

Analysis of Error Concentrations in SNOMED

Michael Halper, PhD¹, Yue Wang, MS², Hua Min, MD, PhD³, Yan Chen, MS^{2,4},
George Hripcsak, MD⁵, Yehoshua Perl, PhD², Kent A. Spackman, MD, PhD⁶

¹Kean University, Union, NJ; ²NJIT, Newark, NJ;

³Fox Chase Cancer Center, Philadelphia, PA; ⁴BMCC, CUNY, New York, NY;

⁵Columbia University, New York, NY; ⁶Oregon Health & Science University, Portland, OR

Abstract

Two high-level abstraction networks for the knowledge content of a terminology, known respectively as the “area taxonomy” and “p-area taxonomy,” have previously been defined. Both are derived automatically from partitions of the terminology’s concepts. An important application of these networks is in auditing, where a number of systematic regimens have been formulated utilizing them. In particular, the taxonomies tend to highlight certain kinds of concept groups where errors are more likely to be found. Using results garnered from applications of our auditing regimens to SNOMED CT, an investigation into the concentration of errors among such groups is carried out. Three hypotheses pertaining to the error distributions are put forth. The results support the fact that certain groups presented by the taxonomies show higher error percentages as compared to other groups. The bootstrap is used to assess their statistical significance. This knowledge will help direct auditing efforts to increase their impact.

Introduction

SNOMED CT has an inherent abstraction framework comprising a collection of top-level hierarchies that partition its concepts [1]. In previous work, we have developed two abstraction networks for a SNOMED hierarchy called the *area taxonomy* and *p-area taxonomy* [2]. Both are derived automatically from partitions of SNOMED’s concepts. Overall, they serve as summarizations for the distribution of SNOMED’s (attribute) relationships and convey aspects of the hierarchical grouping of concepts.

The taxonomies have proven to be particularly useful in the context of auditing, an important part of any terminology’s maintenance cycle [3]. In fact, a number of systematic auditing regimens based on the taxonomies have been formulated [2]. We have found that certain errors at the concept level manifest themselves as anomalies at the taxonomy level, thus giving the auditor clues as to where to search for errors productively. The application of such auditing regimens has proven to be fruitful, as reported in [2].

In this paper, we seek to formally characterize the concentrations of errors with respect to the concept groups presented by our taxonomies and thus further establish their efficacy as vehicles for auditing. Three hypotheses pertaining to error concentrations are set forth. Previously collected data [2] from the auditing process are tabulated in support of these hypotheses.

The results indeed support our hypotheses. The statistical significance of these results is analyzed using the bootstrap [4]. Since auditing resources are typically limited, the hypotheses can guide auditing efforts to concentrate on parts of the taxonomy where more errors are likely to be found.

Background

SNOMED is an important reference terminology being used in a variety of applications worldwide. Its January 2007 edition consists of a collection of more than 308,000 active concepts arranged in a polyhierarchy of subsumption (IS-A) links. These concepts are organized into disjoint top-level hierarchies, including Clinical finding, Procedure, Body Structure, etc. SNOMED’s concepts exhibit a wide range of relationships such as *specimen procedure*, *has active ingredient*, and *measurement method*.

Auditing is an essential part of SNOMED’s maintenance, and various techniques have been proposed. For example, ontological and linguistic approaches have been applied [5,6]. In [7], SNOMED’s native description-logic (DL) formalism is used in an effort to see how well its IS-A hierarchy conforms to a set of accepted ontological principles. In our work, we are not postulating specific principles and then searching for transgressions of these. Instead, we are carrying out an overall structural analysis, building abstraction networks, and hypothesizing about where errors—of any variety—are liable to be discovered. This paper focuses on the analysis of empirical results obtained in SNOMED’s Specimen hierarchy.

