A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships

RECEIVED 18 July 2014
REVISED 9 September 2014
ACCEPTED 20 September 2014
PUBLISHED ONLINE FIRST 20 October 2014

Christopher Ochs¹, James Geller¹, Yehoshua Perl¹, Yan Chen², Ankur Agrawal³, James T Case⁴, George Hripcsak⁵

ABSTRACT

Objective Large and complex terminologies, such as Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), are prone to errors and inconsistencies. Abstraction networks are compact summarizations of the content and structure of a terminology. Abstraction networks have been shown to support terminology quality assurance. In this paper, we introduce an abstraction network derivation methodology which can be applied to SNOMED CT target hierarchies whose classes are defined using only hierarchical relationships (ie, without attribute relationships) and similar description-logic-based terminologies.

Methods We introduce the tribal abstraction network (TAN), based on the notion of a tribe—a subhierarchy rooted at a child of a hierarchy root, assuming only the existence of concepts with multiple parents. The TAN summarizes a hierarchy that does not have attribute relationships using sets of concepts, called tribal units that belong to exactly the same multiple tribes. Tribal units are further divided into refined tribal units which contain closely related concepts. A quality assurance methodology that utilizes TAN summarizations is introduced.

Results A TAN is derived for the Observable entity hierarchy of SNOMED CT, summarizing its content. A TAN-based quality assurance review of the concepts of the hierarchy is performed, and erroneous concepts are shown to appear more frequently in large refined tribal units than in small refined tribal units. Furthermore, more erroneous concepts appear in large refined tribal units of more tribes than of fewer tribes.

Conclusions In this paper we introduce the TAN for summarizing SNOMED CT target hierarchies. A TAN was derived for the Observable entity hierarchy of SNOMED CT. A quality assurance methodology utilizing the TAN was introduced and demonstrated.

Key words: SNOMED CT, terminology quality assurance, abstraction network, terminology without lateral relationships, hierarchical abstraction network, terminology summarization

INTRODUCTION

The Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT, SNOMED for short) is a large medical terminology.¹ Modeling errors and inconsistencies in a terminology of SNOMED's size and complexity are unavoidable. Quality assurance (QA) is an important part of the lifecycle of a terminology.² However, identifying errors in large terminologies is a resource-intensive and error-prone task.^{3,4} In previous research, we developed the paradigm of 'abstraction networks' (AbNs) to support the QA of terminologies. An AbN is a high-level, compact network that summarizes the content and structure of a large, complex terminology. AbNs have been shown to support the identification of concepts with a higher likelihood of errors compared with a control sample.⁵

The AbN paradigm has been successfully applied as the refined semantic network⁶ for the Unified Medical Language

System (UMLS)⁷ and as the Schema⁸ for the Medical Entities Dictionary (MED).⁹ The area and partial-area taxonomy AbNs were developed² for the National Cancer Institute thesaurus (NCIt)^{10,11} and for SNOMED hierarchies^{12,13} with attribute relationships (relationships for short). Furthermore, several types of AbNs^{14–17} were developed for OWL-based ontologies including the Ontology of Clinical Research,¹⁸ Sleep Domain Ontology,¹⁹ Ontology for Drug Discovery Investigations²⁰ and Cancer Chemoprevention Ontology.²¹

In this paper, we introduce the tribal abstraction network (TAN), a new type of AbN designed for SNOMED hierarchies that do not have attribute relationships, assuming only the existence of multiple parents. The TAN summarizes the content and structure of such SNOMED hierarchies and supports their QA, by identifying concepts with a higher likelihood of incorrect or missing IS-A relationships.

Correspondence to Christopher Ochs, Computer Science Department, New Jersey Institute of Technology, Newark, NJ 07102-1982 USA; E-mail: cro3@njit.edu

©The Author 2014. Published by Oxford University Press on behalf of the American Medical Informatics Association.

All rights reserved. For Permissions, please email: journals.permissions@oup.com

For numbered affiliations see end of article.

BACKGROUND

SNOMED is a large terminology curated by the International Health Terminology Standards Development Organization (IHTSDO).²² Its January 2013 release contained 297 801 active concepts in 19 hierarchies. SNOMED is organized as a Directed Acyclic Graph with 542 485 IS-A relationships. In addition, concepts are connected by 912 196 relationships. For example, the concept *Heart sounds abnormal* (in *Clinical finding*) has a relationship *Interprets* with a target concept *Heart sounds* (in *Observable entity*) (concept names and hierarchy names appear in italics).

