Published in final edited form as:
J Biomed Inform. 2012 December ; 45(6): 1042–1048. doi:10.1016/j.jbi.2012.05.006.

A Study of Terminology Auditors' Performance for UMLS Semantic Type Assignments

Huanying (Helen) Gu[1], **Gai Elhanan**[2], **Yehoshua Perl**[2], **George Hripcsak**[3], **James J. Cimino**[4], **Julia Xu**[4], **Yan Chen**[5], **James Geller**[2], and **C. Paul Morrey**[6]

[1]New York Institute of Technology, New York, NY

[2]New Jersey Institute of Technology, Newark, NJ

[3]Columbia University, New York, NY

[4]NIH Clinical Center, Bethesda, MD

[5]BMCC, City University of New York, New York, NY

[6]Utah Valley University, Orem, UT

Abstract

Auditing healthcare terminologies for errors requires human experts. In this paper, we present a study of the performance of auditors looking for errors in the semantic type assignments of *complex* UMLS concepts. In this study, concepts are considered complex whenever they are assigned combinations of semantic types. Past research has shown that complex concepts have a higher likelihood of errors. The results of this study indicate that individual auditors are not reliable when auditing such concepts and their performance is low, according to various metrics. These results confirm the outcomes of an earlier pilot study. They imply that to achieve an acceptable level of reliability and performance, when auditing such concepts of the UMLS, several auditors need to be assigned the same task. A mechanism is then needed to combine the possibly differing opinions of the different auditors into a final determination. In the current study, in contrast to our previous work, we used a majority mechanism for this purpose. For a sample of 232 complex UMLS concepts, the majority opinion was found reliable and its performance for accuracy, recall, precision and the F-measure was found statistically significantly higher than the average performance of individual auditors.

Keywords

Auditing of terminologies; UMLS auditing; Semantic type assignments; Auditor performance; Auditor reliability; Auditing process; UMLS curators; Auditing errors; Quality assurance; Quality assurance evaluation; Auditing performance study

Correspondence Dr. Huanying Gu Computer Science Department New York Institute of Technology 1855 Broadway New York, NY 10023-7692 Tel: (212) 261-1709 Fax: (212) 261-1748 hgu03@nyit.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Over the past twenty years, numerous research papers have described various methodologies for detecting potential errors in healthcare-related terminologies. Whenever a computer-based methodology is used to detect potential errors, be they semantic, lexical or structural errors, human intervention is required to evaluate and make a final determination as to the correctness of the findings and/or the suggested resolutions. Even in well formalized Description Logic (DL) environments it may not be an easy task for a human auditor to determine the "correct" resolution of an error, as it may be domain- and context-sensitive. As has been demonstrated by studies of existing terminologies, there is often more than one "correct" model of a topic of healthcare information. This is especially evident in the Unified Medical Language System (UMLS) [1] where the integration of many source terminologies with common concepts presents a significant challenge.

The Metathesaurus (META) of the UMLS is a repository of over 2.6 million concepts from 161 source terminologies [2]. The integration of the source terminologies is supported by the Semantic Network (SN) [3], which provides a compact abstraction network for the META. Altogether, this large and sophisticated repository is difficult to view and comprehend. Hence, error resolution in the UMLS, even under the most objective conditions, may be complicated by the existence of alternative solutions.

Most prior research is based on the underlying assumption that a domain expert can reliably determine the "best" correction of any given error. Research into inter- and intra-rater reliability of domain experts' performance when auditing terminologies is scarce. In [4], we evaluated the performance of individual auditors against a consensus reference standard. We found that a single auditor is not reliable and that a consensus building process is necessary for producing more reliable results. Only about half of the true errors were detected by individual auditors and only about half of the error reports were correct. Surprisingly, we also found [4] that advanced experience in auditing terminologies and a deeper level of biomedical domain knowledge did not significantly contribute to the quality of the results of individual auditors. These observations regarding individual auditors' performance and the effect of auditing experience (or rather of the lack of it) were unexpected [4]. They are, however, in line with results of related studies by Chute et al. and Fung et al. [5, 6].

The research presented in this paper has the goal to evaluate the impact of aggregating opinions from multiple auditors, using a majority vote, on the reliability of the results. Determining the majority opinion of the auditors is an easy way of aggregating their opinions, as it can be computed automatically without requiring any additional human activity or communication between the auditors. This use of a majority vote has the advantage of simplicity, compared to the consensus-based method of Gu et al. [4]. This study also evaluates the reliability and performance of individual auditors to confirm the results of the pilot study of Gu et al. [4] with a larger sample.

Coincidentally, the UMLS editorial team changed the semantic type assignments of almost all of our sample's concepts in a subsequent release of the UMLS. Thus, we were given the opportunity to use the UMLS editorial team's corrected new release of the UMLS as the gold standard for the performance of our own auditors.

2. Background

The Unified Medical Language System (UMLS) [1] is a large biomedical terminological system. Its large size and complexity make the UMLS prone to errors and make human comprehension very difficult. The Semantic Network (SN) [3, 7] of the UMLS is a compact abstraction network which consists of 133 broad categories called *semantic types*. The semantic types are hierarchically organized in two trees rooted at the semantic types **Entity** and **Event** respectively. Each concept in the UMLS is assigned one or more semantic types.

