

Auditing Concept Categorizations in the UMLS

Huanying (Helen) Gu¹, Yehoshua Perl², Gai Elhanan³, Hua Min², Li Zhang², Yi Peng²

¹Dept. of Health Informatics, University of Medicine & Dentistry of NJ, Newark, NJ 07107

Tel: (973)972-0996 Fax: (973)972-1054 Email: guhy@umdnj.edu

²CS Dept., New Jersey Institute of Technology, Newark, NJ 07102

³Info-X Inc, Northvale, NJ 07647

Abstract

The Unified Medical Language System integrates about 880,000 concepts from 100 biomedical terminologies. Each concept is categorized to at least one semantic type of the Semantic Network. During the integration, it is unavoidable that some categorization errors and inconsistencies will be introduced. In this paper, we present an auditing technique to find such errors and inconsistencies. Our technique is based on an expert reviewing the pure intersections of meta-semantic types of a metaschema, a compact abstract view of the UMLS Semantic Network. We use a divide and conquer approach, handling differently small pure intersections and medium to large pure intersections. By using this approach, we limit the number of concepts reviewed, for which we expect a high percentage of errors. We reviewed all concepts in 657 pure intersections containing one to 10 concepts. Various kinds of errors are identified and the analysis of the results are presented in the paper. Also, we checked the pure intersections containing more than 10 concepts for their semantic soundness, where the semantically suspicious pure intersections are presented in the paper and their concepts are reviewed.

Keywords: Medical terminology, UMLS, auditing, metaschema, semantic network, semantic type, pure intersection, categorization

1 Introduction

The Unified Medical Language System (UMLS) [10, 11, 12], designed by the National Library of Medicine (NLM), combines many well established biomedical information sources

in a unified knowledge representation system. It enables electronic access to a very large compendium of biomedical knowledge expressed as terminologies. The UMLS plays a role in overcoming terminological differences in the integration of healthcare information systems.

The UMLS [21] contains three Knowledge Sources: the Metathesaurus (META) [19, 20], the Semantic Network, and the Specialist lexicon. The META of the 2001 version, used in this research, integrates 99 clinical and biomedical terminologies. It contains knowledge of about 800,000 biomedical concepts and relationships among them. The scope and complexity of the META can make it difficult for users to understand and visualize. The Semantic Network [13, 14, 16] of the UMLS is a network of semantic types. Each concept in the META is associated with one or more of the 135 semantic types (134 in the 2001 version) of the Semantic Network. Assigning semantic types to concepts involves algorithmic procedures as well as extensive review by domain expert based on two assumptions: 1) each concept is assigned to the most specific semantic type available; 2) semantic types are assigned according to the meaning or meanings that the concept has in its source vocabulary [14]. The Semantic Network is a high-level structure that contains information about the semantic types and relationships among them. The Semantic Network provides an abstraction that helps organize the large number of concepts in the biomedical domain.

Due to the way the UMLS was created, it is unavoidable that some categorization errors and inconsistencies were introduced. This is due to the integration of many terminological sources that are not necessarily consistent with one another. Also, since categorization was done by many experts with different backgrounds, views, and priorities, it is possible for the same concept to be categorized in different ways. Assuring consistency and correctness of concept categorization in the UMLS is an ongoing auditing challenge for the NLM. As a matter of fact, an error in the categorization of a concept may expose an error in the definition, naming or relationships of the concept, which is more critical for users of the UMLS.

However, due to the huge size and complexity of the META, such a comprehensive audit is an overwhelming task. Different approaches for auditing the META appear in [1, 2, 5, 6, 8, 9, 17]. For example, in [2], Bodenreider described techniques to support the maintenance of the META by constructing object-oriented models of the UMLS. Cimino [5]

used semantic methods to uncover categorization errors in the UMLS. Hole developed a new method to find missed synonymy in the META [9].

In our previous research [4], we have introduced a technique for partitioning the Semantic Network. This technique groups closely related semantic types of the Semantic Network into semantic-type collections represented as meta-semantic types. We called the network consisting of meta-semantic types connected by hierarchical and semantic relationships a metaschema [18]. This metaschema provides a higher-level abstract view of the Semantic Network.

In this paper, we describe our auditing technique based on the cohesive metaschema of the UMLS. For details about this metaschema see the next section.

Since a concept may be assigned to several semantic types, it may also be assigned to several meta-semantic types. However, it is more likely that a concept will be erroneously assigned to semantic types of different meta-semantic types than to semantic types of the same meta-semantic type because of larger semantic distance. This observation leads to the idea of concentrating the auditing effort on concepts that are assigned to different meta-semantic types. The idea is that such concepts are more prone to be erroneous or inconsistent. Thus, we first identify all concepts of intersections of two or more meta-semantic types. After that, we refine each one of the intersections into multiple pure intersections. A domain expert reviews each pure intersection containing a small number of concepts of similar semantics, to find any categorization errors. Different kinds of categorization errors are exposed in the review. Furthermore, the combination of intersecting semantic types of each pure intersection containing medium and large number of concepts will be reviewed to verify that it is semantically sound. The concepts of the pure intersections, which are not semantically sound, are reviewed by a domain expert. This auditing technique is designed to minimize the effort and maximize the likelihood of finding errors.

2 Background: Metaschema of the Semantic Network of the UMLS

In the Semantic Network, the semantic types are nodes and the relationships between them are the links. The semantic types are arranged in a strict hierarchy through hierarchical IS-A links. In addition, the semantic types are related through non-hierarchical semantic relationships. The semantic relationships defined for a semantic type are generally inherited via the IS-A links by all the children of this semantic type, unless the inheritance is explicitly blocked.

The process of generating a metaschema of the Semantic Network begins with partitioning the Semantic Network. Since every semantic type has a set of relationships that are either defined directly or inherited from its parent, we can partition the Semantic Network based on the distribution of the relationships among the semantic types. All semantic types exhibiting the exact same set of relationships are grouped together [4]. The set of relationships that is shared by all semantic types in a group is the *structure* of those semantic types and their group. Such a group is called a *structural group*. Every semantic type is assigned to one and only one structural group. Therefore, all structural groups are pairwise disjoint and their union yields all the semantic types of the Semantic Network. The partition of the Semantic Network into structural groups is called the *structural partition*.

However, in the structural partition of the Semantic Network, there are cases of structural groups with multiple roots. For an effective partitioning of the Semantic Network, a group should not just be structurally uniform, but also semantically uniform. For this, a group needs to have a unique root i.e., one semantic type that all other semantic types in the group are descendants of, that is, a specialization of, the unique root. In order to obtain semantically uniform groups, we develop rules to transform those structural groups with multiple roots into (cohesive) semantic type groups, each with a unique root. For a detailed explanation, see [18]. Another problem with the structural partition is its large number of leaf singletons. Note that a singleton is a group of one semantic type. A semantic type without children is called a leaf. To avoid it, we developed rules [18] to add a leaf singleton to its parent's structural group. After applying our rules to the structural partition, the