The foundation of our approach is the *area taxonomy*, derived from a partition of SNOMED’s concepts based on their respective sets of relationships. An *area* is defined as a maximal set of concepts that exhibit the exact same set of relationships, irrespective of the targets (or fillers) of those relationships. Collectively, the set of all areas forms a partition because a concept can belong to one and only one area. In the *area taxonomy*, an area is denoted as a node labeled with its unique combination of relationships. Furthermore, these nodes are connected via hierarchical *child-of* links that capture the underlying IS-A links of the terminology. A root of an area is a concept with no IS-A relationships to any other concepts in its area. An excerpt consisting of seven areas of the *area taxonomy* extracted from SNOMED’s Specimen hierarchy can be seen in Figure 1. In the lower left of the figure, we see the area {*specimen source topography*, *specimen*

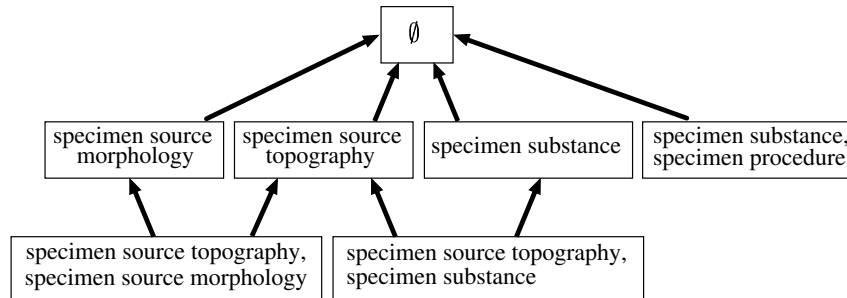


Figure 1: Seven areas from the SNOMED Specimen hierarchy

source morphology}, whose name indicates that there is a group of concepts having *specimen source topography* and *specimen source morphology* as their only relationships. Moreover, the concepts in that area are all the concepts of SNOMED that have exactly those two relationships. Concepts belonging to this area include *Cyst tissue* and *Skin lesion sample*. The uppermost area named \emptyset (the symbol for the empty set) has concepts with no relationships at all. The *child-of* from $\{specimen\ source\ topography, specimen\ source\ morphology\}$ to $\{specimen\ source\ topography\}$ expresses the fact that a root concept in the former area has a parent concept in the latter area from which *specimen source topography* is inherited.

level grouping called a *region* involving collections of p-areas. A region contains a group of p-areas whose roots obtain their (identical) set of relationships in the exact same manner: through explicit introduction, or explicit inheritance, or a combination of both. Figure 2 shows three areas, $\{specimen\ source\ morphology\}$, $\{specimen\ source\ topography\}$, and $\{specimen\ source\ topography, specimen\ source\ morphology\}$, from the p-area taxonomy. The bottom area has four roots, *Cyst tissue*, *Vegetation from heart valve*, *Tissue specimen obtained from ulcer*, and *Skin lesion sample*, and it thus has four p-areas with the corresponding names. The numbers in parentheses indicate the amount of concepts in the p-areas. Moreover, there are two regions, $\{specimen\ source\ topography, specimen\ source\ morphology\}$ and $\{specimen\ source\ topography, specimen\ source\ morphology^*\}$, separated from each other by a dashed line. The latter (on the right side in the figure) contains one p-area, *Skin lesion sample*, that obtains *specimen source topography* via inheritance but explicitly introduces *specimen source morphology*, as indicated by the “*” suffix. The other region contains the three p-areas *Cyst tissue*, *Vegetation from heart valve*, and *Tissue specimen obtained from ulcer*. The roots of each obtain their two relationships strictly via inheritance, as indicated by the lack of the “*” symbol. Each of the top two areas consists of a single region, and, to save space, only one of its p-areas is shown inside.