Considerable research has been carried out on Quality Assurance (QA) of the UMLS. In a 2005 study of UMLS user preferences by Chen et al. [8], users expressed a desire that a significant portion (35%) of a putative UMLS budget be spent on auditing (more than on any other task). Wrong or missing assignments of semantic types from the Semantic Network to concepts of the META were among the top concerns of the study participants. An algorithm was presented by Peng et al. [9] for identifying all redundant semantic type assignments. Such redundant assignments are forbidden by the rules of the UMLS, as described by McCray and Nelson [7].

Semantic techniques complemented by lexical techniques were used by Cimino to detect classification errors [10, 11]. Formal and naïve approaches for identifying and eliminating circular hierarchical relationships in the UMLS have been proposed by Bodenreider [12, 13]. A technique for detecting errors in cycles of three nodes, which minimizes the auditors' efforts, was presented by Halper et al. [14]. Object-oriented models have been employed by Bodenreider to support navigation, maintenance, and auditing of the UMLS [15]. For an extensive review of UMLS auditing and of methods for auditing of medical terminologies in general refer to Zhu et al. [16].

The *extent* of a semantic type is the set of concepts of the UMLS that are assigned this semantic type. Some concepts in an extent are assigned only one semantic type, while others are assigned two or more. Therefore, the extent of a semantic type may contain concepts with different kinds of semantics.

In our previous research [17, 18], we proposed a Refined Semantic Network (RSN) for the UMLS, which can provide a semantically uniform, abstract view for concepts. The RSN promotes the combinations of semantic types into explicit entities of the abstraction network, called *refined semantic types*. The RSN consists of two kinds of refined semantic types: *pure semantic types*, each of which corresponds to one original semantic type from the UMLS Semantic Network, and *intersection semantic types*. An intersection semantic type is based on a combination of two or more semantic types from the UMLS Semantic Network for which there exists a group of concepts assigned exactly this combination of semantic types. This group of UMLS concepts that are assigned the same exact combination of semantic types is then the *extent* of the intersection semantic type. Concepts assigned several semantic types are included in the extent of one and only one intersection semantic type [17].

For example, the two concepts *Cyclic Peptides* and *Peptide hormone* are in the extent of the semantic type **Amino Acid, Peptide, or Protein**. In the Refined Semantic Network, the concept *Cyclic Peptides* is assigned the pure semantic type **Amino Acid, Peptide, or Protein**. In contrast, the concept *Peptide hormone* is assigned the intersection semantic type **Amino Acid, Peptide, or Protein ∩ Hormone,** because *Peptide hormone* is also assigned **Hormone.** (The symbol "∩" is used in mathematics for "intersection").

The concepts of an intersection semantic type are said to have *compound semantics* [17] defined by the combination of the original semantic types. In summary, the extent of each refined semantic type has uniform semantics, because all concepts in this extent have exactly the same semantic types assigned. As an abstraction network, the Refined Semantic Network partitions the META into the disjoint, semantically uniform extents of the refined semantic types. As shown by Chen [19, 20, 21] this semantic uniformity supports effective structural "group auditing" of concepts.

Auditing the UMLS by brute force is a daunting task that is aggravated by the limited availability of trained auditors. Thus, algorithmic approaches for the selection of concepts with a high likelihood of errors are of great utility. For example, we have shown that concepts of intersection semantic types with extents of one to six concepts have a high potential (about 40%) for erroneous semantic type assignments [22]. The error probability declines for larger intersection semantic type extents. Reviewing small intersection semantic type extents allows auditors to focus their attention where it is most effective, since besides a high probability of semantic type assignment errors, other errors are likely to (co)occur for such concepts, as shown in our previous research [4, 19, 23]. However, even with algorithmic tools for the selection of concepts with a high likelihood of errors, discovering and correcting semantic type assignment errors remains a manual process that requires review by auditors with domain knowledge.

In work by Gu et al. [4], we described a process for auditing intersection semantic types with extents of one to six concepts for the purpose of analysis of the performance of the auditors. Four auditors (two domain experts (DEs) and two knowledge engineers (KEs)) first reviewed all concepts independently for semantic type assignment errors. In a second round of processing, the results were aggregated, anonymized and independently reviewed by the two DEs. Each DE affirmed or revised his/her own decision concerning wrong semantic type assignments. For any disagreement on a concept, the DEs consulted with each other and reached a consensus reference standard.

Next, to evaluate the auditors' performance, each KE's first round results were compared to the consensus reference standard. A DE's first round responses were compared only to the second round review by the opposite DE, before reaching the consensus, to avoid experts indirectly judging their own work. Gu et al. found that any individual auditor is unlikely to produce reliable answers and it is necessary to enlist a team of several auditors to achieve reasonable reliability [4]. On average, each individual auditor was able to detect only about half of the true errors, and only about half of the reported errors were indeed errors, as compared to the consensus standard.

The quality of human decision making in vocabulary editing has been questioned in the past [5]. A study about the integration of SNOMED CT into the UMLS by Fung et al. [6] contains similar performance findings regarding synonymy issues as ours [4]. The accuracy of UMLS editors was not significantly better compared to non-editing domain experts and the accuracy was only somewhat better than chance itself.