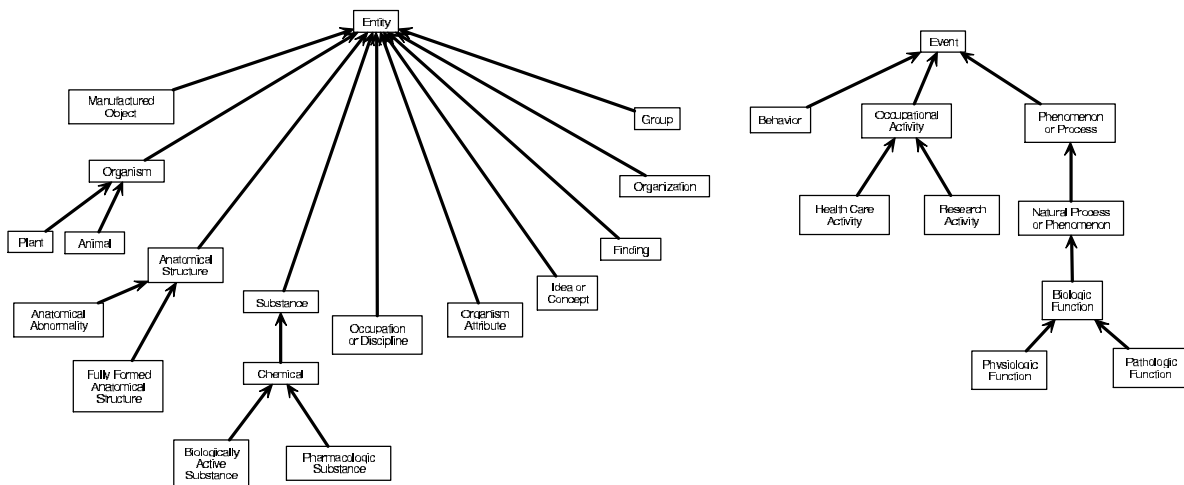


Figure 1: The Cohesive Metaschema Hierarchy of the UMLS Semantic Network

cohesive partition is obtained. It consists of groups, called *semantic-type collections*, with unique roots. Some of these collections are structural groups; others are semi-structural groups (see [18] for details).

From the cohesive partition, we generate the cohesive metaschema of the UMLS. Each semantic-type collection is represented in the metashema as a node, called a *meta-semantic type*. The meta-semantic type is named after the unique root of the corresponding semantic-type collection. The meta-semantic types in the metaschema are connected by two kinds of links, the *meta-child-of* hierarchical relationships and the semantic *meta-relationships*. The hierarchical meta-child-of relationships are induced from the IS-A relationships in the Semantic Network. The meta-relationships are induced from the semantic relationships in the Semantic Network. The meta-child-of hierarchy in the metaschema supports the inheritance of the meta-relationships among meta-semantic types. The cohesive metaschema of the Semantic Network consists of 28 meta-semantic types (see Figure 1 for the metashema hierarchy). It provides an abstract compact view of the Semantic Network.

3 Auditing Method

In the META, each concept is assigned to one or more semantic types, each of which in turn is associated with one meta-semantic type. For example, the concept RETROVIRUS

VECTOR LN¹ is assigned to the three semantic types; *Virus*; *Pharmacologic Substance*; and *Indicator, Reagent, or Diagnostic Aid*, which are partitioned in the metaschema into three meta-semantic types **Organism**, **Pharmacologic Substance**, and **Chemical**, respectively. Therefore, the concept RETROVIRUS VECTOR LN is associated with those three meta-semantic types.

However, a concept that is assigned to two or more semantic types is not necessarily associated with two or more meta-semantic types since multiple semantic types may be grouped into one meta-semantic type. For example, the concepts PULSUS BIGEMINUS, HYPOXEMIA, DNA MARKER, GENETIC MARKERS, ANOXEMIA, CHROMOSOME MARKERS, and RNA MARKER are assigned to the semantic types *Laboratory or Test Result* and *Sign or Symptom*, which are grouped together into one meta-semantic type **Finding** in the Metaschema. Thus, all those 7 concepts are associated with only one meta-semantic type **Finding**.

Our first hypothesis is that the probability of a concept to be erroneously assigned to multiple semantic types from different meta-semantic types is higher than to be erroneously assigned to multiple semantic types of the same meta-semantic type. The reason is that closely related semantic types are grouped together into one meta-semantic type in our metaschema. The chance of a concept to be assigned correctly to 2 closely related semantic types is higher than to be assigned correctly to two semantic types that are not closely related, as is expected for two semantic types of two different meta-semantic types.

This hypothesis leads to the idea of concentrating our auditing effort on concepts that are associated with multiple meta-semantic types, since such concepts may be more error-prone than concepts with a single semantic type.

In order to precisely describe our auditing method, we first need to have a few definitions. **Intersection of semantic types**: An intersection of two or more semantic types is a non-empty set of concepts that are assigned to each of these semantic types and only to them.

Figure 2 shows the intersection of the semantic types *C* and *D*. The concept A is assigned to only two semantic types *C* and *D*. So A is in the intersection of *C* and *D*, denoted $A \in C \cap D$. The notation of an intersection uses the mathematical intersection symbol \cap .

¹”Small cap” will be used to denote concepts; italics will be used for semantic types; bold font will be used for meta-semantic types.

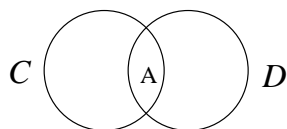


Figure 2: Example of the Intersection of Semantic Types

For example, RETROVIRUS VECTOR LN \in *Virus* \cap *Pharmacologic Substance* \cap *Indicator, Reagent, or Diagnostic Aid*.

According to the definition of intersection of semantic types, each concept in the META will be in at most one intersection of semantic types. Thus, all intersections of semantic types are disjoint. For example, the concept RETROVIRUS VECTOR LN in the previous example will not be a concept in any one of the following three binary intersections: *Virus* \cap *Pharmacologic Substance*; *Pharmacologic Substance* \cap *Indicator, Reagent, or Diagnostic Aid*; or *Virus* \cap *Indicator, Reagent, or Diagnostic Aid*. The reason is that the concept RETROVIRUS VECTOR LN is assigned to three of these semantic types, not two. Thus it can only be a concept in the intersection of all those three semantic types.

Meta-semantic type association: A concept is called associated with a meta-semantic type if it is assigned to at least one of the semantic types in this meta-semantic type.

For example, the semantic types *C* and *D* in Figure 2 and the semantic types *E* and *F* are all grouped into one meta-semantic type **A** (see Figure 3 (a)). As mentioned before, the concept A is assigned to only two semantic types *C* and *D*. The concept B is assigned only to the semantic type *E* and the concept C is assigned only to the semantic type *F*. Thus, all the three concepts A, B, and, C are associated with the meta-semantic type **A** (see Figure 3 (b)).

However, since each concept can be assigned to more than one semantic type, it may also be associated with more than one meta-semantic type if the assigned semantic types are partitioned into different meta-semantic types. For example, the concept ENZYMES is assigned to two semantic types *Organic Chemical* and *Enzyme*. The semantic types *Organic Chemical* and *Enzyme* are partitioned into two meta-semantic types **Chemical** and **Biologically Active Substance** respectively. Therefore, the concept ENZYMES is associated with those two meta-semantic types.

