reviewing complete taxonomies, which may have thousands of partial-areas (e.g., Clinical finding). Such large taxonomies are hard for humans to visualize, which prevents effective taxonomy-based TQA, based on reviewing groups of concepts that have higher error rates (e.g., small partial-areas16 19 20). There are thousands of such concepts in a large hierarchy, for example, 41 450 (14.3%) concepts in ‘small’ partial-areas and 14 220 (14.5%) overlapping concepts in the Clinical finding hierarchy. Available TQA resources do not typically enable a thorough review of so many concepts.

We addressed these difficulties by combining several novel techniques. The first technique is to concentrate on a subject-based subtaxonomy, which is intuitive for terminology curators because it summarizes all descendants of a chosen broad concept, for example, Bleeding or Cancer. This way, the attention of a curator is focused on a comprehensible subtaxonomy that still summarizes a sizable subject-based portion of the hierarchy. Second, we formulate refined hypotheses regarding concepts with a high likelihood of errors. Third, we prioritize the review of concepts according to the ratios for the hypotheses.

When previously developing taxonomy-based TQA methodologies in small hierarchies, we discovered two kinds of groups

---

**Figure 5** The Cancer subject-based subtaxonomy, following the graphical convention of figure 3. Levels have been organized into multiple rows due to space limitations. Areas of the same levels are color coded according to their number of relationships. Child-of-links are not shown for readability. The Cancer subject-based subtaxonomy summarizes 3531 concepts by 125 partial-areas in 19 areas. The 64 partial-areas that do not appear in the complete Clinical finding taxonomy are highlighted in yellow. The concepts inside of the yellow partial-areas are found in the Mass of body structure (7010 concepts) partial-area in the complete taxonomy.

---

**Table 2** Auditing results broken down by disjoint partial-area taxonomy level

<table>
<thead>
<tr>
<th>Level</th>
<th>Sample concepts (n)</th>
<th>Errorneous concepts (n)</th>
<th>Percentage of errorneous concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>9</td>
<td>9%</td>
</tr>
<tr>
<td>2</td>
<td>90</td>
<td>24</td>
<td>26.7%</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>29</td>
<td>40.8%</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>9</td>
<td>50%</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>7</td>
<td>70%</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>5</td>
<td>83.3%</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>100%</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>2</td>
<td>66.7%</td>
</tr>
<tr>
<td>Total</td>
<td>300</td>
<td>87</td>
<td>32.3%</td>
</tr>
<tr>
<td>Total for 4–8</td>
<td>39</td>
<td>25</td>
<td>64.1%</td>
</tr>
</tbody>
</table>
with a higher likelihood of errors in small hierarchies: concepts in small partial-areas and overlapping concepts. The challenges for scalability included whether this still holds true for concepts in subject-based subtaxonomies and prioritizing among the groups’ concepts.

We formulated and tested new hypotheses (H2–H4), while confirming the previously established hypothesis (H1), for the Bleeding subtaxonomy. When there are many overlapping concepts and an extensive level of overlap, as for the Bleeding and Cancer subtaxonomies, resources for reviewing overlapping concepts need to be prioritized.

According to this study, the TQA methodology steps corresponding to the hypotheses should be applied in decreasing error percentage order (table 5).

Thus, an editor will achieve a higher yield for a given effort. Future studies will investigate error rates in other subject-based subtaxonomies, for example, Cancer with 2398 overlapping concepts, to verify this order.

The study confirms most of our hypotheses and the feasibility of the subject-based subtaxonomy paradigm to support scalability of taxonomic-based maintenance of large SCT hierarchies. More experiments will be performed, using other subtaxonomies, where the sample sizes in the Bleeding subtaxonomy were not sufficient to achieve statistical significance (ie, H3).

**CONCLUSIONS**

We introduced the subject-based subtaxonomy, which summarizes a subhierarchy rooted at a subject concept within a large SCT hierarchy. The subject-based subtaxonomy supports effective terminology maintenance based on four hypotheses for groups of concepts expected to have a higher likelihood of errors. We derived the subject-based subtaxonomy and used it for terminology maintenance for the Bleeding subhierarchy of SCT. By directing TQA efforts towards more complex concepts, a higher error yield is achieved.

**Contributors** CO, JG, and YP were the primary authors of this work. YC, JX, and HM reviewed the sample of 300 concepts for this study and participated in the consensus-based evaluation. JTC reviewed and verified our auditing results and provided the feedback for online supplementary appendix IV. ZW assisted with the statistical analysis.