3. Methods

3.1. Sample data preparation and auditor team

Based on our previous research [4, 22], a sample of all intersection semantic types with extent sizes of six or less was selected for this study from the 2007AC UMLS release. We did not include semantic types from the **Chemical** sub-hierarchy of the SN for two reasons: 1) Intersection semantic types are common for chemical concepts, describing both the structural and functional aspects of a chemical concept, as noted in the UMLS Usage Note of the semantic type **Chemical** [24]; and 2) our auditors are not experts in chemistry.

This sample was presented to the auditors in a "concept data form" (see example in Figure 1). For each concept provided to the auditors, the following information (when available) was supplied: the sequential index of the concept in the sample, the UMLS Concept Unique Identifier (CUI), the preferred name, sources, semantic types, definitions, synonyms, parent concepts (with assigned semantic types), and the child concepts (with assigned semantic types) (Figure 1). To collect responses from the auditors, a "response form" was prepared

(Figure 2). Most choices in the response form are self-explanatory. Our instructions to the auditors made it clear that choice 6 (Ambiguity) refers to homonymous terms, i.e., one term with two possible meanings.

The sample contained concepts with between two and four assigned semantic types. The response form allowed the auditors to mark each individual semantic type assignment as wrong. The semantic type numbers in the response form refer to the semantic types in the order they are given in the concept data form. For example, for the data in Figure 1, Semantic Type 1 is **Finding** and Semantic Type 2 is **Organism Function** (Figure 2). In addition, the response form allows for a “no error” response, and for three other error kinds, the last of which is “other” (Figure 2).

We engaged four auditors, all of whom have experience in medical terminology research and in Quality Assurance of terminologies. JC is an MD and a well-known international expert in terminologies. GE is an MD with formal training in Medical Informatics and an NLM-funded Post Doc with focus on controlled medical terminologies. JX obtained her MD from China and an MS degree in Medical Informatics and is currently working on terminologies at the Lister Hill Center at the National Library of Medicine. YC holds a degree in sports medicine from China and a PhD in the field of QA of terminologies. They appear as co-authors of this study.

3.2. Evaluation

To assess the performance of the individual auditors, their auditing results were compared to a reference standard. As mentioned earlier, the modified semantic type assignments of the UMLS 2008AA release, reflecting the changes made by the UMLS editorial team during the time of our own study, were used as a gold standard. The reliability of the four auditors was quantified using Cronbach’s alpha reliability coefficient [25].

Performance was quantified by accuracy (proportion of all answers that matched the reference standard), recall (proportion of errors indicated in the reference standard that the auditor also reported), precision (proportion of errors reported by the auditor that were also indicated in the reference standard), and F-measure (harmonic mean of recall and precision) [26]. Ninety five percent confidence intervals were calculated for all estimates using the bootstrap method [27]. The statistical significance of differences of the estimates was also calculated using the bootstrap method.

The process of resolution of differences of opinions leading to a consensus among auditors, as used in our pilot study [4] and described above, is labor intensive and time consuming and is further complicated by the need for communication among auditors. In this paper, we suggest a more resource-efficient alternative by interpreting auditors’ results as votes and tallying these votes to reach a majority opinion, whenever possible.

A number of possible situations may arise. Ideally, but rarely, there is a unanimous agreement between all four auditors on one combination of errors. For example, all four auditors may agree that there is no error at all, or all may agree in the response form that Semantic Type 1 *and* Semantic Type 2 are wrong assignments for this concept. An approximation to this unanimous agreement exists when not all auditors agree, but there is a majority of opinions in favor of a specific combination of errors. Therefore, for each concept, we first review the response forms of each of the auditors, which may consist of a combination of multiple errors. Any combination of responses across all four auditors, per concept, which has a majority among the answers, will then be considered the majority opinion.

We differentiate among three possibilities of majority. The first case of four equal markings is referred to as unanimous vote. Next we define the case of three auditors agreeing on one error, and one finding no error at all. The mirror image case is when three auditors agree on "no error" and one auditor finds one error. We call such a situation a strong majority opinion and use the notation "3-to-1 majority" for it. Lastly, if, for example, two auditors detect a certain error, one auditor sees another error and the remaining auditor sees no error at all, we call this a weak majority opinion and use the notation "2-to-1 & 1" for this case.

For example, the concept Symptoms, such as flushing, sleeplessness, headache, lack of concentration, associated with the menopause was found to be assigned the intersection semantic type **Organism Function ∩ Sign or Symptom**. All four auditors marked the Semantic Type 1 as error. In the UMLS 2008AA (gold standard), this concept is assigned only the semantic type **Sign or Symptom,** confirming in this case the unanimous agreement. In another example, the concept *Genital system* was assigned the intersection semantic type **Body Part, Organ, or Organ Component ∩ Body System**. Three auditors marked ST1 as error, that is, the concept should only be assigned **Body System**. One auditor indicated no error. Thus, this is a case of a strong majority opinion (3-to-1). In the UMLS 2008AA, the concept *Genital system* is only assigned the semantic type **Body System**. Hence, the reference standard confirmed the strong majority opinion for this concept. Another example is the concept *Thermal Factors* that was assigned the intersection semantic type **Finding ∩ Natural Phenomenon or Process**. Two auditors marked ST1 as error, one auditor marked ST2 as error, and one marked no error. Thus, this is a case of a weak majority opinion (2-to-1 & 1). In the UMLS 2008AA, this concept is only assigned the semantic type **Natural Phenomenon or Process**. Hence, the reference standard confirms, in this case, the weak majority opinion.