A large terminology diagram, where nodes represent concepts and edges represent relationships, would be overwhelming. In addition, viewing a terminology through a browser, such as CliniClue,²³ hides the overall context of the concept. Often, only parents and children will be displayed alongside a selected concept. AbNs summarize an entire SNOMED hierarchy and have been shown to support QA for various terminologies.^{2,8,14–17,24–26}

In previous work, we derived the area and partial-area taxonomies for SNOMED by using relationships.^{12,27} These AbNs were shown to support auditing of SNOMED.^{5,28,29} Wei and Bodenreider³⁰ showed that taxonomies support finding errors that cannot be discovered by classifiers such as Hermit.³¹ Various semantic, structural and ontological QA techniques for SNOMED are offered by Rector,^{32,33} Schulz^{34–36} and Zhu *et al.*³⁷

The (partial-)area taxonomies require relationships. Within SNOMED, 12 hierarchies have no relationships and serve only as targets for relationships ('target hierarchies' for short). Thus, an alternative paradigm is suggested to design an AbN for target hierarchies. In SNOMED, 102 826 concepts (34.5%) have multiple parents, and the average number of parents is 1.822. Online supplementary appendix 1 shows the number of concepts in each hierarchy having multiple parents and their percentage of each hierarchy. The number of concepts with multiple parents varies widely, with almost half (45.39%) of the concepts in *Clinical finding* compared with only 5.33% in *Observable entity*. We present a new AbN for SNOMED target hierarchies with multiple parents, supporting auditing. Future applications are discussed.

METHODS

The TAN is derived as follows. The children of a hierarchy's root are named 'patriarchs'. A 'tribe' consists of a patriarch and all its descendants. The use of tribe and patriarch follows the family tree paradigm. A tribe is named after its patriarch, since all its concepts are specializations of the patriarch. Every concept in a hierarchy, except for the hierarchy root, belongs to at least one tribe. In a TAN, all concepts belonging to a common set of tribes are grouped together. A necessary but not sufficient condition for a hierarchy to have concepts in multiple tribes is that there are concepts with multiple parents. We illustrate these definitions using the *Observable entity* hierarchy.³⁸ In the January 2013 release, it contains 8274 concepts and 8726 IS-A relationships.

Figure 1 shows 20 concepts, represented as nodes labeled with their respective names. Each of the children of *Observable entity*—for example, *Process*, *Function* and *Clinical history/examination observable* (shortened to *Clinical history/exam*)—is a patriarch of a tribe, abbreviated to P, F and C, respectively. IS-A links are represented as arrows. For example, *Digestive system function* IS-A *Function*. *Physiological action*, *Activity*, *Ingestion*, *Drinking*, *Feeding* and *Breastfeeding (mother)* belong to the *Process* tribe since they are descendants of *Process*.

Each concept is labeled by its set of tribes, called 'tribal set'. From the view point of the tribes, each concept is assigned to all tribes that it belongs to. To assign all concepts in a hierarchy to tribes, the hierarchy is traversed using 'topological sort'³⁹ starting from the patriarchs. In a topological sort procedure, any non-patriarch concept is processed only after all of its parents have been processed. If a concept *c* has one parent *p_a* belonging to the tribe A and another parent *p_b* belonging to B, *c* belongs to both tribes A and B, because it is a descendant of both patriarchs A and B. Once all parents of a concept *c* have been processed, *c* is assigned the union of its parents' tribal sets.

This procedure is generally more efficient than performing a separate graph traversal from each hierarchy's patriarch, since each concept is only processed once. In a standard graph traversal, such as breadth first search,³⁹ concepts would be processed multiple times, according to the number of tribes to which they belong.

Figure 1 shows the results of applying the tribal assignment process to an excerpt of 20 concepts. Tribal sets are shown in braces below each concept's name. Figure 2 groups together the concepts with common tribal sets. Each group is represented by a dashed bubble and is labeled with the name(s) of the tribes.

Concepts that are descendants of only one patriarch belong only to its tribe. In figure 2 *Large bowel function* belongs only to the *Function* tribe. However, *Ingestion*, *Breastfeeding (mother)*, *Activity of daily living* and *Defecation* all belong to more than one tribe, because each has multiple parents in different tribes. For example, *Ingestion* has two parents, *Physiological action* and *Digestive system function*, which belong to the *Process* and *Function* tribes, respectively. *Ingestion* therefore belongs to both the *Process* and *Function* tribes. *Defecation* belongs to all three tribes. Even though *Drinking*, *Feeding*, *Basic activity of daily living* and *Toileting* each have only one parent, they belong to multiple tribes because each has an ancestor that belongs to multiple tribes.

Generally, concepts that belong to more than one tribe are more complex than those belonging to only one tribe, since they are specializations of several patriarchs. A concept that belongs to multiple tribes is called a 'joint' concept. Joint-ness can be used to group concepts into sets. These sets can be used to derive two kinds of TANs: the band TAN and the more refined cluster TAN.