Intersection of meta-semantic types: An intersection of two or more meta-semantic

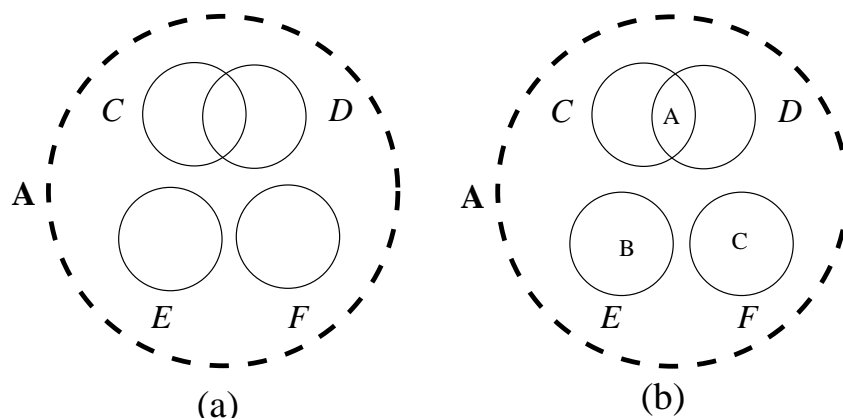


Figure 3: Example of Meta-Semantic Type Association (Semantic types are represented by circles and meta-semantic types are represented by bold dash circles)

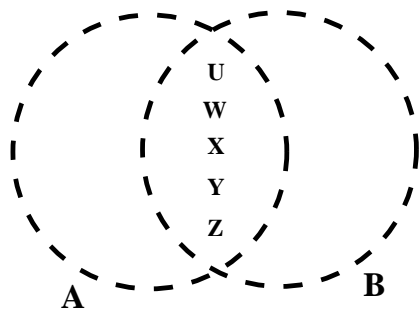


Figure 4: Example of the Intersection of Meta-Semantic Types

types is a non-empty set of concepts that are associated with each of these meta-semantic types and only with them.

Figure 4 shows the intersection of meta-semantic types **A** and **B**. The concepts *u*, *w*, *x*, *y*, and *z* are all associated with the meta-semantic types **A** and **B**. Thus, all of them are in the intersection of the meta-semantic types **A** and **B**.

We use the same notation for the intersection of meta-semantic types. For example, $\text{ENZYMES} \in \text{Chemical} \cap \text{Biologically Active Substance}$.

As with the intersection of semantic types, each concept can only be in at most one intersection of meta-semantic types. Therefore, all the intersections of meta-semantic types are disjoint.

A concept in the intersection of meta-semantic types must be in an intersection of semantic types. However, a concept in one intersection of semantic types may not necessarily be in any intersection of meta-semantic types. The reason is that the intersected semantic types

may be grouped into the same meta-semantic type. In a previous example, the seven concepts PULSUS BIGEMINUS, HYPOXEMIA, DNA MARKER, GENETIC MARKERS, ANOXEMIA, CHROMOSOME MARKERS, and RNA MARKER are in the intersection of semantic types *Laboratory or Test Result* \cap *Sign or Symptom*. But both semantic types *Laboratory or Test Result* and *Sign or Symptom* are grouped together into one meta-semantic type **Finding** in the Metaschema. In this case, all seven concepts are just in the meta-semantic type **Finding**, not in any intersections of meta-semantic types. Therefore, not all intersections of semantic types are intersections of meta-semantic types. Thus, the effort to review the intersections of meta-semantic types should be smaller than the effort of reviewing the intersections of all semantic types.

Note that concepts in one intersection of meta-semantic types are not necessarily in one intersection of semantic types. An intersection of meta-semantic types may consist of several intersections of semantic types. When the domain expert examines the concepts in one such intersection of meta-semantic types, it is difficult for the expert to analyze all those concepts together since they belong to different intersections of semantic types and thus have different compound semantics [8]. This makes the review job more complicated. In order to make the auditing job simpler and more efficient, we need to partition each one of the intersections of meta-semantic types into multiple pure intersections, defined as follows.

Pure intersection of meta-semantic types: A pure intersection of meta-semantic types is a subset of the intersection of the corresponding meta-semantic types, containing all concepts in one intersection of semantic types.

According to the definition, all pure intersections of one intersection of meta-semantic types are disjoint and their union yields the intersection of the meta-semantic types. In other words, the collection of all pure intersections of an intersection of meta-semantic types is a partition of the intersection of meta-semantic types.

The graphical representation of Figures 2, 3, and 4 uses the standard Venn Diagram [7]. However, the graphical representation of the pure intersections is not straightforward as the intersections of semantic types and the intersections of meta-semantic types. The reason is that each pure intersection involves three kinds of entities: meta-semantic types, semantic types, and, concepts. From the semantic type point of view, all meta-semantic types

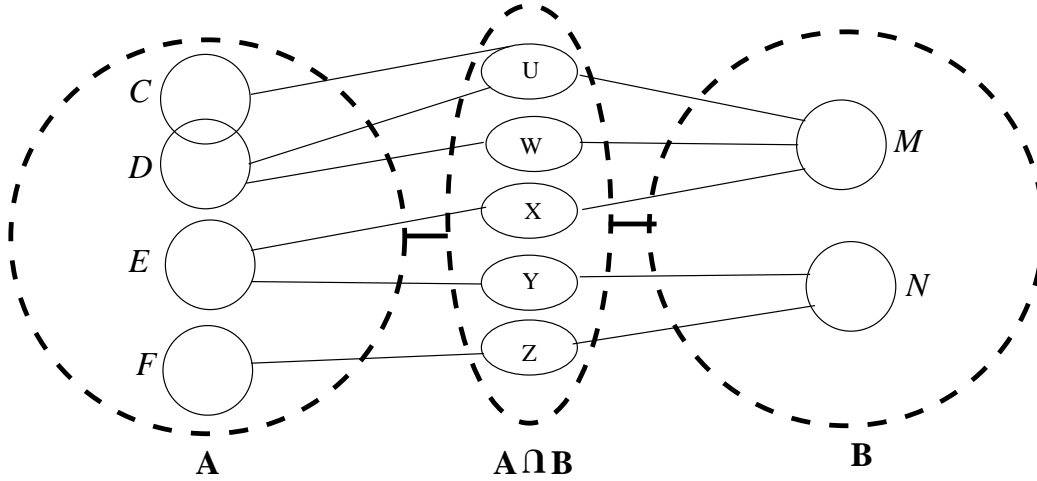


Figure 5: Example of Pure Intersections (The bold dash oval represents the intersection of meta-semantic types and the ovals inside the bold dash oval represent pure intersections)

should be disjoint since each semantic type is in only one meta-semantic type. But from the concept point of view, some meta-semantic types are not disjoint because some concepts can be associated with multiple meta-semantic types. Therefore, in order to represent all these details, we use another way to represent the pure intersections. In Figure 5, we show five pure intersections of the intersection of meta-semantic types $\mathbf{A} \cap \mathbf{B}$ of Figure 4. The intersection of meta-semantic types appears as the bold dash oval that is connected to the meta-semantic types with bold lines, while the pure intersections appear as ovals inside the intersection of meta-semantic types and are connected to the semantic types drawn as circles inside their meta-semantic types.

Figure 5 shows that the concepts u , w , x , y , and, z are all in the same intersection of meta-semantic types $\mathbf{A} \cap \mathbf{B}$. The meta-semantic type \mathbf{A} consists of four semantic types C , D , E , and, F , while the meta-semantic type \mathbf{B} consists of two semantic types M and N . However, each of the concepts u , w , x , y , and, z is in a different intersection of semantic types. The concept u is in the semantic type intersection $C \cap D \cap M$; the concept w is in the semantic type intersection $D \cap M$; the concept x is in the semantic type intersection $E \cap M$; the concept y is in the semantic type intersection $E \cap N$; and finally the concept z is in the semantic type intersection $F \cap N$.