If no combination of errors reaches a majority for a concept, then the four response forms for the same concept are decomposed into their individual errors. At this point, individual errors will be aggregated, as opposed to combinations of errors. For example, if two auditors chose two different errors for a concept and two other auditors chose only one error for that concept, then all six errors will be considered for the majority computation. The most popular individual error(s), if such exist in this case, will then be considered as the majority opinion error.

The decomposed individual errors across auditors are grouped into the same majority categories as the combined answers before: unanimous, strong majority, and weak majority. Overall, the majority opinion was generated by picking the most popular error for each case. Ties were broken by random choice. For example, the concept *Digital Video Recording* is assigned two semantic types, **Human-caused Phenomenon or Process** (ST1) and **Manufactured Object** (ST2). When reviewing this concept, the four auditors had four different answers, 1) no error; 2) ST2 error; 3) Ambiguity; and 4) ST1 error & ST2 error & Add ST. Clearly, there is no combined majority opinion for this concept, as all four answers are different. Thus, we decomposed the answer 4) into three individual answers and thus achieved a partial majority opinion, with two auditors in favor of the ST2 error choice. This response results in a weak majority opinion (2-to-1 & 1). However, there was no change of the semantic type assignment of this concept in the UMLS 2008AA. Thus, this weak majority determination was not confirmed by the gold standard.

4. Results

Based on the UMLS 2007AC release, all 103 intersection semantic types (excluding those from the subhierarchy of **Chemical**) with an extent size of six or less were selected. All their 232 concepts were reviewed by the four auditors.

In Table 1, each individual auditor's performance was evaluated by comparing the audit results with the 2008AA reference standard. For example, Auditor 1 reported 159 concepts with erroneous semantic type assignments. The various average performance measures for three auditors were all below 50%. Only the fourth auditor displayed a performance that was slightly higher than 50%. These results are in line with those of our pilot study [4].

We found that the per-rater reliability for auditors to designate errors was .34 (95% CI .30 to .38), and the reliability of the combined group answer was .67 (95% CI .63 to .71). Thus, four reviewers as a group achieved a reliability close to the target value of .7. This is consistent with our previous results on rater reliability [4].

As described in the Methods Section, we defined three majority categories: unanimous (4-to-0), strong majority (3-to-1) and weak majority (2-to-1 & 1). Table 2 summarizes the performance results for the different majority opinions. Columns 2, 3, and 4 list the numbers of concepts agreed on by the auditors according to the various levels of majority, in decreasing order of majority strength. Columns 5 and 6 provide information regarding the combinations of the stricter two and of all three such majority levels, respectively.

Rows 2 and 4 report the numbers of concepts identified by the auditors without and with errors, respectively. The corresponding rows 3 and 5 indicate how many out of those concepts were confirmed by the gold standard. Row 6 provides the total number of concepts identified by the proper level of majority either as erroneous or not, representing the sum of the corresponding entries in rows 2 and 4. Rows 7 and 8 show the breakdown of these same concepts into erroneous and correct, as reflected by the gold standard.

Using these numbers, four measures of performance, Accuracy, Recall, Precision, and F-measure were calculated (Table 3). The calculation uses numbers from rows 2 to 8 in Table 2. For example, the strong majority (Column 3) shows 30 concepts reported without errors by the auditors, only eight of which were confirmed with no errors by the NLM. The corresponding numbers of erroneous concepts are 58 and 40, respectively. Hence a strong majority opinion exists for 88 concepts. Thus, the accuracy for strong majority is (8+40) / (30 + 58) = 0.55. In the last two columns of Table 3, the average performance of the four individual auditors is reported for the same 178 concepts for which some level of majority was found (Column 7) and for the whole sample of 232 concepts (Column 8). The corresponding averages are very close with a slight advantage for the averages for the 178 concepts.

In Table 3, there is an increase in the various performance measures, with the decrease of the required majority level. That is, Column 4 for the weak majority has the highest performance of the three. The weak majority contributes only to Column 6 (unanimous + strong majority + weak majority), but not to Column 5 (unanimous + strong majority). Thus, there is a higher performance in Column 6 than in Column 5, for all measures.

It is instructive to compare the performance of the "all majorities" case (Column 6: unanimous + strong + weak) to that of the average performance of all four auditors for those 178 concepts (Column 7 in Table 3, taken from Table 1). In Table 3, there is an improvement of 0.53 – 0.42 = 0.11 in accuracy from Column 7 to Column 6, which corresponds to a 26% relative improvement (0.11 / 0.42) * 100 = 26%. For recall, the improvement is 0.62 – 0.50 = 0.12, which corresponds to a relative improvement of (0.12 / 0.5) * 100 = 24%.