Band TAN

A 'tribal band', or 'band' for short, is a tribal unit consisting of the set of all concepts that are members of the exact same

Figure 1: An excerpt of 20 concepts from the *Observable entity* hierarchy with abbreviated tribal names in braces.

tribes. A band is named after the set of tribes that each concept within the band belongs to. A band may have multiple roots, where a root of a band is a concept that has no parents within the band, although it may have parents in other bands. In figure 2, root concepts are identified by a red outline. Each set of concepts, surrounded by a dashed bubble, defines a band.

A band TAN consists of one node for each band. These nodes are linked by hierarchical 'child-of' relationships derived from the underlying IS-A hierarchy. A band A is a child-of another band B if and only if every root concept in A has an IS-A link to a concept in B. A band may be child-of multiple bands. The band TAN provides a compact view of a target hierarchy.

Figure 3 shows the band TAN for figure 1 obtained using the tribal sets from figure 2. The number of concepts is listed under each band's name. The four concepts *Ingestion*, *Feeding*, *Drinking* and *Breastfeeding (mother)* belong to the band {*Process*, *Function*}f. *Ingestion* and *Breastfeeding (mother)* are the roots of the {*Process*, *Function*} band, because neither has parents in the {*Process*, *Function*} band. The band {*Process*, *Function*} is a child-of two bands, {*Process*} and {*Function*}, because both roots *Ingestion* and *Breastfeeding (mother)* have parents in both these bands.

The band {*Process*, *Function*, *Clinical history/exam*} is a child-of both {*Process*, *Clinical history/exam*} and {*Function*} because its root *Defecation* has two parents, *Toileting* in {*Process*, *Clinical history/exam*} and *Large bowel function* in {*Function*}.

Each band has a degree of 'joint-ness' according to the number of tribes that its members belong to. Bands containing concepts of only one tribe consist of the patriarch and all of its descendants that are not descendants of a second patriarch.

In visualizations of TANs (figures 3 and 5), tribal bands are organized into levels according to their degrees of joint-ness and are color-coded. Bands of degree 1 are located at the top. Bands of degree 2, with concepts that belong to two tribes, are below.

Cluster TAN

A tribal band may have multiple roots. Each root defines a sub-hierarchy of concepts within the band. A 'tribal cluster', or 'cluster' for short, is a refined tribal unit consisting of a root of a band and all its descendants within the same band. A tribal cluster is named after its root, because all other concepts in the cluster are specializations of the root.

Figure 2: The concepts from figure 1 grouped by common tribal sets. Root concepts are outlined in red.

Figure 3: The band tribal abstraction network derived from figure 2. Each box represents a band. Child-of links are represented using arrows between bands.

Clusters further refine the band TAN into the ‘cluster TAN’. In a cluster TAN, all the clusters of a band are drawn within that band’s node. Clusters, like bands, are linked by child-of relationships based on the underlying IS-A hierarchy. A cluster A is a child-of another cluster B if the root concept of A has an IS-A link to a concept in B. A cluster may be a child-of multiple clusters.

In figure 2, *Ingestion* and *Breastfeeding (mother)* are the two roots of the {*Process*, *Function*} band. In visualizations of a

cluster TAN (figures 4 and 6), clusters are represented as white boxes within a band box, labeled by their roots, with their numbers of concepts below the root names. The root, *Ingestion*, and its two descendants are represented as a cluster named *Ingestion* of three concepts in {*Process*, *Function*} (figure 4). The *Ingestion* cluster is a child-of the *Process* and *Function* clusters, because the root concept *Ingestion* has parents in these clusters.

TANs for QA

QA of large terminologies is difficult and time consuming. By focusing QA efforts on a subset of concepts that are likely to be more error prone, QA resources can be utilized more effectively. In previous research,^{2,12,14–17} we have shown that AbNs support QA by identifying such concepts. TANs can also be used for SNOMED QA, by identifying concepts likely to have more hierarchical errors. Such errors were deemed the most problematic in a study of SNOMED users.⁴⁰ IS-A relationships play an important role in SNOMED. For target hierarchies, the correctness of the IS-A hierarchy is important, because the concepts of these hierarchies serve as targets for relationships with source concepts in other hierarchies. There are 18 839 defining relationships in other SNOMED domains where an *Observable entity* concept is used as a target concept. Proper placement of target concepts in a hierarchy is crucial, since the target of a relationship should be as specific as possible.

Hypothesis 1: In a cluster TAN, concepts in large clusters are more likely to have errors than concepts in small clusters.

Figure 4: The cluster tribal abstraction network derived from figure 2. Child-of links are represented by arrows between clusters.

