Complexity Measures to Track the Evolution of a SNOMED Hierarchy

Duo Wei, BS[1], Yue Wang, MS[1], Yehoshua Perl, PhD[1], Junchuan Xu, MD[1], Michael Halper, PhD[2], Kent A. Spackman, MD, PhD[3]

[1]NJIT, Newark, NJ; [2]Kean University, Union, NJ; [3]IHTSDO, Copenhagen, Denmark

Abstract

SNOMED CT is an extensive terminology with an attendant amount of complexity. Two measures are proposed for quantifying that complexity. Both are based on abstraction networks, called the area taxonomy *and the* partial-area taxonomy*, that provide, for example, distributions of the relationships within a SNOMED hierarchy. The complexity measures are employed specifically to track the complexity of versions of the Specimen hierarchy of SNOMED before and after it is put through an auditing process. The pre-audit and post-audit versions are compared. The results show that the auditing process indeed leads to a simplification of the terminology's structure.*

Introduction

SNOMED CT [1] is large and complex, with its July 2007 release containing about 376,000 concepts organized into 19 hierarchies. Due to its creation via the integration of SNOMED RT and the CTV3, it is unavoidable that errors have been introduced during SNOMED's design and ongoing evolution.

Lateral relationships are defined between hierarchies to govern the relationship structure of concepts of the respective hierarchies. In [2,3], we utilized these relationships, and their inheritance patterns within SNOMED hierarchies, to introduce structural methodologies for auditing those hierarchies. The methodologies utilize two abstraction networks, called the *area taxonomy* and *partial-area taxonomy*, to capture the structure of a hierarchy in a compact manner. The taxonomies highlight where errors tend to concentrate [3] and offer techniques to detect them [2]. The errors reported in [2,3] were subsequently corrected in later releases of SNOMED.

In this paper, we investigate whether, in addition to the elimination of the errors, their correction simplifies the structure of the hierarchy. Our hypothesis is that errors contribute to structural disorderliness. Hence, we expect to see a simplification of the hierarchy structure due to the reduction of such disorderliness after an auditing regimen has been carried out. For this, we need to posit a way to assess the complexity/simplicity of a hierarchy. Our previously defined taxonomies offer a natural solution to this problem since they are derived via a structural analysis of the underlying SNOMED hierarchy. Two proposed assessment measures are used to track changes in complexity. In particular, the measures are applied to the Specimen hierarchy to track its evolution from its July 2004 version (the one prior to our corrections) to its July 2007 version. The results show a reduction in the hierarchy's complexity.

Background

Auditing is essential to any terminology's maintenance [4]. Various techniques have been proposed and applied to SNOMED. For example, ontological and linguistic techniques have been utilized [5]. An analysis has been carried out to determine how well SNOMED's IS-A hierarchy adheres to certain ontological principles [6]. This latter work makes use of SNOMED's description-logic (DL) formalism. Such DL representations have also been used for the development of algorithms to detect terminological inconsistencies [7] and synonymy [8].

Our own auditing approaches are based on two abstraction networks that have been designed to sit above SNOMED's concept hierarchy: the *area taxonomy* and the *partial-area taxonomy* [2–4]. Both of these are derived automatically from the respective lateral (i.e., non-IS-A) relationships exhibited by the concepts. The latter also relies on local configurations of the IS-A hierarchy itself. In the following, we give details of each of these two networks.

While other auditing techniques follow the hierarchical design and the DL model of SNOMED, our techniques follow the lateral relationship aggregation of a SNOMED hierarchy as captured by the taxonomies. This property enables our techniques to expose errors that may remain hidden using hierarchical techniques similar to SNOMED's underlying design. These errors may thus have escaped the attention of the SNOMED editors in the design process.

The area taxonomy has as its foundation the notion of *area*. Each SNOMED hierarchy has a given set of relationships that can be defined for its concepts. An area, defined with respect to a given subset of the hierarchy's relationships, is the entire set of terminology concepts that exhibit exactly this subset of the hierarchy's relationships. A concept's membership in an area is based on the domain of its relationships and is irrespective of its targets (or fillers) of those particular relationships. From the definition, we can see that the areas of a SNOMED hierarchy form a partition. That is, each concept belongs to one and only one area.