For precision the improvement is 0.64 – 0.46 = 0.18, which corresponds to a 39% relative improvement, as (0.18 / 0.46) * 100 = 39%. The increase of the F measure is 0.63 – 0.48 =

0.15, which corresponds to a relative increase of (0.15 / 0.48) * 100 = 31%, the median of the increases for recall and precision.

For the evaluation of data in Column 8 of Table 3, ties were broken randomly to obtain a final determination for all 232 concepts. Accuracy, recall, precision, and F-measure were each statistically significantly higher for the “all majorities” vote (Column 6), than for the average of the auditors for the 232 concepts (Column 8).

5. Discussion

5.1 Interpretation

The current study confirms the finding of our pilot study [4] that a single auditor cannot reliably detect, determine, and correct semantic type assignment errors for complex concepts in the UMLS, i.e., for concepts assigned combinations of semantic types, with a substantially larger sample than in [4]. Furthermore, the current study also confirms that for the various performance measures, the average performance of the auditors is lower than 0.50. That is, for example, that an auditor finds fewer than half of the errors, and the corrections suggested for half of the errors by one auditor are not appropriate. Our motivation for performing this study with a larger sample (232 concepts) versus the smaller sample of 70 concepts used in [4] was that on a superficial level the prior results on the low reliability of single auditors were surprising. One would assume that an experienced auditor would find the majority of the existing errors, and would suggest the correct changes for a majority of the errors detected. However, the pilot results [4] were confirmed in the current larger study.

As mentioned in the Background section, similar results regarding the performance of editing of terminologies were found in the work on SNOMED integration into the UMLS by Fung et al. [6] and in the context of the Mayo Clinic’s clinical terminology development by Chute et al. [5].

This study included only concepts that have multiple semantic type assignments, which makes these concepts more *complex* than concepts with a single semantic type. Such concepts are more likely to have erroneous semantic type assignments. In our previous study [22], the concepts assigned combinations of semantic types of small extents (extents with between one and six concepts) had double the probability of errors, compared to a reference set of concepts from intersections of larger extents. We did not study the percentage of errors of concepts assigned only one semantic type, which we expect to be low. Thus, one cannot generalize this study to deduce similar low performance of auditors for concepts with single semantic types. Similarly, this conclusion cannot be generalized to *editors* integrating a new source into or updating an old source in the UMLS. Hence, the findings of our study should not be interpreted as a general statement that low performance of editors, when assigning semantic types to concepts, is to be expected.

The importance of this study is not in the number of errors found, which is relatively small in proportion to the whole UMLS, but in its implications for auditing per se. The findings of this study are important for *auditing*, since the complex META concepts with combinations of semantic types provide an example of good candidates for auditing due to their high likelihood of errors. Scarce (human) auditing resources should be expended on concepts for which relatively more errors are expected.

Accepting that in some circumstances several auditors might need to perform the same auditing task to produce a reliable joint result, a method is needed to aggregate individual results. In our study, a unanimous decision was reached by the auditors for fewer than 18%

of the concepts. Hence, for more than 82% of the complex concepts, a process was required to resolve differences between the auditors. Thus, assigning several auditors to the same task requires further resources to obtain a resolution for the many cases without unanimous agreement. Among the possible solutions in such cases are consensus building and voting. The aggregated opinions reported in our pilot project [4] were obtained by consensus between the two experts based on the opinions of all four auditors. The aggregated opinions were submitted as an audit report to the NLM.

An analysis of the consensus method uncovered several issues [4]. Finding a consensus is time consuming. When striving for a consensus result, the two expert auditors had to weigh their own original answers against those of other auditors, a process in which they cannot be assumed to be unbiased. Additional anecdotal evidence obtained independently from both these auditors suggests that reaching a consensus was not only related to scientific issues and arguments but also to some “social give and take” interactions. Some of these interactions resulted in “brand new” resolutions, which were not based on any of the four original solutions in front of them. As a result of these issues, the consensus arrived at did not necessarily constitute an objective aggregate solution.

We note that the same consensus opinion which was used in [4] for preparing the aggregated auditing report (submitted to the NLM) was also used as a reference standard for the evaluation of the performance of the individual auditors. Accurate reference standards rarely exist to support evaluations [5, 28]. In their absence, pooled human expert opinions are used [28] and in our pilot study [4] we followed this practice.

In retrospect, we acknowledge that the consensus opinion used as a reference standard in [4] suffers from the same deficiencies we listed above for its role as an aggregated opinion for the group of auditors. Nevertheless, the conclusions of the pilot study [4] regarding the reliability and performance of individual auditors were confirmed by the larger current study using an objective reference standard.

We were presented with an excellent opportunity by the UMLS editorial team’s extensive efforts to correct semantic type assignments, providing us with our gold standard. Using this objective reference standard avoids personal biases of experts and the influence of “social give and take” communication. Furthermore, this reference standard can be used to evaluate not just the individual auditors but also the aggregated opinions found in the current study.

The values of the four measures (accuracy, etc.) increase, as we move from the strongest agreement level of unanimous agreement to strong majority and on to weak majority. Table 3 also shows the performance of two levels of cumulative agreement, the combination of unanimous agreement and strong majority (Column 5), and the combination of unanimous, strong and weak majority (Column 6). A similar increase in the four measures is observed with a decrease of the level of cumulative agreement.