The rationale is as follows. For a concept in a target hierarchy, errors can occur only in the hierarchy. An IS-A relationship may be either wrong or missing and the concept is misplaced in the hierarchy. There is a greater chance of such situations in large clusters, because, as the number of hierarchically closely related concepts increases, the chance of a concept being misplaced also increases. In clusters with fewer concepts, there is less chance of a concept being misplaced. We tested this hypothesis in the *Observable entity* hierarchy.

To reiterate, our goal is to minimize the number of concepts that are reviewed, by selecting concepts with a high likelihood of errors. Such a review should yield a large number of erroneous concepts, relative to the effort spent. However, auditing all large clusters is not practical, because of their large numbers of concepts. Therefore, we introduce hypothesis 2. (Reminder: level numbers grow higher when moving downward in a band diagram.)

Hypothesis 2: Among the large clusters, those concepts belonging to higher-numbered levels are more likely to have errors.

The rationale is that concepts belonging to more tribes tend to be more complex because of their specialization of more patriarchs. The modeling of more complex concepts is more prone to errors. Assuming there is support for these two hypotheses, the following auditing methodology emerges. Start reviewing the larger clusters of the highest-numbered level. As long as QA resources remain, continue to review larger clusters moving up in the diagram.

RESULTS

A cluster TAN was derived for the July 2011 version of the *Observable entity* hierarchy. Even though *Observable entity* has relatively few concepts with multiple parents (439, see [online supplementary appendix I](#)), a cluster TAN well summarizes this hierarchy ([table 1](#)). *Observable entity* has 27 children and therefore 27 tribes, with 16 of these tribes having joint concepts. The maximum number of tribes that a concept belongs to is three, while 6627 (80.5%) of the concepts of a unique tribe belong to the 27 tribal bands on the first level. The second level comprises 1236 concepts (15%) and the third level 368 (4.47%). The percentage of concepts with multiple parents is much higher in levels 2 and 3 (14% and 20%) than in level 1 (2.5%). [Figures 5 and 6](#) show the band TAN and cluster TAN, respectively.

The bands of level 1 indicate the major types of concepts in a hierarchy; level 1 of [figure 5](#) contains many *Clinical history/examination* and *Function* concepts. Levels 2 and 3 show how the bands of level 1 intersect in the hierarchy—for example, the *Clinical history/examination observable* band intersects with most other bands. [Figure 6](#) allows common concept groups of multiple tribes to be identified. For example, looking at the large clusters, such as *Female genital feature* (152), *Cardiac feature* (145), *Eye observable* (143), followed by *Blood pressure* (86), *Activity of daily living* (79), *Joint movement* (86) and *Feature of upper limb* (84), provides a summarization of the major types of concepts in the *Observable entity* hierarchy. For a finer summary, one should view the ‘medium’ sized clusters of 25–50 concepts—for example, *Tumor size* (35). The 15 clusters with at least 25 concepts summarize 1084 concepts (68.3%) of the major subjects in levels 2 and 3.

[illegible]

Figure 6: The cluster tribal abstraction network for *Observable entity*. Child-of edges are hidden for readability. Each level is organized into multiple rows because of width limitations. Level 1 (not shown) is the same as in figure 5. The tribal clusters from figure 4 are outlined in red.

To test our hypotheses, one of the authors (YC) reviewed 1160 concepts (14.1%) from *Observable entity*, 420, 474 and 266, from levels 1–3, respectively. YC is trained in medicine and experienced in terminology auditing. In particular, the reviewer has performed extensive QA reviews of SNOMED for several previous studies.^{5,12,28,29,41–44} At each level, we audited all concepts from clusters of nine concepts or fewer (284 in total) and randomly selected concepts from larger clusters (876 total). Because of the large difference in the number of concepts between the three levels of the *Observable entity* TAN, the sizes of the samples from each level were not selected evenly but with a higher percentage for the smaller number of higher indexed levels, to enable statistical significance for this study.

The reviewer was blind to the TAN methodology and was only provided with a list of concepts sorted alphabetically according to their fully specified names. The reviewer was not aware of what level or which cluster a given sample concept is from.

We found 39 erroneous concepts (3.36%) in our sample (listed in [online supplementary appendix II](#)). Twenty-one and 18 concepts had incorrect and missing IS-A relationships, respectively. These errors were submitted to JTC (co-author) for review and inclusion in the US extension of SNOMED.⁴⁵ Only three corrections were not accepted by JTC. IHTSDO

accepted all of these corrections (included in July 2014 release).

The 39 erroneous concepts exhibited 42 errors. These erroneous concepts served as targets for 42 different relationships from source hierarchies. We performed a follow-up review of these concepts, using the January 2013 release of SNOMED and all errors were still present.