The Specimen hierarchy has five relationships that can be defined for its concepts. They describe different aspects of the specimen represented by a concept. They are *substance*, *morphology*, *identity*, *procedure*, and *topography*. For example, the concept *Breast cyst fluid sample* has three relationships, *topography* pointing to the concept *Breast* (of the Body Structure hierarchy), *morphology* to *Cyst*, and *substance* to *Fluid*.

An area name is the list of the respective relationships exhibited by the area. For example, the concept *Breast cyst fluid sample* belongs to the area

{*morphology, topography, substance*}. The concept *Blood bag specimen* has, for example, a relationship *identity* to the concept *Blood bag* of the Physical Object hierarchy, since the device (blood bag) used in this sample has to be identified. The concept belongs to the area {*identity*}.

The areas form the nodes of the area taxonomy, which essentially is a hierarchical graph structure. The edges of the graph are hierarchical relationships called *child-of*'s. These are derived from SNOMED's own IS-A links. The area taxonomy serves as a means of conveying the overall distribution of relationships throughout SNOMED's hierarchy.

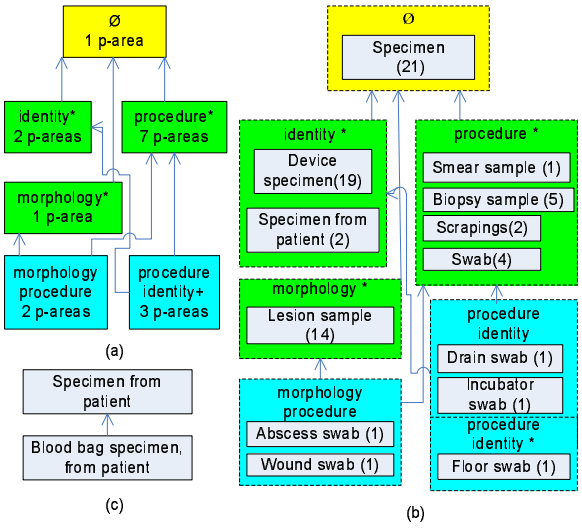

Figure 1: (a) Areas from SNOMED's Specimen hierarchy; (b) areas and partial-areas; (c) partial-area's concept hierarchy

A portion of the area taxonomy for SNOMED's Specimen hierarchy is shown in Figure 1(a). The area boxes are color-coded according to their levels (i.e., numbers of relationships). This excerpt contains six areas. The top area (at "Level 0") is denoted using ∅ ("empty set") because it has no relationships at all. At Level 1, displayed in green, we find the areas {*morphology*}, {*identity*}, and {*procedure*}. On Level 2, highlighted in blue, are {*morphology*, *procedure*} and {*procedure*, *identity*}. Note that no concepts are shown in the diagram of an area taxonomy. But examples of some concepts belonging in these areas are: *Specimen* (∅); *Lesion sample* ({*morphology*}); *Device specimen* ({*identity*}); *Swab* ({*procedure*}); and *Drain swab* ({*procedure*, *identity*}).

The *child-of* relationships—the arrows in the figure—are derived from the IS-As of the *roots* of the areas. A root of an area is a concept, of the area, whose parents all reside in other areas. In other words, a root has no IS-A to another concept in its area. An area may have more than one root. A *child-of* from an area, say, *X* to an area *Y* indicates that some root of *X* has a parent in *Y* (which is not necessarily a root of *Y*). As an example, the *child-of* from {*morphology*, *procedure*} to {*morphology*} expresses the fact that a root (namely, *Wound swab*) in the former has a (non-root) parent (*Specimen from wound*) in the latter. The relationship *morphology* is in fact inherited via that IS-A.

The partial-area taxonomy extends the area taxonomy by further refining areas when there are multiple roots. In addition to areas, the partial-area taxonomy includes *partial-areas*, each being a set of concepts comprising a single root and all its descendants within one area. The nodes representing the partial-areas are embedded in the nodes representing their respective areas. The node-label of a partial-area is its constituent root, which hierarchically sits atop all its other concepts and thus generalizes them. Note that while the root concepts' names are shown in the partial-area taxonomy, the names of non-root concepts are hidden.