The notation for a pure intersection is the list of names of each meta-semantic type followed by its corresponding semantic type (or intersection of semantic types) in curly

brackets, where the intersection symbol \cap appears between any two meta-semantic types in the intersection list. The intersection of meta-semantic types $\mathbf{A} \cap \mathbf{B}$ in Figure 4 is partitioned into five pure intersections in Figure 5. They are $\mathbf{A} \{ C \cap D \} \cap \mathbf{B} \{ M \}$, $\mathbf{A} \{ D \} \cap \mathbf{B} \{ M \}$, $\mathbf{A} \{ E \} \cap \mathbf{B} \{ M \}$, $\mathbf{A} \{ E \} \cap \mathbf{B} \{ N \}$, and $\mathbf{A} \{ F \} \cap \mathbf{B} \{ N \}$. The union of all five pure intersections is $\mathbf{A} \cap \mathbf{B}$.

Consider the following example. In the metaschema, the semantic types *Event*, *Activity*, *Daily or Recreation Activity*, and *Machine Activity* are grouped into the meta-semantic type **Event**; while the semantic types *Idea or Concept*, *Functional Concept*, *Temporal Concept*, *Qualitative Concept*, *Quantitative Concept*, *Spatial Concept*, and some others are grouped into the meta-semantic type **Idea or Concept**. The concepts STRESSFUL EVENTS, HOUSEHOLD CONSUMPTION, and WITHDRAWING CARE are all in the intersection of meta-semantic types **Event** \cap **Idea or Concept**. However, they are in different pure intersections. The concept STRESSFUL EVENTS is in the pure intersection **Event** $\{Event\} \cap$ **Idea or Concept** $\{Qualitative Concept\}$, HOUSEHOLD CONSUMPTION is in **Event** $\{Activity\} \cap$ **Idea or Concept** $\{Quantitative Concept\}$, and WITHDRAWING CARE is in **Event** $\{Activity\} \cap$ **Idea or Concept** $\{Idea or Concept\}$.

After the pure intersections of the meta-semantic types are generated, the domain expert can now review all the concepts in one pure intersection together, which is easier since they all have the same compound semantics as expressed by the associated semantic types [8].

An effective auditing process should expose many errors with limited efforts. With this in mind, we concentrate on reviewing the pure intersections of meta-semantic types containing very few concepts. Our second hypothesis is that the likelihood of a mistake for a small pure intersection is higher than in the case of a large pure intersection. The reason is that if a combination of semantic types makes sense semantically, then there would probably be quite a few or at least several concepts associated with it. For example, the pure intersection **Chemical** $\{Organic Chemical\} \cap$ **Pharmacologic Substance** $\{Pharmacologic Substance\}$ is a reasonable combination, since many drugs are composed of organic chemicals. This pure intersection contains the largest number of concepts (70,436) among all pure intersections. On the other hand, the case where a pure intersection contains only one or two concepts may indicate an erroneous categorization, where no concepts should be associated with such

a combination.

The process of our auditing technique is as follows. First, we identify all intersections of meta-semantic types of the metaschema. All those intersections are refined to generate the pure intersections. Now we apply a divide and conquer approach in order to limit the number of concepts reviewed by domain expert, while covering the concepts with high likelihood of a wrong categorization. Hence, we minimize the effort while trying to maximize the impact of the audit. On one side, a domain expert reviews concepts of each pure intersection containing relatively small number of concepts. The total number of concepts reviewed is limited due to the low cardinality of the pure intersections considered. On the other hand, a domain expert reviews the semantic soundness of the intersected semantic types of all medium and large size pure intersections, looking for combinations of semantic types, which are not semantically sound. This is one review per pure intersection independent of the number of concepts assigned to this intersection. Only for those unlikely pure intersections, the expert will review their concepts independent of their number. This way the number of concepts reviewed is limited due to the small number of semantically unsound pure intersections, and their likelihood to have erroneous categorization is high due to their unsound compound semantics.

4 Results

Using our auditing process, we first identified all intersections of meta-semantic types, which contain total 170,179 concepts. Then, each one of them was partitioned into pure intersections to create 874 pure intersections. Table 1 describes the distribution of the number of concepts for all pure intersections. Reviewing Table 1, we found that most of the pure intersections are small sets of one or two concepts. For example, there are 332 pure intersections containing only one concept, 113 pure intersections containing only two concepts. On the other extreme, the pure intersection that contains the largest set of concepts has 70,436 concepts. The average number and median number of concepts for a pure intersection are 195 and 2 respectively. Note that the median is small due to the large number of very small pure intersections. On the other hand, the weighted median number of concepts is 27,002

due to the size of the two largest pure intersections.

4.1 Analysis of Small Pure Intersections

Our domain experts were Dr. Gai Elhanan and Hua Min. Dr. Elhanan is an experienced MD with Medical Informatics training and experience in Columbia Presbyterian Medical Center. Hua Min is a Ph.D. student at the Computer Science Department of NJIT who obtained her MD training and internship in China.

After examining all pure intersections containing one to 10 concepts, covering a total of 657 pure intersections and 1,680 concepts, we found a high percentage of incorrect categorizations that can be divided into four categories. They are 1) Polysemy, 2) Inconsistency, 3) Miscategorization, and 4) Redundant categorization. We discussed the redundant categorizations in another paper [17]. All other three categories are discussed in this section.

In Table 2, we list some examples of those three types of incorrect categorizations. First let us review examples of polysemy. The first indication of a polysemy error is an intersection of semantic types, which is not semantically sound. For example, the concept TALIPES CAVUS is the only concept of the pure intersection **Anatomical Abnormality** $\{ \textit{Congenital Abnormality} \cap \textit{Acquired Abnormality} \} \cap \textbf{Finding}\{ \textit{Sign or Symptom} \}$. How can a congenital abnormality be an acquired abnormality at the same time? These two semantic types are mutually exclusive siblings in the semantic network. In order to disambiguate polysemous concepts, one can replace the polysemous concept by several new different concepts according to the different intersecting semantic types, and let each one of the created concepts be associated with only one of the semantic types. In the above case, one possible solution is to create two alternative concepts, TALIPES CAVUS <1>² that belongs to the pure intersection **Anatomical Abnormality** $\{ \textit{Congenital Abnormality} \} \cap \textbf{Finding}\{ \textit{Sign or Symptom} \}$, and TALIPES CAVUS <2> that belongs to the pure intersection **Anatomical Abnormality** $\{ \textit{Acquired Abnormality} \} \cap \textbf{Finding}\{ \textit{Sign or Symptom} \}$. Another possible solution, instead of creating two concepts for TALIPES CAVUS, is to recategorize this concept with the parent semantic type *Anatomical Abnormality* of the two semantic types currently assigned to it.

²Following the UMLS notation, we denote the different meanings of a polysemous concept by < 1 > and < 2 >.