The implied conclusion is that even a weak majority of a determination should suffice for accepting it as a recommended aggregate opinion for the group of auditors. Naturally, the less demanding levels of agreement cover more cases.

These results are encouraging for the suitability of the majority mechanism to offer an effective method for aggregating the opinions of auditors. For some concepts a majority could not be reached. We decided to make random choices for such cases.

5.2. Future research

Our study considered one kind of complex META concepts, those assigned combinations of semantic types. An interesting research problem is whether similar reliability and performance measures of auditors will be observable for other kinds of complex concepts of the UMLS or its source terminologies. If so, will the utilization of several auditors and the use of their majority opinions improve the resulting reliability and performance?

We noted at the end of the Methods Section, "if no combination of errors reaches a majority… then the four response forms for the same concept are decomposed into their individual errors." This decomposition raises the possibility of a simpler analysis methodology, where the unit of analysis is not a concept but an assignment of one ST to a concept. We did not use this methodology in this study, because our auditing concentrated on each concept as a unit. However, this alternative analysis methodology has the advantage of greater simplicity. Future research should compare these two methodologies to determine whether there is a negative tradeoff for this simplicity.

For an example of complex META concepts, one may consider pairs of concepts with a hierarchical relationship, e.g., parent-of, connecting them, versus the hierarchical relationships in the Semantic Network between the semantic types assigned to the concepts of this pair. In [29] we considered three different kinds of configurations, consistent configuration, lack of ancestry and semantic type inversion.

The child concept Y in a semantic inversion configuration is complex due to having a parent X with a semantic type that is more specific than the semantic type of Y itself. Out of 100 randomly selected parent-child pairs with semantic type inversion that were analyzed, a domain expert determined that 84 contained errors. Thus, the error percentage for a sample with complex concepts was considerably higher than for consistent configurations, for which a sample of 100 concepts had only a single error. For similar, related conditions for complex concepts, see also [19, 20, 30].

In the context of SNOMED auditing [31], an example of complex concepts was defined by those concepts appearing in the overlap of two or more *partial areas*. Partial areas are groups of concepts of similar structure and semantics in the *partial area taxonomy* abstraction network of a SNOMED hierarchy, as used in a study by Wang et al. [32]. In that study, a high likelihood of errors for the concepts in the overlapping partial areas was observed, compared to other concepts. It would be interesting to investigate whether for such cases of complex concepts there is also a phenomenon of low reliability and low performance of a single auditor. Does the performance improve when similar majority rules are applied to obtain an aggregate opinion of several auditors?

Another research issue would be to determine the minimum number of auditors needed to obtain a desired level of reliability and performance for the aggregate opinions. The availability of domain expert auditors is very limited, which dictates effective (minimal) use of such human resources, but one still wants to guarantee acceptable levels of reliability and performance when auditing complex concepts.

The minimum number of auditors enabling the determination of a simple majority is three. The finding of our study that better performance is obtained with an aggregated opinion, even when a weak majority is used as a criterion, suggests that in case of three auditors, a majority of two against one will suffice to obtain an acceptable performance. More studies of the performance of auditors working with complex concepts are needed to determine whether, indeed, three auditors are sufficient for this purpose.

6. Conclusions

When auditing complex concepts in the UMLS, a single auditor is not reliable and the average performance of auditors is quite low. Instead, multiple auditors should be assigned to each such task in order to achieve an acceptable performance level. These findings should significantly affect the allocation of resources to auditing tasks, resulting in the assignment of several auditors to auditing the same complex concepts. Since such human resources are scarce, a resource-efficient aggregation method is needed for auditing to be productive. In this study, significantly better performance was demonstrated by a vote-based method than for individual auditors. Even a simple majority protocol proved to perform better than the average performance of individual auditors. Such a majority opinion can be derived automatically, is objective, does not require time for communication among auditors, and does not depend on social interactions among them.

More work is need to establish whether these conclusions can be generalized to auditing tasks of complex concepts other than those defined by combinations of semantic type assignments and beyond the UMLS to cases where no semantic types exist. For resource-efficient and productive auditing, we propose further research into whether auditing by three auditors, using a simple majority vote, will achieve acceptable performance levels.

Acknowledgments

This work was partially supported by the United States National Library of Medicine under grant R01LM008445-01A2.

References

[1]. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: An Informatics Research Collaboration. Journal of the American Medical Informatics Association. 1998; 5(1):1–11. [PubMed: 9452981]

[2]. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. Apr; 1993 81(2):217–22. [PubMed: 8472007]

[3]. McCray AT. An Upper Level Ontology for the Biomedical Domain. Comp Funct Genom. 2003; 4:80–84.

[4]. Gu H, Hripcsak G, Chen Y, Morrey CP, Elhanan G, Cimino JJ, Geller J, Perl Y. Evaluation of Auditing process of UMLS Semantic Type Assignment. Proceedings of 2007 AMIA annual symposium. 2007:294–8.