Figure 1(b) is the portion of the Specimen hierarchy's partial-area taxonomy derived from the area-taxonomy in Figure 1(a). (Not all partial-areas are shown.) For example, we see that in the area {*procedure*} four partial-areas are shown, *Scrapings*, *Smear sample*, *Biopsy sample*, and *Swab*, corresponding to four of its roots. There are actually three more roots, and thus three more partial-areas, which will be seen in the full partial-area taxonomy. The number in parentheses alongside a partial-area name indicates the number of concepts contained within the partial-area. For example, in the area {*identity*}, we see a partial-area *Specimen from patient (2)* whose second non-root (hidden) concept is *Blood bag specimen from patient*, mentioned above. (See Figure 1(c) for the concept hierarchy of this partial-area.)

As can be seen for {*procedure*, *identity*} in Figure 1(b), the partial-area taxonomy includes an additional grouping of partial-areas that is used to convey the way in which root concepts obtain their relationships. In general, a concept can obtain its set of relationships via explicit introduction, or explicit inheritance, or a combination of these. In the context of auditing, we have found this knowledge to be useful in identifying concepts with high probability of errors [2,3]. Thus, we have defined the notion of *region* (of an area), which is a group of partial-areas whose roots obtain their relationships in an identical manner. The three partial-areas of {*procedure*, *identity*} are divided into two regions: {*procedure*, *identity*} and {*procedure*, *identity**}. The first holds the partial-areas *Drain swab* and *Incubator swab*. Both roots obtain the two relationships via inheritance. The concept *Drain swab* inherits the *procedure* relationship from the root *Swab* in {*procedure*}. It inherits *identity* from *Drain device specimen* residing in the partial-area *Device specimen* of {*identity*}. The root of *Floor swab* in the second region also gets *procedure* through inheritance; however, it obtains *identity* by explicit introduction, as indicated by the "*" suffix on the relationship's name within the region's name. The corresponding area {*procedure*, *identity*$^+$} in Figure 1(a) uses the symbol "+" to indicate that some roots inherit *identity* while others introduce it.

We note that partial-areas are not necessarily disjoint. For example, *Blood bag specimen from patient* in the area {*identity*} is in both the partial-areas *Device specimen* and *Specimen from patient*. Hence, {*identity*} has 20 ($= 19 + 2 - 1$) concepts, not 21.

In our previous work [2–4], we have proposed the use of the two abstraction networks as the basis for various auditing regimens. As we have shown, these networks afford the auditor novel views of the terminology's content and also help to highlight portions that are ripe for deeper investigation. A variety of kinds of errors have been discovered in this way, including redundant concepts, incorrect IS-A configurations, erroneous relationships, and general modeling errors. Note that we assume the correctness of the relationship distribution in the creation of the taxonomies. When incorrect relationships are found and corrected, the re-created taxonomies will typically have a different structure.

In this paper, we are further exploiting the two taxonomies as a means for measuring the complexity of SNOMED's concept network. We are interested in examining its change in complexity following auditing.

Methods

The issue we are investigating in this paper is how to measure the simplicity of a SNOMED hierarchy. One natural criterion is a global weighting function for a hierarchy, such as size or height (number of levels in the longest hierarchical path). Indeed, in the comparison of such measures following the first audit of the 2004 SNOMED, the number of concepts was reduced from 1,056 to 1,044, and the height was reduced from 12 to 10. At the same time, SNOMED's total concepts went from 357,134 to 364,461. Furthermore, only two hierarchies of SNOMED decreased in size during this period, the second of which was the huge Clinical Finding hierarchy. We attribute the decrease in size to the errors of duplicate concepts such as *Ear sample* and *Specimen from ear* [2]. Such errors were caused by failing to identify the synonymy of "sample" and "specimen" when integrating SNOMED RT and CTV3. The reduction in height can be attributed to finding errors in some of the most complex concepts in the hierarchy, from which the longest hierarchical path originates.

However, such global measures fail to take into account the role of lateral relationships in the complexity of the concepts. The size measure accounts only for erroneous concepts eliminated from the hierarchy (due to duplicates or improper concepts) but not for other errors that were corrected. It is also influenced by concepts added to the hierarchy as part of normal maintenance. The height measure reflects only a few concepts at the bottom of the longest hierarchical path.