Table 1: Distribution of Number of Concepts for Pure Intersections

No. of Concepts	No. of Pure Intersections	No. of Concepts	No. of Pure Intersections	No. of Concepts	No. of Pure Intersections	No. of Concepts	No. of Pure Intersections
1	332	39	1	113	1	440	1
2	113	40	3	116	1	453	1
3	64	42	1	118	1	466	1
4	35	43	2	119	1	484	1
5	28	47	2	120	1	522	1
6	25	48	2	122	1	534	1
7	18	49	2	125	1	541	1
8	17	50	1	127	2	543	1
9	17	51	2	128	1	549	1
10	8	52	1	130	1	568	1
11	9	53	2	131	1	587	1
12	8	54	1	135	1	603	1
13	3	55	2	142	1	648	1
14	12	56	1	148	1	649	1
15	4	57	2	150	1	678	1
16	5	60	1	154	1	688	1
17	6	62	1	161	1	787	1
18	4	67	1	169	1	815	1
19	3	68	1	176	1	880	1
20	6	69	1	185	1	883	1
21	1	70	1	197	1	1096	1
22	3	74	1	213	1	1187	1
23	4	76	1	230	1	1219	1
24	3	77	1	234	1	1290	1
25	1	79	1	237	1	1364	1
26	2	80	1	242	1	1460	1
27	2	81	1	243	1	2132	1
28	1	85	1	247	1	2339	1
29	1	87	1	279	1	3074	1
30	3	88	1	287	1	3126	1
31	3	90	1	289	1	4299	1
32	2	93	2	296	1	4937	1
33	1	96	1	304	1	8061	1
35	2	98	2	328	1	10407	1
36	2	106	1	339	1	27002	1
37	1	107	1	341	1	70436	1
38	2	111	1	354	1		

Table 2: Examples of Various Types of Incorrect Categorizations

CONCEPTS	PURE INTERSECTIONS
Polysemy	
TALIPES CAVUS	Abnormality { <i>Congenital Abnormality</i> \cap <i>Acquired Abnormality</i> } \cap Finding { <i>Sign or Symptom</i> }
TOXICODENDROM (POISON IVY)	Plant { <i>Plant</i> } \cap Pathologic { <i>Disease or Syndrome</i> }
Inconsistency	
MUSSELS PRAWNS SCALLOP, NOS	Animal { <i>Invertebrate</i> } \cap Substance { <i>Food</i> }
THIRSTY PHYSICAL EXHAUSTION	Physiologic Function { <i>Physiologic Function</i> } \cap Finding { <i>Sign and Symptom</i> }
Miscategorization	
CYTARABINE	Chemical { <i>Nucleic Acid, Nucleoside, or Nucleotide</i> \cap <i>Biomedical or Dental Material</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> }
LAMINARIA TENTS MARINE ALGAE	Plant { <i>Alga</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> }
DIPHThERIA-TETANUS VACCINE	Health Care Activity { <i>Therapeutic or Preventive Procedure</i> } \cap Natural Phenomenon or Process { <i>Natural Phenomenon or Process</i> }
BILISCAN GLUCOSE RANDOM	Chemical { <i>Chemical</i> } \cap Substance { <i>Body Substance</i> }
SUPPORT HOSPITAL-PATIENT RELATIONS FACILITY-PATIENT RELATIONS	Behavior { <i>Social Behavior</i> } \cap Health Care Activity { <i>Health Care Activity</i> }
Polysemy + Inconsistency	
ADULTHOOD OLD-AGE	Idea or Concept { <i>Temporal Concept</i> } \cap Group { <i>Age Group</i> }
Polysemy + Miscategorization	
RHOGAM SCREEN TOTAL BODY CLEARANCE RATE SPECIFIC GRAVITY MEASUREMENT OXYGEN MEASUREMENT, -PARTIAL PRESSURE, ARTERIAL	Finding { <i>Lab or Test Result</i> } \cap Health Care Activity { <i>Laboratory Procedure</i> }

This way is consistent with the representation of this concept in the source terminologies of the UMLS, which the UMLS needs to conserve.

Similarly, the concept TOXICODENDROM (POISON IVY) has been assigned to the pure intersection **Plant**{*Plant*} \cap **Pathologic Function**{*Disease or Syndrome*}. This is also a polysemy error. The same concept is used for the plant and for the disease caused by the plant. To resolve this polysemy, the current version of the UMLS contains two concepts TOXICODENDROM (POISON IVY) <1>, that is a plant, and TOXICODENDROM (POISON IVY) <2> that is a disease, caused by the plant. Both above intersections then disappear with the change in the categorization of the polysemous concept.

In Table 2, we also list some examples of inconsistent categorization. The concepts MUSSELS; SCALLOP, NOS; and PRAWNS are the only three concepts of the pure intersection **Animal**{*Invertebrate*} \cap **Substance**{*Food*}. However, they are not the only invertebrates that are food. Many others, e.g., SHRIMP, LOBSTER, and, OCTOPUS are food as well. But they are assigned only to *Invertebrate*, not *Food*. Thus, if those three concepts are categorized as food, some other invertebrates such as shrimp, lobster, and octopus should also be categorized as food. This is an inconsistent categorization case. Note that in this case, one semantic type (e.g., *Invertebrate*) represents a sort type while the other (e.g., *Food*) represents a role type. Another example occurs with the concepts THIRSTY and PHYSICAL EXHAUSTION that are the only two concepts of the pure intersection **Physiologic Function**{*Physiologic Function*} \cap **Finding**{*Sign and Symptom*}. As in the previous example, other concepts, e.g., STARVATION and DEHYDRATION should have been also in this pure intersection. However, they are categorized only as *Sign and Symptom*, not *Physiologic Function*.

Table 2 also lists some examples of miscategorization. We distinguish between 2 cases.

Case 1: All concepts in one pure intersection are categorized incorrectly.

Case 1.1: The concepts should be categorized only to part of the intersecting semantic types, not all of them.

For example, the concept CYTARABINE is the only concept in the pure intersection **Chemical** {*Nucleic Acid, Nucleoside, or Nucleotide* \cap *Biomedical or Dental Material*} \cap **Pharmacologic Substance**{*Pharmacologic Substance*}. However, CYTARABINE is a nucleoside analog. It is a pharmacologic substance and potentially a hazardous substance. It

should be assigned to neither the semantic type *Nucleic Acid, Nucleoside, or Nucleotide* nor the semantic type *Biomedical or Dental Material*. Thus, this pure intersection will not exist after this miscategorization error is resolved. In another example, the pure intersection **Pharmacologic Substance**{*Pharmacologic Substance*} \cap **Plant**{*Alga*} contains two concepts LAMINARIA TENTS and MARINE ALGAES. However, LAMINARIA TENTS is a pharmacologic substance produced from a type of marine algae. It should not be assigned to the semantic type *Alga*, while the concept MARINE ALGAES should be only assigned to the semantic type *Alga*. Hence, after correcting the errors, there will be no such a pure intersection.

Case 1.2: All concepts should not have been categorized as any one of the intersecting semantic types.

For example, the pure intersection **Health Care Activity**{*Therapeutic or Preventive Procedure*} \cap **Natural Phenomenon or Process**{*Natural Phenomenon or Process*} contains only one concept DIPHTHERIA-TETANUS VACCINE. However, a vaccine is a pharmacologic substance and immunologic factor. It is neither a procedure nor a process. So it should not be assigned to any one of those two semantic types. This intersection becomes empty.

In another example, the concepts BILISCAN and GLUCOSE RANDOM are the only two concepts of the pure intersection **Substance**{*Body Substance*} \cap **Chemical**{*Chemical*}. However, the concepts BILISCAN and GLUCOSE RANDOM are neither body substance nor chemical. They are just laboratory procedures. Thus, both concepts should not be assigned to any one of those semantic types. Again, the intersection becomes empty.

Case 2: Some of the concepts in the pure intersection are categorized incorrectly.