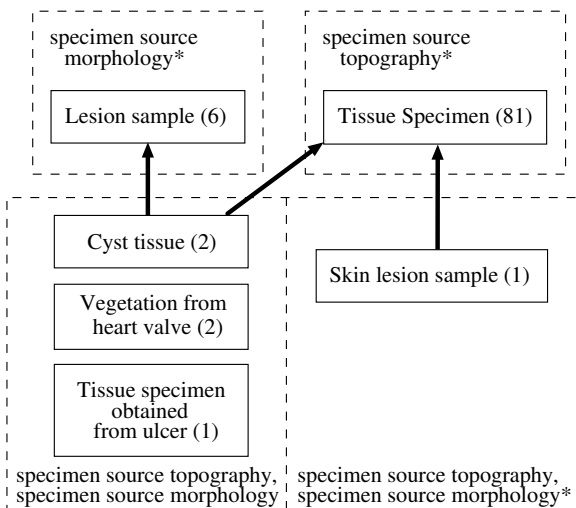


Figure 2: Excerpt of p-area taxonomy showing three areas, four regions, and six p-areas

An area can have multiple roots, and the p-area taxonomy offers a further division of SNOMED based on these important concepts. The foundational concept group of the p-area taxonomy is the *p-area* (short for *partial-area*), consisting of a single root together with all its descendants within one area. This finer-grained abstraction network thus comprises concept groups of structural similarity and common hierarchical ancestry. A p-area is named after its root since all other concepts in the p-area are specializations of the root. The area nodes of the area taxonomy are augmented in the p-area taxonomy to include their constituent p-areas. The p-area taxonomy also includes an intermediate-

The two taxonomies have demonstrated their worth in the context of auditing, where we have previously formulated a number of systematic regimens based on them. Certain kinds of errors at the concept level tend to reveal themselves as anomalies at the taxonomy level. The taxonomies also support *group-based* auditing, where concepts of purported structural similarity and hierarchical proximity can be reviewed in concert in order to more easily find inconsistencies and omissions. In this paper, we are interested in focusing on the efficacy of auditing done with respect to certain regions and p-areas.

Methods

The auditing regimens under investigation are based on specific concept groups presented automatically by our partitioning and abstraction methodologies. In particular, three regimens focused respectively on two kinds of regions and small-sized p-areas are applied to a top-level hierarchy of SNOMED. Let us note that our au-

ding and the subsequent analysis carried out here are based on the inferred (distributed) view of the terminology, i.e., the results after the DL classifier has computed all entailed subsumption relationships.

There are actually three distinct kinds of regions, two of which are pertinent to auditing. A *strict inheritance* region is one where all the relationships are obtained (by the roots) via inheritance. The region {*specimen source topography, specimen source morphology*} is an example. A *strict introduction* region is characterized by the lack of inheritance at the roots. In other words, all exhibited relationships are introduced explicitly at each of the roots. Each of the areas in Figure 1 that is *child-of* the area \emptyset consists of one region which is a strict introduction region since there are no relationships to inherit from \emptyset . A *mixed* region is one that is neither strict inheritance nor strict introduction. That is, some relationships are inherited and some are introduced. An example is {*specimen source topography, specimen source morphology**}.

For the sake of comparison, all the concepts in the chosen SNOMED hierarchy are reviewed for errors. Based on the overall outcomes of these efforts, the validity of the following three hypotheses pertaining to the efficacy of our auditing regimens is investigated.

Hypothesis 1: There is a higher likelihood for the existence of concept errors in strict inheritance regions than in strict introduction regions or mixed regions. ■

Hypothesis 2: There is a higher likelihood for the existence of concept errors in mixed regions than in strict introduction regions. ■

The idea underlying these two hypotheses has to do with hierarchical complexity accumulated in the inheritance process. When a relationship is inherited, it comes down through a path of ancestors who contribute—in addition to the relationship—their accumulated definitional knowledge to the descendant.

Typically, at each level, a constraint or limiting scope is added. Such additional knowledge is sometimes manifested as a more detailed concept name. For example, consider the path from *Specimen* to *Cyst tissue*. It goes through the concepts *Lesion sample* (introducing *specimen source morphology*) and *Specimen from cyst*. Naturally, each concept along the path is more specialized than its parent. We refer to the specialized knowledge accumulated along the path as the hierarchical complexity.