To illustrate the difficulty of using the size measure, the Specimen hierarchy grew from 1,044 concepts in July 2005 to 1,052 in January 2007, while no special auditing was applied. Finally, the number grew back to 1,056 (the original number in 2004) in July of 2007, following the second auditing effort.

As a more appropriate way of measuring the complexity of a hierarchy, we are suggesting to utilize our area taxonomy and partial-area taxonomy. The area taxonomy reflects the lateral relationships (just "relationships," for short) of all the concepts in the underlying SNOMED hierarchy. The partial-area taxonomy further shows hierarchical cohesiveness [9], where concepts subsumed under a common root concept are clustered into a *partial-area*. All these concepts elaborate the semantics of their root.

We assert that a concept with one relationship is simpler than a concept with two or more relationships since it is more general and expresses less detailed knowledge. For example, *Drain swab* is more complex than *Swab* or *Drain device specimen*. Similarly, *Skin ulcer swab* in the area {*morphology*, *topography*, *procedure*} is more complex than the one-relationship concepts *Swab*, *Skin sample*, and *Specimen from ulcer* from the respective partial-areas *Swab* of the area {*procedure*}, *Dermatological sample* of the area {*topography*}, and *Lesion sample* of the area {*morphology*}. The area levels of the area taxonomy serve to partition concepts of the hierarchy according to their numbers of relationships. If, as a result of an auditing phase, we see an increase in the number of concepts in a lower area level (consisting of areas with a lower number of relationships) at the expense of a decrease in the number of concepts in an higher area level (with a higher number of relationships), then this change can be interpreted as a simplification of the terminology structure. Of course, a concept must first be modeled with all its necessary relationships. A simpler representation is seen as a desired quality but is only secondary to correctness. Hence, the auditing process should not seek to delete required relationships just for the sake of simplification. However, as a result of auditing, we expect such simplification.

To formalize this measure, we define the function *num_cncpts_w_relshp_count*(n) to be the number of concepts in the hierarchy with exactly n relationships. This function is a structural measure as it depends solely on the number of relationships, not on their kind. It is a global structural measure for the complexity of the hierarchy because it is dependent on all concepts and their respective structures.

Another complexity measure concentrates on what is happening inside an area. An area may have several roots. Those roots are, in a sense, semantically independent of one another since none sits in an ancestor/descendant relationship to any other. Each root defines a partial-area, named after it, which contains all concepts that are its specializations in the area. Each partial-area expresses an overarching semantics for its constituent concepts, each being a kind of the root concept. For example, all 19 concepts in the partial-area *Device specimen* are concepts that are specimens derived from various devices, such as *Catheter specimen*. That is, the division of an area into partial-areas serves

to divide all concepts of the same structure (expressed by the area's name) into groups of semantically similar concepts. The semantics of each group is expressed explicitly by the partial-area's name.

Thus, an area with fewer partial-areas for the same number of concepts is an area with a smaller number of sets with different semantics. Such an area is considered simpler than an area with the same number of concepts in more partial-areas. Similarly, an area with more concepts, but with the same number of partial-areas as before, is considered simpler. The ratio of the number of concepts to the number of partial-areas in an area can be used as a good measure of the simplicity of the area. In fact, we define the *simplicity ratio* of an area X as follows. Let $E(X)$ denote the extent (i.e., set of concepts) of X and let $P(X)$ be the *set* of partial-areas in X. Then the simplicity ratio $S(X) = \frac{|E(X)|}{|P(X)|}$, where $|E(X)|$ ($|P(X)|$) is the number of concepts (partial-areas) in X (i.e., the cardinality of $E(X)$ ($P(X)$)). For example, in SNOMED 2004, the area $\{substance\}$ had 56 concepts distributed across 15 partial-areas, for a simplicity ratio $S(\{substance\}) = \frac{56}{15} = 3.73$. In SNOMED 2007, the same area has 81 concepts in ten partial-areas, for a ratio of 8.10. Hence, this area became simpler between 2004 and 2007. One can calculate the simplicity ratio for a whole level of areas having the same number of relationships. There were 399 concepts in 153 partial-areas exhibiting exactly one relationship in 2004. Hence, the simplicity ratio is 2.61. The ratio for the same level for 2007 is $\frac{468}{45} = 10.4$. Therefore, as a whole, the level of all areas of one relationship became simpler. We will compare all levels.