For example, the concepts SUPPORT, HOSPITAL-PATIENT RELATIONS, and FACILITY-PATIENT RELATIONS are the only concepts in the pure intersection **Behavior**{*Social Behavior*} \cap **Health Care Activity**{*Health Care Activity*}. For the concepts HOSPITAL-PATIENT RELATIONS and FACILITY-PATIENT RELATIONS, they are categorized perfectly. However, the concept SUPPORT is not necessarily related to health care. Thus, it should not be assigned to the semantic type *Health Care Activity*.

However, not all concepts in one pure intersection always demonstrate the same kind of errors. Sometimes, we see different kinds of errors for various concepts in the same pure

intersection.

Mixed Case 1: Polysemy + inconsistency

For example, consider the concepts ADULTHOOD and OLD-AGE, which are the only two concepts of the pure intersection **Idea or Concept**{*Temporal Concept*} \cap **Group**{*Age group*}. This is an inconsistent categorization since many other concepts, e.g., the concepts CHILDHOOD, JUVENILE, and, YOUNG ADULTS should have been also in this pure intersection. However, they are assigned only to *Age group*. Also, this is a polysemy error because each of those two concepts refers to two different concepts, one is the state of age, and the other is the group of people in that state. To disambiguate this concept, we could replace the concept ADULTHOOD by ADULTHOOD <1> that is a temporal concept and ADULTHOOD <2> that is an age group. Similarly, we can replace the concept OLD-AGE by OLD-AGE <1> that is a temporal concept and OLD-AGE <2> that is an age group. Again the intersection becomes empty, while the polysemy and inconsistency are resolved.

Mixed Case 2: Polysemy + miscategorization

The example in this case occurs with the pure intersection **Health Care Activity**{*Laboratory Procedure*} \cap **Finding**{*Lab or Test Result*}, which contains four concepts TOTAL BODY CLEARANCE RATE; SPECIFIC GRAVITY MEASUREMENT; OXYGEN MEASUREMENT, PARTIAL PRESSURE, ARTERIAL; and RHOGAM SCREEN. However, the concept TOTAL BODY CLEARANCE RATE is found polysemous because it refers to two concepts, one is a laboratory procedure and the other is its result. To disambiguate this polysemous concept, two concepts have to be created. One is assigned to *Laboratory Procedure* and the other is assigned to *Lab or Test Result*. The concepts SPECIFIC GRAVITY MEASUREMENT; OXYGEN MEASUREMENT, PARTIAL PRESSURE, ARTERIAL; and RHOGAM SCREEN are laboratory procedures and should not be assigned to the semantic type *Lab or Test Result*. Therefore, the intersection will become empty.

Table 3 lists the results of the analysis for pure intersections containing 1 to 10 concepts. In this table, we compute the error percentages for both erroneous pure intersections (i.e., with some incorrectly categorized concepts) and incorrectly categorized concepts. The percentage of erroneous pure intersections and the percentage of incorrect categorizations are quite high for the pure intersections containing small number of concepts up to the inter-

Table 3: Analysis of Errors in Small Pure Intersections

No. of concepts	No. of pure intersections	No. of pure intersections with errors	Percentage of intersections with errors	Total No. of concepts	No. of erroneous concepts	Percentage of erroneous concepts
1	332	120	34	332	120	34
2	113	56	50	226	105	46
3	64	27	42	192	75	39
4	35	13	37	140	43	31
5	29	13	45	145	55	48
6	25	9	36	150	44	29
7	18	2	11	126	13	10
8	17	3	18	136	11	8
9	17	4	23	153	13	8
10	8	0	0	80	0	0
Total	658	247	38%	1680	479	29%

sections of six concepts. It decreases when the size increases above six concepts. For all the pure intersections, with up to 10 concepts, 38% contain erroneous categorization and 29% of the concepts have incorrect categorization.

4.2 Analysis of Large Pure Intersections

For large and medium sized pure intersections, the domain expert will not review their large number of concepts. The domain expert will just check the semantic soundness of medium to large size pure intersections. Analysis of the concepts is limited only to the pure intersections judged semantically suspicious. The domain experts reviewed the pure intersections containing more than 10 concepts. There are 217 such pure intersections. Almost all of them are semantically sound. For example, the pure intersection **Chemical**{*Organical Chemical*} \cap **Pharmacologic Substance**{*Pharmacologic Substance*}, which contains the largest number of concepts (70,436), is a reasonable combination, since many drugs are also organic chemicals. Similarly, for the pure intersection **Chemical**{*Amino Acid, Peptide or Protein*} \cap **Biologically Active Substance**{*Enzyme*} that contains 27,002 concepts.

In Table 4, we list the 16 largest pure intersections that are associated with the meta-semantic type **Chemical** and the 16 largest pure intersections that are not associated with

the meta-semantic type **Chemical**. The 16th pure intersection associated with the meta-semantic type **Chemical** is an interesting case. As a matter of fact, it is a case of redundant categorization[17]. All these 883 concepts should not be categorized as **Organic Chemical**. After removing this redundant categorization, those 883 concepts should join the 9th pure intersection in the left column of Table 4. A similar situation exists for the 10th pure intersection in the left column of Table 4. The concepts actually belong to the second largest pure intersection in this column. All others are semantically sound.

A few of the pure intersections are semantically suspicious. For example, the pure intersection **Manufactured Object**{*Manufactured Object*} \cap **Organization**{*Organization*} contains 70 concepts. However, no concepts can be a manufactured object, as well an organization simultaneously. Basically, the semantic types *Manufactured Object* and *Organization* are mutually exclusive. Therefore, the pure intersection **Manufactured Object**{*Manufactured Object*} \cap **Organization**{*Organization*} is semantically suspicious and probably should not exist. All 70 concepts are reviewed and found polysemous. For example, the concept DAY CARE CENTERS FOR CHILDREN is in this pure intersection. However, it refers to two concepts. One is a organization and the other is a manufactured object that includes buildings and facilities in day care centers. All other concepts in this pure intersection such as PRIMARY SCHOOLS, LABORATORIES, INFORMATION CENTER, etc. have the same polysemy errors. To disambiguate these polysemous concepts, we suggest to create two concepts for each polysemous concept. The original one is assigned to the semantic type *Organization* and the other with the word "building" added is assigned to the semantic type *Manufactured Object*. After disambiguating these polysemous concepts, the original pure intersection will not contain any concepts and should not exist. Similarly, since the semantic types *Inorganic Chemical* and *Organic Chemical* are mutually exclusive, the pure intersection **Chemical**{*Organic Chemical* \cap *Inorganic Chemical*} \cap **Pharmacologic Substance**{*Pharmacologic Substance*} should not exist. The 247 concepts inside this pure intersection are reviewed. All of them should not be categorized as inorganic chemical. Thus, they are in the largest pure intersection **Chemical**{*Organic Chemical*} \cap **Pharmacologic Substance**{*Pharmacologic Substance*}.