[5]. Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. Proc AMIA Symp. 1998:795–9. [PubMed: 9929328]

[6]. Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inform Assoc. Jul-Aug;2005 12(4):486–94. [PubMed: 15802483]

[7]. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med. 1995; 34:193–201. [PubMed: 9082131]

[8]. Chen Y, Perl Y, Geller J, Cimino JJ. Analysis of a study of the users, uses and future agenda of the UMLS. Journal of the American Medical Informatics Association. 2007; 14(2):221–31. [PubMed: 17213497]

[9]. Peng, Y.; Halper, M.; Perl, Y.; Geller, J. Auditing the UMLS for redundant classifications. Proceedings of 2002 AMIA annual symposium; San Antonio, TX. Nov. 2002 p. 612-6.

[10]. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. Journal of the American Medical Informatics Association. 1998; 5:41–51. [PubMed: 9452984]

[11]. Cimino JJ. Overhage JM. Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS Metathesaurus. Proceedings of 2001 AMIA annual symposium. 2001:120–4.

[12]. Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proceedings of AMIA symposium. 2001:57–61.

[13]. Mougin F, Bodenreider O. Approaches to eliminating cycles in the UMLS metathesaurus: naive vs. formal. Proceedings of 2005 AMIA annual symposium. 2005:550–4.

[14]. Halper M, Morrey CP, Chen Y, Elhanan G, Hripcsak G, Perl Y. Auditing Hierarchical Cycles to Locate Other Inconsistencies in the UMLS. Proc AMIA Annual Symp. 2011:529–36.

[15]. Bodenreider, O. An object-oriented model for representing semantic locality in the UMLS. Proceedings of Medinfo 2001; London, UK. Sep. 2001 p. 161-5.

[16]. Zhu X, Fan J, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. Journal of Biomedical Informatics. 2009; 42(3): 413–425. [PubMed: 19285571]

[17]. Gu H, Perl Y, Geller J, Halper M, Liu L, Cimino JJ. Representing the UMLS as an OODB: Modeling Issues and Advantages. Journal of the American Medical Informatics Association. 2000; 7(1):66–80. [PubMed: 10641964] Haux, R.; Kulikowski, C., editors. Yearbook of Medical Informatics. International Medical Informatics Association; Rotterdam: 2001. p. 271-285.Selected for reprint in

[18]. Geller J, Gu H, Perl Y, Halper M. Semantic Refinement and Error Correction in Large Terminological Knowledge bases. Data & Knowledge Engineering. 2003; 45(1):1–32.

[19]. Chen Y, Gu H, Perl Y, Geller J, Halper M. Structural Group Auditing of a UMLS Semantic Type's Extent. Journal of Biomedical Informatics. 2009; 42(1):41–52. [PubMed: 18619563]

[20]. Chen Y, Gu H, Perl Y, Geller J. Structural Group-Based Auditing of Missing Hierarchical Relationships in UMLS. Journal of Biomedical Informatics. 2009; 42(3):452–467. [PubMed: 18824248]

[21]. Chen Y, Gu H, Perl Y, Halper M, Xu J. Expanding the Extent of a UMLS Semantic Type via Group Neighborhood Auditing. Journal of the American Medical Informatics Association. 2009; 16(5):746–57. [PubMed: 19567802]

[22]. Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artif Intell Med. May; 2004 31(1):29–44. [PubMed: 15182845]

[23]. Morrey CP, Geller J, Halper M, Perl Y. The Neighborhood Auditing Tool: A hybrid interface for auditing the UMLS. Journal of Biomedical Informatics. 2009; 42(3):468–489. [PubMed: 19475725]

[24]. [accessed 2/15/1012] http://web.njit.edu/~kh8/UMLS/SemanticNetwork3.php?2

[25]. Dunn, G. Design and analysis of reliability studies. Oxford University Press; New York: 1989.

[26]. Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005; 12:296–8. [PubMed: 15684123]

[27]. Efron, B.; Tibshirani, R. An Introduction to the Bootstrap. Chapman & Hall; New York: 1993. R.

[28]. Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc. Jan-Feb;2002 9(1):1–15. [PubMed: 11751799]

[29]. Geller, J.; Morrey, CP.; Xu, J.; Halper, M.; Elhanan, G.; Hripcsak, G. Comparing Inconsistent Relationship Configurations Indicating UMLS Errors. Proc 2009 AMIA Annual Symp; San Francisco, CA. 2009. p. 193-197.

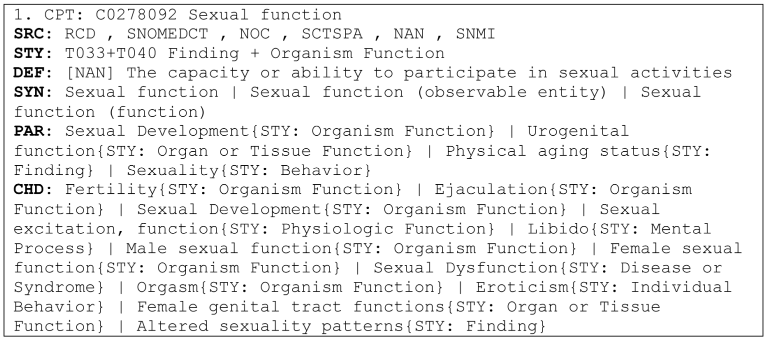
[30]. Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. J Biomed Inform. Dec; 2003 36(6):450–61. [PubMed: 14759818]