To test hypothesis 1, we studied the relationship between cluster size and error rate as follows. To handle correlation of concepts within clusters, we analyzed the data at the cluster level⁴⁶ by calculating the error rate per cluster (ie, for each cluster, the total number of erroneous concepts divided by the total number of sample concepts in the cluster). To better visualize the effect of cluster size, and because the relation between cluster size and error rate might not be linear, we stratified clusters into six bins. The per-cluster analysis is shown in [online supplementary appendix III](#). [Table 2](#) shows the distribution of clusters, concepts, sample concepts and erroneous concepts among the six bins. The mean cluster error rate column shows the average error rate of clusters in each bin.

We calculated the pairwise statistical differences of mean cluster error rates among the bins. The error rates and 95% CIs versus cluster size were calculated between all bins. Bin 1 (clusters with more than 150 concepts) had an error rate

Table 1: Level summary of the *Observable entity* hierarchy's band and cluster TANs

Level	Bands (n)	Clusters (n)	Concepts (n)	Concepts with multiple parents (n (%))	Average parents (n)
1	27	27	6643	169 (2.5)	1.03
2	23	101	1220	170 (13.9)	1.14
3	13	52	368	73 (19.8)	1.21
Total	63	180	8231	412 (5.3)	1.06

TAN, tribal abstraction network.

Table 2: Distribution of concepts, errors and error rates among the six bins

Bin	Cluster size	Clusters (n)	Concepts (n)	Concepts/ clusters	Sample concepts (n)	Erroneous concepts (n (%))	Mean cluster error rate (%)
1	>150	5	6,198	1239.6	219	10 (4.56)	5.1
2	86–150	6	665	110.83	221	16 (7.24)	4.3
3	46–85	7	482	68.86	186	3 (1.08)	1
4	11–45	27	572	21.19	231	5 (2.16)	1
5	2–10	46	225	5	214	3 (1.40)	1.8
6	1	89	89	1	89	2 (2.25)	2.3
Total		180	8,231	45.98	1160	39 (3.36)	2.0

significantly higher than bin 3 (46–85 concepts) and bin 4 (clusters with 11–45 concepts), with $p = 0.019$ and $p = 0.009$, respectively. Furthermore, bin 2 (85–150 concepts) had an error rate significantly higher than bin 4 ($p = 0.039$). Error rates between other pairs of bins were not significantly different. However, in general, bin 1 and 2 clusters have higher mean error rates than clusters in bins 3–4.

To test hypothesis 2, we analyzed the mean error rates among the ‘large’ clusters in the three levels. Various boundaries between small and large were tested. No boundary resulted in significance because of the relatively small number of ‘large’ clusters at level 3 (see online [supplementary appendix III](#)). However, we observed that, at higher levels, larger clusters tended to have higher error rates. Using the result from hypothesis 1, we define a large cluster as a cluster in bin 1 or 2, and a small cluster as a cluster in bins 3–6.

Table 3 provides a breakdown of auditing results by level and by large versus small clusters. We observe that higher

level large clusters have a higher error rate. The mean error rates are 12.79%, 4.95% and 3.08% for levels 3, 2 and 1, respectively. For large clusters, the error rate among concepts (E) also increases with their level (ie, 3.28%, 5.26% and 12.8%). Table 4 provides examples of errors identified.

In [online supplementary appendix IV](#) we report on a comparative QA study involving 900 *Observable entity* concepts which reports error ratios, precision, specificity and sensitivity for three additional QA techniques which are applicable for a hierarchy without attribute relationships (eg, Agrawal *et al*⁴¹ for the SNOMED Problem List⁴⁷).

DISCUSSION

The TAN addresses the need for summary methodologies for the eight target hierarchies of SNOMED with multiple parents. We demonstrated how a TAN can be used for QA. We note that the number of concepts with multiple parents in a hierarchy is not as important for deriving a TAN as where such concepts

Table 3: An analysis of the three levels broken down by bins 1 and 2 (‘large clusters’) and bins 3–6 (‘small clusters’)

Level	Large clusters (bins 1 and 2)					Small clusters (bins 3–6)				
	NCI	C	S	E (%)	Mean cluster error rate (%)	NCI	C	S	E (%)	Mean cluster error rate (%)
1	6	6251	183	6 (3.28)	3.08	21	392	237	7 (2.95)	1.11
2	4	526	171	9 (5.26)	4.95	97	694	303	4 (1.32)	1.88
3	1	86	86	11 (12.8)	12.79	51	282	180	2 (1.11)	2.12
Total	11	6863	440	26 (5.9)	4.64	169	1368	720	13 (1.81)	1.86

C, number of concepts; E (%), number of erroneous concepts and percentage that are erroneous ($\% = E/S$); NCI, number of clusters; S, number of sample concepts.