A possible impact of auditing is discovering that the root of a small partial-area, especially a singleton (i.e., a one-concept partial-area) should be a child of a concept in another partial-area. Hence, the small partial-area will be absorbed into the new parent's partial-area. For example, in 2004, $\{morphology\}$ had nine partial-areas, four of which are: *Specimen from abscess*, *Specimen from ulcer*, *Specimen from wound*, and *Lesion sample*. In 2007, the first three of these became part of the expanded *Lesion sample*, which now consists of 14 concepts. $S(\{morphology\})$ went from $\frac{15}{9} = 1.67$ to $\frac{14}{1} = 14.0$, reflecting a simplification that this area underwent when it was realized that abscess, ulcer, and wound were all kinds of lesions.

To summarize, we have introduced two structural measures of complexity, one lateral and one hierarchical. We utilize them to assess our hypothesis that following the application of an auditing process [2,3], the Specimen hierarchy got simpler.

Results

In the July '07 release of SNOMED, the area taxonomy for the Specimen hierarchy has a total of 24 areas. In the partial-area taxonomy, there are 361 partial-areas.

In comparison, the 2004 Specimen hierarchy had 22 areas and 451 partial-areas. Table 1 summarizes the number of concepts, partial-areas, and simplicity ratios for the areas of different levels (i.e., numbers of relationships). Columns 2 and 5 (both labeled "C") give the values of "*num_cncpts_w_relshp_count*" for the various levels in 2004 and 2007, respectively. In Level 0, the reduction reflects the discovery of hidden relationships for concepts (in area $\emptyset$) having no relationships previously. At Level 1, *num_cncpts_w_relshp_count* reflects a large increase in concepts with exactly one relationship in 2007, representing many more concepts with lower structural complexity. A similar increase occurs for Level 2. These increases are balanced by the decrease in concepts on Level 3 with higher structural complexity. This measure reflects our finding many concepts with unnecessary relationships during auditing. For example, in 2004, the partial-area *Specimen from digestive system* had an extraneous *identity* relationship which was subsequently removed from its 38 concepts.

Table 1: # Concepts ("C") / # partial-areas ("PA") for levels (2004 vs. 2007)

Level	2004 Version			2007 Version		
	C	PA	C/PA	C	PA	C/PA
0	29	1	29.0	21	1	21.0
1	399	153	2.6	468	45	10.4
2	430	186	2.3	517	269	1.9
3	194	107	1.8	48	44	1.1
4	4	4	1.0	2	2	1.0
Total:	1,056	451	2.3	1,056	361	2.9

Table 2: # Concepts / # partial-areas for areas with one relationship (2004 vs. 2007)

Area Name	2004 Version			2007 Version		
	C	PA	C/PA	C	PA	C/PA
substance	56	15	3.7	81	10	8.1
morphology	15	9	1.7	14	1	14.0
topography	297	112	2.7	333	25	13.3
procedure	12	9	1.3	20	7	2.9
identity	19	8	2.4	20	2	10.0

The simplicity ratio column in Table 1 gives these measures with respect to the taxonomy's various levels. The whole hierarchy became simpler with a decrease in the number of partial-areas. Level 1 became much simpler, while Levels 2 and 3 became somewhat more complex from 2004 to 2007. In Level 2, the increase in the number of concepts is with a similar increase in the number of partial-areas of a single concept which lost a wrong relationship. Table 2 presents in details the dramatic simplification of Level 1 by examining the ratio of each area separately. The large decrease in complexity occurs with the large decrease in the number of partial-areas for the areas $\{morphology\}$, $\{topography\}$, and $\{identity\}$.