[31]. Wang Y, Halper M, Wei D, Gu H, Perl Y, Xu J, Elhanan G, Chen Y, Spackman KA, Case JT, Hripcsak G. Auditing Complex Concepts of SNOMED using a Refined Hierar-chical Abstraction Network. J Biomed Inform. Feb; 2012 45(1):1–14. [PubMed: 21907827]

[32]. Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. J Biomed Inform. 2007; 40(5):561–81. [PubMed: 17276736]

Highlights

- The performance of auditors on auditing the semantic type assignments of complex UMLS concepts is studied.
- The results indicate that individual auditors are not reliable.
- The reliability of the automatically computed majority opinion from multiple auditors is evaluated.
- The results indicate the majority opinion is reliable.
- The performance of majority opinion is statistically significantly higher than the average performance of individual auditors.



1. CPT: C0278092 Sexual function

SRC: RCD , SNOMEDCT , NOC , SCTSPA , NAN , SNMI

STY: T033+T040 Finding + Organism Function

DEF: [NAN] The capacity or ability to participate in sexual activities

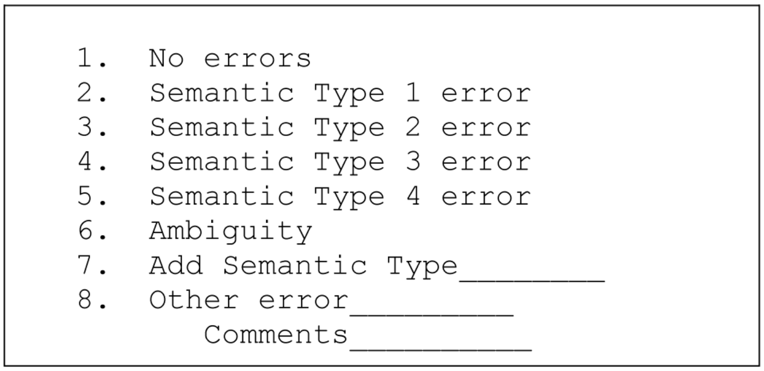
SYN: Sexual function | Sexual function (observable entity) | Sexual function (function)

PAR: Sexual Development{STY: Organism Function} | Urogenital function{STY: Organ or Tissue Function} | Physical aging status{STY: Finding} | Sexuality{STY: Behavior}

CHD: Fertility{STY: Organism Function} | Ejaculation{STY: Organism Function} | Sexual Development{STY: Organism Function} | Sexual excitation, function{STY: Physiologic Function} | Libido{STY: Mental Process} | Male sexual function{STY: Organism Function} | Female sexual function{STY: Organism Function} | Sexual Dysfunction{STY: Disease or Syndrome} | Orgasm{STY: Organism Function} | Eroticism{STY: Individual Behavior} | Female genital tract functions{STY: Organ or Tissue Function} | Altered sexuality patterns{STY: Finding}

Figure 1.

Concept data form



1. No errors
2. Semantic Type 1 error
3. Semantic Type 2 error
4. Semantic Type 3 error
5. Semantic Type 4 error
6. Ambiguity
7. Add Semantic Type________
8. Other error_________

Comments__________

Figure 2.
Response form format for marking error(s)

Table 1

Performance of auditors measured by the 2008AA reference standard

Auditors	# of Erroneous Concepts	Accuracy	Recall	Precision	F
Auditor 1	159	0.31	0.39	0.42	0.40
Auditor 2	164	0.34	0.41	0.43	0.42
Auditor 3	164	0.37	0.45	0.47	0.46
Auditor 4	217	0.47	0.64	0.51	0.57
Average	176	0.37	0.47	0.46	0.46

Table 2

Performance for various majority opinion levels measured by the 2008AA reference standard

Experimental Condition		**Unanimous**	**Strong majority**	**Weak majority**	**Unanimous + Strong Majoriyt**	**Unanimous + Strong Majority + Weak Majority**
# of concepts identified without error by auditors		9	30	10	39	49
# of the concepts in row above confirmed without error by NLM		0	8	3	8	11
# of concepts identified with error by auditors		31	58	40	89	129
# of the concepts in row above confirmed with error by NLM		14	40	29	54	83
Total # of concepts with a majority		40	88	50	128	178
2008AA reference standard	**without error**	13	22	10	35	45
	with error	27	66	40	93	133

Table 3

Accuracy, Recall, Precision and F-measure for various majority opinion levels measured by the 2008AA reference standard

Performance Measure	**Unanimous**	**Strong majority**	**Weak majority**	**Unanimous + Strong Majority**	**Unanimous + Strong Majority + Weak majority**	**Average of all auditors on 178 concepts**	**Average of all auditors on 232 concepts**
Accuracy	0.35	0.55	0.64	0.48	0.53	0.42	0.37
Recall	0.52	0.61	0.72	0.58	0.62	0.50	0.47
Precision	0.45	0.69	0.72	0.61	0.64	0.46	0.46
F	0.48	0.65	0.72	0.59	0.63	0.48	0.46