Table 4: A sample of five errors taken from our auditing results with errors separated into errors of omission and errors of commission

Concept(s)	Current parent(s)	Error	Suggested solution
Errors of omission			
<i>Autonomic bladder function</i>	<i>Autonomic nervous system function</i>	Missing parent: <i>Bladder function</i>	Add IS-A to <i>Bladder function</i>
<i>Date chemotherapy completed</i>	<i>Date chemotherapy completed</i>	Missing parent: <i>Temporal observable</i>	Add IS-A to <i>Temporal observable</i>
<i>Sitting systolic blood pressure and Sitting diastolic blood pressure</i>	<i>Systolic blood pressure and Diastolic blood pressure, respectively</i>	Missing parent: <i>Sitting blood pressure</i>	Add IS-A relationships from <i>Sitting systolic blood pressure</i> and <i>Sitting diastolic blood pressure</i> to <i>Sitting blood pressure</i>
Errors of commission			
<i>Ankle joint temperature</i>	<i>Body temperature and Feature of ankle joint</i>	Incorrect parent: <i>Body temperature</i>	Replace IS-A to <i>Body temperature</i> with IS-A to <i>Joint temperature</i> , which is a grandchild of <i>Body temperature</i>
<i>Dorsalis pedis arterial pressure</i>	<i>Blood pressure</i>	Incorrect parent: <i>Blood pressure</i>	Replace IS-A to <i>Blood pressure</i> with IS-A to <i>Arterial blood pressure</i> , which IS-A <i>Blood pressure</i>

appear. Only 412 (5.33%) of the concepts in *Observable entity* have multiple parents, a relatively small number compared with other hierarchies (see [online supplementary appendix I](#)), but a TAN is successfully derived, since 153 such concepts are located ‘at the crossroads’ of tribe combinations.

The goal of using a TAN is to limit the resources and increase the yield of QA. We found that concepts in the *Observable entity* hierarchy are more likely to be erroneous if they belong to larger clusters (bins 1 and 2) in the TAN rather than to smaller clusters (bins 3–6). Furthermore, the percentage of errors is highest in our sample for larger clusters at level 3 and slightly higher in larger clusters in level 2 than level 1.

Following our methodology, the 86 and 526 concepts in large clusters of levels 3 and 2, respectively, should be reviewed. Among the 86 concepts in the larger level 3 cluster, 11 erroneous concepts were found. The number of erroneous concepts expected in reviewing the 526 concepts in larger level 2 clusters is 28 ($=0.0526 \times 526$) (based on E in [table 3](#)). Hence, a total of 39 ($=11+28$) errors are expected from reviewing 612 ($=86+526$) concepts in the large clusters of levels 2 and 3, according to the methodology. Coincidentally, we found 39 erroneous concepts when reviewing our sample of 1160 concepts. Hence, our methodology would probably yield the same number of erroneous concepts while saving the review of 548 ($=1160-612$) concepts. Research on other target hierarchies is required in view of the small number of concepts in level 3 and the relatively small number of large clusters on all levels.

Future work

One issue arising from the placement of concepts with multiple parents in a hierarchy is the emergence of ‘super-large’ level 1 clusters, such as *Clinical history/examination observable* (4096) (CHEO for short) and *Function* (1384), together containing 67% of the *Observable entity* hierarchy. These clusters require further summarization. We can recursively derive a TAN for each such cluster, with its patriarch treated as a hierarchy root, thus creating a TAN to summarize its contents. In the future, we will derive TANs for these two clusters and perform QA for them.

Recursive application of the TAN summarization on the super-large clusters of level 1 has the potential to overcome a problem regarding the QA of the large level 1 clusters, consisting of 80% of the *Observable entity* hierarchy. Assuming that hypothesis 2 holds true for recursively derived TANs, then, in the CHEO subhierarchy’s TAN, more erroneous concepts will concentrate in the large clusters of higher levels than in the large clusters of level 1 of the TAN for the CHEO subhierarchy. Repeating the recursive process as necessary will hopefully enable the desired review of large tribal clusters of levels higher than 1 with higher concentration of erroneous concepts, for the proper subhierarchies. Of course, this future research plan needs testing, which is beyond the current research. We further describe this approach in [online supplementary appendix V](#).

Similarly to deriving a TAN for a super-large cluster, we can derive a TAN for a super-large root partial-area of a

partial-area taxonomy.^{2,12} For example, the partial-area *Procedure* contains 2518 concepts without lateral relationships.²⁹ A TAN for such a root area will provide a summary of its content. We will investigate QA methods for its concepts.