When a concept inherits a relationship, the path has to go through an area where that relationship is introduced. Traversing an area may mean visiting several concepts (e.g., two from the *Lesion sample* p-area above). If a concept introduces a relationship instead, then a subpath going through an area for the sake of picking up the relationship can be avoided, making the overall path shorter. For example, *Hematological sample*, the root of the only p-area (of 26 concepts) in the strict introduction region of the area {*specimen substance, specimen procedure*} (the rightmost area in Figure 1). That concept has just one parent *Specimen*, belonging to the area \emptyset , and introduces its own two re-

lationships without gaining hierarchical complexity. In general, an inherited relationship implies more hierarchical complexity than an introduced relationship.

A strict inheritance region implies more paths, each of which must travel through areas where inherited relationships are introduced and collected from. This in turn implies that concepts in such a region will, in general, have more ancestors and more hierarchical complexity. The case of strict introduction is of lower hierarchical complexity due to the fact that no extra path is needed to deliver the relationship. A mixed region has an intermediate hierarchical complexity as it inherits some relationships (via an ancestor path) but introduces others (without going through extra areas). Our underlying assumption and motivation for the two hypotheses is that concepts with higher hierarchical complexity are more prone to modeling errors.

An example of this can be found in the context of the p-areas in Figure 2. The concept *Skin lesion sample*, the root of its p-area in the region {*specimen source topography, specimen source morphology**}, has a single parent *Skin tissue specimen*, residing in the p-area *Tissue specimen*, from which it inherits *specimen source topography*. *Skin lesion sample* explicitly introduces *specimen source morphology*, providing further hierarchical complexity. The concept *Cyst tissue* in the neighboring region {*specimen source topography, specimen source morphology*} inherits those two relationships respectively from its parents *Tissue specimen* (the root of its p-area) and *Specimen from cyst* (in the p-area *Lesion sample*). Two ancestor paths through these two parents lead to *Cyst tissue*. The one through the latter parent was described above. Therefore, *Skin lesion sample* obtains its relationships in a simpler hierarchical configuration than that needed for *Cyst tissue* and is thus less complex.

Hypothesis 3: There is a higher likelihood for the existence of concept errors in small p-areas than in large p-areas. ■

This hypothesis indicates the expectation that a small group of concepts similar in their structure and semantics is less likely to be properly modeled and have proper classifications than a similarly constituted large group with a common structure and semantics. That is, the high incidence of a combination of a structure and semantics supports its feasibility, while a rarely seen combination raises questions about whether it is the correct structure and root for its few elements. (Let us note that a similar hypothesis was proposed and verified [3] in the context of the NCI Thesaurus [8].)

We use the bootstrap [4] to assess the statistical significance of the hypotheses while accounting for the clustering of concepts within p-areas.

Results

The auditing regimens pertaining to strict inheritance and strict introduction regions and small p-areas were applied to the Specimen hierarchy of SNOMED, and the resulting error counts with respect to these vari-

ous groups have been tabulated (Table 1 and Table 2). For example, within the Specimen hierarchy, there are nine strict inheritance regions encompassing 28 p-areas and a total of 83 concepts (see the second row of Table 1). Among those concepts, 16 errors were discovered, amounting to a percentage of 19.28. The percentages of errors for the other two kinds of regions are: mixed: 12.60%; and strict introduction: 3.28%. (Note that the first row in Table 1 shows the data for the area \emptyset whose only region is a special case of a region without any relationships at all.) With respect to the Specimen hierarchy's overall 1,056 concepts, 97 (9.19%) concept errors were found. These figures confirm Hypotheses 1 and 2.

The error totals found in the context of p-areas of various sizes can be seen in Table 2. The table, in fact, breaks the space of p-areas into two: those with seven or fewer concepts and those with eight or more. P-areas in the former range are deemed to be "small"; those in the latter, large. As can be seen from the table, 10.68% of the concepts in small p-areas are in error, while the number is only 6.83% for large p-areas. This result confirms Hypothesis 3.