Discussion

The first measure we introduced for complexity of a SNOMED hierarchy comes from the idea that a con-

cept with less lateral relationships is simpler than a concept with more. Our comparison of the 2004 and 2007 Specimen hierarchies indeed reveals a decrease in the number of concepts exhibiting three or four relationships (i.e., complex concepts) and a corresponding increase in the number of concepts with one or two relationships (i.e., simpler concepts). Table 1 can be used to calculate the global number of relationships in the Specimen hierarchy to find it is reduced from 1,857 in 2004 to 1,654 in 2007. (This count does not include occurrences of multiple targets for the same relationship with respect to the same concept.) The reduction of 203 erroneous relationships in a hierarchy of 1,056 concepts is a meaningful improvement in both quality and simplicity. Indeed, most of the errors discovered in the auditing process had to do with relationships. And the amount of deleted incorrect relationships goes up if one also considers the relationships that were found to be missing and were subsequently added (e.g., for $29 - 21 = 8$ concepts of area $\emptyset$ in Level 0) since those cancel the impact of the same number of deleted relationships. (Obviously, it is imperative that concepts have the correct relationships, even if it makes them more complex.)

Table 2 concentrates on the simplicity ratios of areas with one relationship. Not only did the number of concepts on that level grow from 399 to 468 (due to removing incorrect relationships), but the number of partial-areas was reduced from 153 to 45. That is, the concepts with one relationship were grouped in a better manner as a result of auditing. This was achieved primarily due to the corrections of faulty or omitted IS-As.

However, for Levels 2 and 3, the simplicity ratios decreased slightly in 2007. The main reason for this is the explosion in the number of partial-areas in two areas: $\{$*topography*, *substance*$\}$ and $\{$*topography*, *procedure*$\}$. There are many body parts and various procedures and substances. Even when grouped on Level 1, they have ten, seven, and 25 partial-areas, respectively. On Level 2, where many body parts are combined with many procedures, they produce 193 partial-areas, which are mostly singletons.

In this paper, we availed ourselves of the two audits that were applied to one hierarchy, namely, Specimen [2,3]. The results would be further strengthened by analyses of other, perhaps larger, SNOMED hierarchies and other similar terminologies, e.g., the NCI Thesaurus [10]. In future work, we plan to extend the monitoring of the evolution of the Specimen hierarchy over more stages of auditing, carried out using our own methodologies and those of the SNOMED maintenance personnel.

Conclusion

In this paper, we set out to track complexity measures of a terminology following its evolution. In particular, we explored whether, as a result of an audit, a SNOMED hierarchy became less complex in its structure. The underlying rationale is that errors correlate with disorderliness. Correction of errors is manifested by the terminology exhibiting a more orderly—indeed, simpler—structure.

Two measures were introduced to quantify the complexity/simplicity of a SNOMED hierarchy. They are based on characteristics of the area taxonomy and partial-area taxonomy abstraction networks that we previously introduced. Both networks are derived automatically via analysis of structural aspects of the hierarchy. The finding in this paper is that not only do the abstraction networks form the basis for fruitful auditing regimens, but they also help to provide the means for measuring the simplification that comes about as a result of the auditing.

Acknowledgment

This work was partially supported by the NLM under grant R-01-LM008912-01A1.

References

1. IHTSDO: SNOMED CT. Available at http://-www.ihtsdo.org/our-standards/snomed-ct. Accessed December 31, 2007.
2. Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. JBI. 2007;40(5):561–581.
3. Halper M, Wang Y, Min H, Chen Y, Hripcsak G, Perl Y, Spackman KA. Analysis of error concentrations in SNOMED. In: Teich JM, Suermondt J, Hripcsak G, eds. Proc. 2007 AMIA Annual Symposium. Chicago, IL; 2007. p. 314–318.
4. 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.
5. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. In: Fieschi M, Coiera E, Li YC, eds. Proc. Medinfo 2004. San Francisco, CA; 2004. p. 482–486.
6. 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.
7. Schlobach S, Huang Z, Cornet R, Van Harmelen F. Debugging incoherent terminologies. J. Autom. Reasoning. 2007;39:317–349.
8. Cornet R, Abu-Hanna A. Auditing description-logic-based medical terminological systems by detecting equivalent concept definitions. Int J. Med. Informatics. 2008;77(5):336–345.
9. Perl Y, Chen Z, Halper M, Geller J, Zhang L, Peng Y. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. JBI. 2003;35(3):194–212.
10. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. JBI. 2007;40(1):30–43.