Among the 217 pure intersections containing more than 10 concepts, only 6 medium

Table 4: Largest Pure Intersections

CHEMICAL	NON-CHEMICAL
Chemical{ <i>Organic Chemical</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (70,436)	Anatomical Abnormality{ <i>Congenital Abnormality</i> } \cap Pathologic Function{ <i>Disease or Syndrome</i> } (1,096)
Chemical{ <i>Amino Acid, Peptide, or Protein</i> } \cap Biologically Active Substance { <i>Enzyme</i> } (27,002)	Anatomical Abnormality{ <i>Acquired Abnormality</i> } \cap Pathologic Function{ <i>Disease or Syndrome</i> } (815)
Chemical{ <i>Amino Acid, Peptide, or Protein</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (10,407)	Finding{ <i>Finding</i> } \cap Physiologic Function { <i>Organ or Tissue Function</i> } (787)
Chemical{ <i>Amino Acid, Peptide, or Protein</i> } \cap Biologically Active Substance { <i>Biologically Active Substance</i> } (8,061)	Anatomical Abnormality{ <i>Anatomical Abnormality</i> } \cap Pathologic Function { <i>Disease or Syndrome</i> } (603)
Chemical{ <i>Amino Acid, Peptide, or Protein</i> } \cap Biologically Active Substance { <i>Immunologic Factor</i> } (4,937)	Health Care Activity{ <i>Therapeutic or Preventive Procedure</i> } \cap Occupational Activity{ <i>Educational Activity</i> } (549)
Chemical{ <i>Amino Acid, Peptide, or Protein</i> } \cap Biologically Active Substance { <i>Receptor</i> } (4,299)	Finding{ <i>Finding</i> } \cap Pathologic Function { <i>Pathologic Function</i> } (534)
Chemical{ <i>Nucleic Acid, Nucleoside, or Nucleotide</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (3,126)	Finding{ <i>Finding</i> } \cap Pathologic Function { <i>Disease or Syndrome</i> } (339)
Chemical{ <i>Steroid</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (3,074)	Pathologic Function { <i>Disease or Syndrome</i> } \cap Finding{ <i>Sign or Symptom</i> } (328)
Chemical{ <i>Carbohydrate</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (2,339)	Event{ <i>Daily or Recreational Activity</i> } \cap Health Care Activity { <i>Therapeutic or Preventive Procedure</i> } (243)
Chemical{ <i>Organic Chemical</i> \cap <i>Amino Acid, Peptide, or Protein</i> } \cap Pharmacologic Substance{ <i>Pharmacologic Substance</i> } (2,132)	Health Care Activity{ <i>Health Care Activity</i> } \cap Occupational Activity { <i>Educational Activity</i> } (197)
Chemical{ <i>Lipid</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (1,460)	Manufactured Object { <i>Manufactured Object</i> } \cap Entity{ <i>Intellectual Product</i> } (176)
Chemical{ <i>Organic Chemical</i> } \cap Pharmacologic Substance { <i>Antibiotic</i> } (1,364)	Pathologic Function { <i>Pathologic Function</i> } \cap Finding{ <i>Sign or Symptom</i> } (161)
Chemical{ <i>Inorganic Chemical</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (1,290)	Organism Attribute { <i>Organism Attribute</i> } \cap Finding{ <i>Finding</i> } (135)
Chemical{ <i>Amino Acid, Peptide, or Protein</i> } \cap Pharmacologic Substance{ <i>Pharmacologic Substance</i> } \cap Biologically Active Substance{ <i>Immunologic Factor</i> } (1,219)	Plant{ <i>Plant</i> } \cap Substance{ <i>Food</i> } (125)
Chemical{ <i>Organophosphorus Compound</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (1,187)	Health Care Activity { <i>Diagnostic Procedure</i> } \cap Entity{ <i>Intellectual Product</i> } (111)
Chemical{ <i>Organic Chemical</i> \cap <i>Carbohydrate</i> } \cap Pharmacologic Substance { <i>Pharmacologic Substance</i> } (883)	Physiologic Function{ <i>Cell Function</i> } \cap Pathologic Function { <i>Cell or Molecular Dysfunction</i> } (106)

sized are judged semantically suspicious. After reviewing all 405 concepts in these 6 semantically suspicious pure intersections, we found only two pure intersections that should exist. One is the pure intersection **Physiologic Function**{*Organ or Tissue Function*} \cap **Pathologic Function**{*Pathologic Function*}. Despite the suspicious semantic type combination, all 12 concepts in it are categorized correctly. For example, the concepts MESIAL MOVEMENT OF TEETH, SKIN WRINKLING, and, OSTEOLYSIS are organ or tissue function, but they are pathologic function as well. The other semantically suspicious pure intersection that should exist is **Chemical**{*Amino Acid, Peptide, or Protein \cap Element, Ion or Isotope*} \cap **Biologically Active Substance**{*Immunologic Factor*} \cap **Pharmacologic Substance**{*Pharmacologic Substance*}. For example, the concepts IODINE I 131 MONOCLONAL ANTIBODY 3F8, IODINE I 131 MONOCLONAL ANTIBODY ANTI-B1, and IODINE I 131 MONOCLONAL ANTIBODY G-250 are in this pure intersection. It is true that an antibody cannot be element, ion or isotope. However, each concept in this pure intersection is not an antibody produced naturally, but rather an antibody engineered for therapeutic purposes, coupled with a radioactive substance in order to selectively target the tissue to that the antibody is directed. Therefore, each such concept is categorized to both *immunologic factor* and *element, ion, or isotope*.

However, all other 377 concepts in the other 4 semantically suspicious pure intersections are erroneously assigned to some semantic types. Table 5 lists all semantically suspicious pure intersections, the number of concepts categorized to them, and the pure intersections to that the concepts should belong.

Out of the 405 concepts in these 6 pure intersections reviewed, 377 concepts, about 93%, have erroneous categorizations. Note that these reviews are much easier than those of the small pure intersections, since all of the concepts of a large or medium pure intersection typically share the same semantics and have the same categorization error. Hence, our method of auditing medium to large pure intersections is an example of a successful auditing process, finding many errors with a limited review effort.

Table 5: Semantically Suspicious Medium Sized Pure Intersections

Semantically Suspicious Pure Intersections	No. of Concepts	Correct semantic types or Pure Intersections
Physiologic Function <i>{Organ or Tissue Function}</i> \cap Pathologic Function <i>{Pathologic Function}</i>	12	Same as original
Chemical <i>{Amino Acid, Peptide, or Protein}</i> \cap <i>Element, Ion or Isotope}</i> \cap Biologically Active <i>{Immunologic Factor}</i> \cap Pharmacologic Substance <i>{Pharmacologic Substance}</i>	16	Same as original
Chemical <i>{Inorganic Chemical} \cap Organic Chemical \cap Amino Acid, Peptide, or Protein}</i> \cap Pharmacologic Substance <i>{Pharmacologic Substance}</i>	20	Chemical <i>{Organic Chemical} \cap Amino Acid, Peptide, or Protein}</i> \cap Pharmacologic Substance <i>{Pharmacologic Substance}</i>
Chemical <i>{Organic Chemical} \cap Element, Ion or Isotope}</i> \cap Pharmacologic Substance <i>{Pharmacologic Substance}</i>	40	Chemical <i>{Organic Chemical}</i> \cap Pharmacologic Substance <i>{Pharmacologic Substance}</i>
Manufactured Object <i>{Manufactured Object}</i> \cap Organization <i>{Organization}</i>	70	Organization <i>{Organization}</i>
Chemical <i>{Organic Chemical} \cap Inorganic Chemical}</i> \cap Pharmacologic Substance <i>{Pharmacologic Substance}</i>	247	Chemical <i>{Organic Chemical}</i> \cap Pharmacologic Substance <i>{Pharmacologic Substance}</i>

5 Discussion

The task of checking the correctness of all the concepts, and their related data, in a large terminology is overwhelming. Usually, there are not enough resources for such a task. Furthermore, the tendency of terminology designers is to invest most of the available resources in extending the terminology.