While strict inheritance had a nominally greater error rate than mixed or strict introduction, the differences were not statistically significant, most likely due to the relatively small number of strict inheritance p-areas. Mixed was greater than strict introduction, and the difference in this case was statistically significant.

The error rate for smaller p-areas was nominally higher than that for larger p-areas, but again the difference was not statistically significant, perhaps due to the small number of large p-areas.

Table 3 presents a sample of 15 errors discovered with the use of our taxonomy auditing regimens in the context of the 2004 release of SNOMED. In each case, the concept's region, p-area, kind of error, and required correction are listed. The table is subdivided with respect to the different kinds of regions (SIT = strict introduction; SIH = strict inheritance; MIX = mixed). The second row, for example, shows that the concepts *Body fluid sample* and *Body fluid specimen* were found to be independent concepts, when in fact they should be synonyms of each other. Furthermore, the fifth row indicates the discovery of a missing IS-A between the child *Body fluid sample* and the parent *Fluid sample*.

All the errors in Table 3 were confirmed by one of the authors (KAS) who is the Scientific Director of SNOMED. Most of the errors have already been corrected as of the 2007 release. The others will be dealt with in the upcoming release.

Discussion

The hypotheses suggest that ever-limited auditing resources be concentrated on small p-areas of strict inheritance and mixed regions in order to try to maximize the number of errors found for a given amount of effort. The scope of the auditing experiments was limited to the Specimen hierarchy, which represents a relatively small portion of SNOMED. While the tabulated per-

centages support our hypotheses, the current numbers are too small to achieve statistical significance for two out of the three hypotheses. There is thus a need to apply our auditing methodologies to additional hierarchies to further examine our hypotheses and especially to further support their statistical analysis. We expect similar results for other hierarchies.

Each hierarchy of SNOMED is different in its size, height, width, number of defined relationships, and pattern of relationship introduction. These characteristics will naturally be reflected in the taxonomies that abstract the hierarchies. It is not clear how those differences will affect the distribution of errors among regions and p-areas. While the reasoning for the hypotheses suggests a general phenomenon, further experiments are required for verification. In particular, it is difficult to predict the range of small and large p-areas for the context of the hypotheses. We followed an empirical approach suggesting 7 as the size threshold.

Since SNOMED uses a DL formalism, it can be fruitful to go outside that realm in an effort to uncover errors. As we saw, SNOMED's DL-classifiers failed to find certain errors (such as the fact that *Eye fluid sample* is a child of *Body fluid sample*) that were found with our structural methodologies.

Conclusion

The area taxonomy and related p-area taxonomy were previously presented as important abstraction networks for SNOMED. In particular, these automatically-generated, high-level views have formed the bases of a number of systematic auditing regimens. Focusing on various concept groupings presented by the two taxonomies has shown to be productive in searches for errors. In this paper, we have done an analysis of error concentrations with respect to the taxonomies in an effort to formally characterize the efficacy of the auditing regimens. The results of auditing SNOMED's Specimen hierarchy were presented and analyzed. Three kinds of concept groups, strict-inheritance regions, mixed regions, and small p-areas, proved to be fruitful in bringing errors to light. Three hypotheses were proposed and confirmed in this regard.

References

1. SNOMED: SNOMED CT. Available at <http://www.snomed.org/snomedct/index.html>. Accessed February 20, 2007.
2. Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural Methodologies for Auditing SNOMED. *Journal of Biomedical Informatics* (in press; available at the Web site).
3. Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as Part of the Terminology Design Life Cycle. *JAMIA* 2006;13(6):676-690.
4. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press; 1993.
5. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT.