One could derive a TAN for each partial-area of a taxonomy.¹² What is common to all concepts of a partial-area is that they share the same root and set of relationships. Hence, for such groups, with multiple parents, it is not possible to use relationships to obtain a further division. However, one can ignore the relationships and derive a TAN for the partial-area, summarizing its concepts. Examples of super-large partial-areas in *Procedure* include *Procedure by method* (3684) and *Measurement of substance* (3980). In future work, we will investigate the use of TANs to complement partial-area taxonomy-based QA of large source hierarchies.²⁹ To support this research, we will create a tool for automatically deriving and visualizing TANs, similar to the BLUSNO tool for SNOMED partial-area taxonomies.¹⁴

In the future, we will analyze the phenomenon of concepts that overlap between clusters. While bands are strictly disjoint, a concept may belong to multiple clusters. We hypothesize that concepts in multiple clusters are more likely to be erroneous because of being specifications of the roots of multiple clusters. While the *Observable entity* hierarchy has no such concepts, there are over 18 000 concepts that overlap between multiple clusters located throughout SNOMED’s other hierarchies.

CONCLUSIONS

The TAN summarizes the content of hierarchies without relationships in SNOMED. We derived a TAN for the *Observable entity* hierarchy. We found that concepts in large clusters have a statistically significantly higher likelihood of being erroneous than concepts in small clusters. Furthermore, for large clusters, concepts of more tribes are more likely to be erroneous than concepts belonging to fewer tribes.

CONTRIBUTORS

CO, JG and YP were the primary authors of this work. YC reviewed the sample of 1160 concepts used in the TAN study and the comparative study described in [online supplementary appendix V](#). AA managed the comparative study described in [online supplementary appendix V](#). JTC verified our auditing results. GH performed the statistical analysis.

COMPETING INTERESTS

None.

PROVENANCE AND PEER REVIEW

Not commissioned; externally peer reviewed.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at <http://jamia.oxfordjournals.org/>.

REFERENCES

1. SNOMED CT. <http://www.ihtsdo.org/snomed-ct/>

2. Min H, Perl Y, Chen Y, et al. Auditing as part of the terminology design life cycle. *J Am Med Inform Assoc*. 2006;13: 676–90.
3. Gu H, Elhanan G, Perl Y, et al. A study of terminology auditors' performance for UMLS semantic type assignments. *J Biomed Inform*. 2012;45:1042–8.
4. Gu HH, Hripcsak G, Chen Y, et al. Evaluation of a UMLS Auditing Process of Semantic Type Assignments. *AMIA Annual Symposium Proceedings* 2007:294–8.
5. Halper M, Wang Y, Min H, et al. Analysis of error concentrations in SNOMED. *AMIA Annual Symposium Proceedings* 2007:314–18.
6. Gu H, Perl Y, Geller J, et al. Representing the UMLS as an object-oriented database: modeling issues and advantages. *J Am Med Inform Assoc* 2000. 2000;7:66–80.
7. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32 issue):D267–70.
8. Gu H, Halper M, Geller J, et al. Benefits of an object-oriented database representation for controlled medical terminologies. *J Am Med Inform Assoc*. 1999;6:283–303.
9. Cimino JJ, Clayton PD, Hripcsak G, et al. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J Am Med Inform Assoc*. 1994;1:35–50.
10. Sioutos N, de Coronado S, Haber MW, et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform*. 2007;40:30–43.
11. Fragoso G, de Coronado S, Haber M, et al. Overview and utilization of the NCI thesaurus. *Comp Funct Genomics*. 2004; 5:648–54.
12. Wang Y, Halper M, Min H, et al. Structural methodologies for auditing SNOMED. *J Biomed Inform*. 2007;40:561–81.
13. Wang AY, Sable JH, Spackman KA. The SNOMED clinical terms development process: refinement and analysis of content. *Proc AMIA Symp*. 2002:845–9.
14. Geller J, Ochs C, Perl Y, et al. New Abstraction Networks and a New Visualization Tool in Support of Auditing the SNOMED CT Content. *AMIA Annu Symp Proc*. 2012;2012: 237–46.
15. Ochs C, He Z, Perl Y, et al. Choosing the Granularity of Abstraction Networks for Orientation and Quality Assurance of the Sleep Domain Ontology. *Proceedings of the 4th International Conference on Biomedical Ontology*. 2013; 2013:84–9.
16. He Z, Ochs C, Soldatova L, et al. Auditing Redundant Import in Reuse of a Top Level Ontology for the Drug Discovery Investigations Ontology 2013 Workshop on Vaccine and Drug Ontology Studies. 2013.
17. He Z, Ochs C, Agrawal A, et al. A Family-Based Framework for Supporting Quality Assurance of Biomedical Ontologies in BioPortal. *Proc AMIA Annu Symp*. 2013;2013:581–90.
18. Tu S, Carini S, Rector A, et al. OCRE: An Ontology of Clinical Research. 11th International Protege Conference; 2009.
19. Arabandi S, Ogbuji C, Redline S, et al. Developing a Sleep Domain Ontology. *AMIA Clinical Research Informatics Summit*. 2010.
20. Qi D, King RD, Hopkins AL, et al. An Ontology for Description of Drug Discovery Investigations. *J Integr Bioinform*. 2010;7:126–39.
21. Zeginis D, Hasnain A, Loutas N, et al. A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. *Semantic Web*. 2013:1–16.
22. IHTSDO. International Health Terminology Standards Development Organization (IHTSDO). 2012 [cited 2013 9 September 2013]; <http://www.ihtsdo.org/>
23. CliniClue Xplore. <http://www.cliniclue.com/software>
24. Gu H, Perl Y, Elhanan G, et al. Auditing concept categorizations in the UMLS. *Artif Intell Med*. 2004;31:29–44.
25. Chen Y, Gu H, Perl Y, et al. Structural group auditing of a UMLS semantic type's extent. *J Biomed Inform*. 2009;42:41–52.
26. Chen Y, Gu H, Perl Y, et al. Expanding the extent of a UMLS semantic type via group neighborhood auditing. *J Am Med Inform Assoc*. 2009;16:746–57.
27. Wang Y, Halper M, Wei D, et al. Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED. *J Biomed Inform*. 2012;45:15–29.
28. Wang Y, Halper M, Wei D, et al. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. *J Biomed Inform*. 2012;45:1–14.
29. Ochs C, Perl Y, Geller J, et al. Scalability of abstraction-network-based quality assurance to large SNOMED hierarchies. *AMIA Annu Symp Proc*. 2013;2013:1071–80.
30. Wei D, Bodenreider O. Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies—a case study in SNOMED CT. *Stud Health Technol Inform*. 2010;160(Pt 2):1070–4.
31. Shearer R, Motik B, Horrocks I. HermiT: a highly-efficient OWL reasoner. *Proceedings of the 5th International Workshop on OWL: Experiences and Directions*. 2008.
32. Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. *J Am Med Inform Assoc*. 2011;18:432–40.
33. Rector AL, Iannone L. Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. *J Biomed Inform*. 2011:199–209.
34. Schulz S, Hahn U, Rogers J. Semantic Clarification of the Representation of Procedures and Diseases in SNOMED((R))CT. *Stud Health Technol Inform*. 2005;116: 773–8.
35. Schulz S, Hanser S, Hahn U, et al. The semantics of procedures and diseases in SNOMED CT. *Methods Inf Med*. 2006;45:354–8.
36. Schulz S, Suntisrivaraporn B, Baader F, et al. SNOMED reaching its adolescence: ontologists' and logicians' health check. *Int J Med Inform*. 2009;78(Suppl 1):S86–94.
37. Zhu X, Fan JW, Baorto DM, et al. A review of auditing methods applied to the content of controlled biomedical terminologies. *J Biomed Inform*. 2009;42:413–25.
38. SNOMED CT User Guide.