**Competing interests** None.

**Provenance and peer review** Not commissioned; externally peer reviewed.

### Table 3 Five examples of errors reported by the auditors

<table>
<thead>
<tr>
<th>Concept</th>
<th>Error</th>
<th>Proposed solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bleeding varices of prostate</td>
<td>Missing relationships: Associated morphology and Finding site, with target concepts Varix and Venous structure, respectively</td>
<td>Add the two new relationships in a role group</td>
</tr>
<tr>
<td>Hemorrhage of cervix</td>
<td>Incorrect parent: Hemorrhage of abdominal cavity structure</td>
<td>Remove IS-A to Hemorrhage of abdominal cavity structure (corrected independently in Jan 2014 release)</td>
</tr>
<tr>
<td>Hematoma of pinna</td>
<td>Missing child: Chronic hematoma of pinna (which is incorrectly a synonym of the concept Cauliflower ear).</td>
<td>Add Chronic hematoma of pinna concept and remove the synonym from Cauliflower ear</td>
</tr>
<tr>
<td>Peptic ulcer with hemorrhage AND obstruction</td>
<td>Incorrect relationship target: Associated morphology relationship with a target concept Hemorrhage</td>
<td>Make target concept of Associated morphology relationship Bleeding ulcer to be consistent with Esophageal bleeding</td>
</tr>
<tr>
<td>Bleeding gastric varices</td>
<td>Missing parent: Venous hemorrhage</td>
<td>Add IS-A to Venous hemorrhage</td>
</tr>
</tbody>
</table>

JTC confirmed all of these errors and forwarded the corrections to the International Health Terminology Standards Development Organisation (IHTSDO).

### Table 4 Subject-based subtaxonomy metrics for the 10 leading causes of death in the USA

<table>
<thead>
<tr>
<th>Rank</th>
<th>Cause of death</th>
<th>Subject-based subtaxonomy concept</th>
<th>Concepts (n)</th>
<th>Partial-areas (n)</th>
<th>Areas (n)</th>
<th>Relative size (concepts/partial-areas)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Heart disease</td>
<td>Heart disease</td>
<td>2402</td>
<td>316</td>
<td>61</td>
<td>2.4%/3.0%</td>
</tr>
<tr>
<td>2</td>
<td>Cancer</td>
<td>Malignant neoplastic disease</td>
<td>3531</td>
<td>125</td>
<td>19</td>
<td>3.6%/1.2%</td>
</tr>
<tr>
<td>3</td>
<td>Chronic lower respiratory diseases</td>
<td>Disorder of lower respiratory system</td>
<td>1414</td>
<td>354</td>
<td>51</td>
<td>1.4%/3.4%</td>
</tr>
<tr>
<td>4</td>
<td>Stroke</td>
<td>Cerebrovascular disease</td>
<td>262</td>
<td>75</td>
<td>15</td>
<td>0.3%/0.7%</td>
</tr>
<tr>
<td>5</td>
<td>Accidents</td>
<td>Injury due to exposure to external cause</td>
<td>267</td>
<td>65</td>
<td>11</td>
<td>0.3%/0.6%</td>
</tr>
<tr>
<td>6</td>
<td>Alzheimer’s disease</td>
<td>Disorder of brain</td>
<td>2300</td>
<td>396</td>
<td>67</td>
<td>2.3%/3.8%</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes</td>
<td>Diabetes mellitus</td>
<td>112</td>
<td>30</td>
<td>14</td>
<td>0.1%/0.2%</td>
</tr>
<tr>
<td>8</td>
<td>Nephritis, nephrotic syndrome, and nephrosis</td>
<td>Kidney disease</td>
<td>909</td>
<td>243</td>
<td>47</td>
<td>0.9%/2.3%</td>
</tr>
<tr>
<td>9</td>
<td>Influenza and pneumonia</td>
<td>Pneumonitis</td>
<td>334</td>
<td>73</td>
<td>24</td>
<td>0.3%/0.7%</td>
</tr>
<tr>
<td>10</td>
<td>Suicide</td>
<td>Suicide</td>
<td>16</td>
<td>9</td>
<td>2</td>
<td>0.4%/29%</td>
</tr>
</tbody>
</table>

### Table 5 Order of TQA methodology steps

<table>
<thead>
<tr>
<th>Order</th>
<th>Hypothesis</th>
<th>Overlap</th>
<th>Errors expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H4</td>
<td>Overlap levels 4–8</td>
<td>64.1%</td>
</tr>
<tr>
<td>2</td>
<td>H4</td>
<td>Overlap level 3</td>
<td>40.8%</td>
</tr>
<tr>
<td>3</td>
<td>H3</td>
<td>Small overlapping disjoint partial-areas</td>
<td>40.2%</td>
</tr>
<tr>
<td>4</td>
<td>H4</td>
<td>Overlap level 2</td>
<td>26.7%</td>
</tr>
<tr>
<td>5</td>
<td>H2</td>
<td>Small non-overlapping disjoint partial-areas</td>
<td>20.6%</td>
</tr>
</tbody>
</table>

H, hypothesis; TQA, terminology quality assurance.
REFERENCES