Table 1: Errors across kinds of regions

Kind of Region	#	# P-areas	# Concepts	# Errors	% Errors
∅	1	1	30	2	6.67
Strict inheritance	9	28	83	16	19.28
Mixed	12	266	516	65	12.60
Strict introduction	6	157	427	14	3.28
Total:	28	452	1,056	97	9.19

Table 2: Errors across ranges of p-area size (# of concepts)

P-area Size	# P-areas	# Concepts	# Errors	% Errors
1-7	427	646	69	10.68
8 or more	25	410	28	6.83
Total:	452	1,056	97	9.19

- In: Fieschi M, et al, eds. Proc. Medinfo 2004. San Francisco, CA; 2004. p. 482-486.
- Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in Medical Ontologies: Where Do They Come From and How Can They Be Detected? In: Pisanelli DM, ed. Ontologies in Med.: Proc. Workshop on Med. Ontologies. Rome; 2003. p. 145-164.
 - Bodenreider O, Smith B, Kumar A, Burgun A. Investigating Subsumption in DL-Based Terminologies: A Case Study in SNOMED CT. In: Hahn U, Schulz S, Cornet R, eds. Proc. KR-MED 2004. Whistler, Canada; 2004. p. 12-20.
 - NCI Terminology Browser. Available at <http://nciterns.nci.nih.gov/NCIBrowser/Dictionary.do>. Accessed February 27, 2007.

Table 3: Sample of errors discovered in SNOMED (sp = specimen; src = source)

Concept Name	Region	P-area	Error type	Correction
Mushroom specimen	∅	Sample	Missing relationship	Missed Relationship: sp substance
Body fluid specimen/Body fluid sample	sp substance* (SIT)	Body fluid specimen	Synonym problem	Body fluid specimen/Body fluid sample made synonyms
Specimen from ear / Ear sample	sp src topography* (SIT)	Specimen from ear/ Ear sample	Synonym problem	Specimen from ear / Ear sample are synonyms
Surgical excision sample	sp procedure* (SIT)	Surgical excision sample	Missing child	Missed child: Specimen obtained by standard surgical excision
Fluid sample	sp substance* (SIT)	Fluid sample	Missing child	Missed child: Body fluid sample
Tendon biopsy sample	sp src topography* (SIT)	Musculoskeletal sample	Missing relationship	Missed relationship: specimen procedure
Saliva specimen	sp src topography, sp substance (SIH)	Saliva specimen	P-area root with wrong parent (Fluid Sample)	Right parent: Body fluid Sample
Specimen from lung obtained by fine needle aspiration procedure	sp src identity, sp src topography, sp procedure (SIH)	Specimen from lung obtained by fine needle aspiration procedure	Wrong parent (Specimen from lung obtained by biopsy (specimen))	Right Parent: Specimen obtained by fine needle aspiration procedure
Respiratory fluid specimen	sp src topograph, sp substance (SIH)	Respiratory fluid specimen	Wrong concept name for root	Right root concept name: Upper respiratory fluid specimen
Throat washings (specimen)	sp src topograph, sp substance (SIH)	Respiratory fluid specimen	Missing relationship	Missed relationship: sp src procedure
tissue specimen from pancreas	sp src identify, sp src topograph* (MIX)	Specimen from digestive system	Wrong target (Specimen source topography: large intestinal structure)	Right targets: pancreatic structure and body tissue structure
Leukocyte specimen from patient	sp src identify, sp src topograph* (MIX)	Specimen from digestive system	Wrong target (Specimen source topography: large intestinal structure)	Right targets: specimen source topography: leukocyte
Gallstone sample	sp src morphology, sp substance* (MIX)	Biliary stone sample	Wrong relationship (sp src morphology)	Right relationship: sp substance
Gastric washings	sp src topography, sp substance* (MIX)	Gastrointestinal fluid sample	Missing relationship	Missed relationship: sp procedure
Eye fluid sample	sp src topography, sp substance* (MIX)	Eye fluid sample	Missing parent	Missed parent: Body fluid sample

SIT: Strict Introduction Region

SIH: Strict Inheritance Region

MIX: Mixed Region