39. Cormen TH, Leiserson CE, Rivest RL, *et al.* *Introduction to algorithms*. MIT Press and McGraw-Hill, 2001.
40. Elhanan G, Perl Y, Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. *J Am Med Inform Assoc*. 2011;18(Suppl 1): i36–44.
41. Agrawal A, Perl Y, Chen Y, *et al.* Identifying Inconsistencies in SNOMED CT Problem Lists using Structural Indicators. AMIA Annual Symposium Proceedings 2013:17–26.
42. Wei D, Halper M, Elhanan G, *et al.* Auditing SNOMED relationships using a converse abstraction network. AMIA Annual Symposium Proceedings 2009:685–9.
43. Wang Y, Wei D, Xu J, *et al.* Auditing complex concepts in overlapping subsets of SNOMED. AMIA Annual Symposium Proceedings 2008:273–7.
44. Agrawal A, He Z, Perl Y, *et al.* The readiness of SNOMED problem list concepts for meaningful use of EHRs. *Artif Intell Med* 2013;58:73–80.
45. US Edition of SNOMED CT. 2013 September 2013 [cited 2013 9 September]. http://www.nlm.nih.gov/research/umls/Snomed/us_edition.html
46. Chuang J-H, Hripcsak G, Heitjan DF. Design and Analysis of Controlled Trials in Naturally Clustered Environments Implications for Medical Informatics. *J Am Med Inform Assoc*. 2002;9:230–8.
47. Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. *J Am Med Inform Assoc*. 2010;17: 675–80.

AUTHOR AFFILIATIONS

¹Computer Science Department, New Jersey Institute of Technology, Newark, New Jersey, USA

²Computer Information Systems Department, BMCC, CUNY, New York, New York, USA

³Department of Computer Science, Manhattan College, Riverdale, New York, USA

⁴NLM/NIH, Bethesda, Maryland, USA

⁵Department of Biomedical Informatics, Columbia University, New York, New Jersey, USA