However the accuracy of a terminology is critical for its mission in overcoming terminological differences among various health care information systems. Thus, auditing techniques for terminologies, similar to auditing techniques in other fields, e.g., finance, are designed in an effort to expose as many errors as possible with a limited effort.

Our auditing technique is designed in the same approach by checking only a limited number of concepts such that their probability to be erroneous is high. Our technique is based on our two hypotheses. Our first hypothesis is that the probability of a concept to be incorrectly assigned to a pure intersection is higher than to be incorrectly assigned to an intersection of semantic types that are in one meta-semantic type. In order to confirm this hypothesis, we checked 100 concepts that are in the intersections, containing 1 to 6 concepts of semantic types, from the same meta-semantic type. The error percentage is about 20% versus about 40% in our method. This result confirms our first approach that using intersections of meta-semantic types is an effective auditing approach.

Our approach also differentiates between the treatment of small intersections and large intersections. This approach is based on our second hypothesis that the probability of incorrect categorizations is high for small pure intersections. As we see in Table 3, the results confirm our second hypothesis. The percentage of erroneous categorizations for pure intersections of up to six concepts, about 40%, is high. The percentage decreases for medium size pure intersection of 7 to 10 concepts, and is further reduced for large ones where most pure intersections are judged as semantically sound. This observation confirms our second hypothesis and shows that our auditing approach to concentrate on the concept-based analysis of small pure intersections is justified.

These results support the design of the audit technique as a divide and conquer technique applying different processing to small intersections and to large intersections.

We need to stress that the fact that there is an intersection of two semantic types assigned to two different meta-semantic types, does not automatically imply that there is an error in the categorization of the concepts of this intersection. The interdisciplinary nature of medicine implies that medical knowledge is also interdisciplinary. Thus, it is quite natural for a concept to be categorized to several semantic types. Such semantic types assigned the same concepts may or may not be closely related; and thus in the same meta-semantic type or not.

Thus, we do not imply that all concepts in pure intersection are erroneously categorized. Even for the small pure intersections, 60% of the concepts are properly categorized. Actually the fact that many concepts are assigned to a specific combination of semantic types, i.e., to their intersection, supports the idea that this combination is semantically sound in spite of the fact that the intersecting semantic types are assigned to different meta-semantic types.

Table 4 lists the 16 largest pure intersections of semantic types that are descendants of the semantic type *Chemical* and 16 largest pure intersections that are not descendants of *Chemical*. The reason for this distinction is the dominance of the first kind among the largest pure intersections. The soundness of the combination of the semantic types in these pure intersection is straightforward. The two exceptions are the 10th and 16th pure intersections in the left column, which are redundant categorization cases.

As reported in Table 5 only 4 pure intersections of medium size out of the 6 suspicious ones are actually found semantically unsound and the categorization of their concepts need some modification.

As a matter of fact, one can apply our auditing approach for a partition of the Semantic Network rather than for a metaschema. While every metaschema is based on a partition, not every partition fits for the construction of a metaschema. For example, in the partition [15, 3] called semantic partition ³, not all 15 groups consist of a connected subtree of the Semantic Network, a necessary condition, for constructing a metaschema, based on a partition.

Hence, applying our audit technique for a partition rather than a metaschema can broaden its usefulness. We note that out of the 32 pure intersections of Table 4, only 5 would be pure intersections when applying our technique to the partition of [15, 3], instead of our

³O. Bodenreider, personal communication.

metaschema. All 5 are on the right column of Table 4. Hence, part of the expert work of checking the semantic soundness of the pure intersection is saved when using the semantic partition. On the other hand, using the semantic partition some of the erroneous small pure intersections would not be detected. For example, the errors in the categorizations of the concepts TALIPES CAVUS and CYTARABINE (see Table 2) as well as three of the four semantically unsound medium sized pure intersections (see Table 5) would not be detected. There seem to be a tradeoff between the recall and the precision. It is interesting to note that one of the principles underlying the semantic partition is exclusivity, to minimize the number of concepts associated with different groups. The exclusivity and proximity qualities coupled with the flexibility regarding the connectivity of the groups, enable to avoid detecting many of the large pure intersections that are actually semantically sound. The cohesive metaschema does not share these qualities. Using it in our audit, generates the large pure intersections that are semantically sound. On the other hand, the cohesive metaschema helps us uncover many erroneous small pure intersections that would be missed if the semantic partition would be used instead.

Finally we note that in the design of a metaschema and its underlying partition, one can choose various granularities, resulting various number of meta-semantic types. The choice of granularity seems to influence the tradeoff between recall and precision. In our view, the emphasis on recall is more important. One reason is that the effort on checking the soundness of a pure intersection is independent on the size of the intersection and is easier than checking the categorizations of a concept since it is done with the broad categorizations of semantic types.

We note that for some errors exposed in our audit, the actual error was not due to categorization of a concept to two semantic types in different meta-semantic types but into two semantic types in the same meta-semantic type, which are exclusive since they are semantically incompatible. Examples of exclusive pairs of semantic types are (*Congenital Abnormality*; *Acquired Abnormality*) and (*Organic Chemical*; *Inorganic Chemical*). Each of the two examples is a pair of siblings, where a concept should not be assigned to both, due to their semantic incompatibility. However a pair of *Inorganic Chemical* and any descendant of *Organic Chemical* will also be an exclusive pair. Our audit technique does not consider

intersections of exclusive pairs of semantic types. But this is a natural potential extension to complement our technique. One could enumerate all exclusive pairs of semantic types and check their intersections. For every concept assigned to two exclusive semantic types, one should consider whether it is just a categorization error or a case of polysemous categorization. In the later case, one can create two concepts with the different meanings, one for each semantic type, or recategorize the polysemous concept to the parent semantic type of the two exclusive sibling semantic types to preserve consistency with the source terminology of the UMLS, from which these concepts come. The later case is demonstrated earlier regarding TALIPES CAVUS.

In our auditing methods, we identify three kinds of incorrectly categorized concepts, polysemy, inconsistency, and miscategorization. For some concepts we even see a combination of various kinds. Typically, errors of the first kind stem from polysemy in the terminology used by health care workers, in verbal communication. Humans overcome such polysemy cases due to the context in which they are used. However, a concept entry in the META should be non-polysemous. In some cases of inconsistency, one semantic type represents a sortal type while the other represents a role type. The sortal type categorization seem to be the consistent one while the role type appears only in some of the cases. See for example the intersection of *Invertebrate* and *Food* in Table 2. For such cases of inconsistent categorization, a decision is needed whether to add the missing categorization to all other concepts qualified or to remove the extra categorization for the concepts that had it. Either way will lead to a consistent categorization.

6 Conclusions

The UMLS integrated many biomedical terminologies. During the integration, each concept was assigned to at least one semantic type. However, due to the size and complexity of the UMLS, it is unavoidable that some incorrect associations have been generated. To find and correct such wrong associations, we introduce in [8] the notion of intersection semantic types. The more complex concepts, those with compound semantics [8], are associated with intersection semantic types. These are concepts, which are likely to have errors in

their modeling or categorization. Hence, the review of these concepts will provide effective auditing. However, the number of such concepts is quite large and only a small sample of them were reviewed in [8] to provide a proof of concept. The comprehensive review of all such concepts is an overwhelming task.

In this paper we set up to design an effective auditing technique to review a substantial part of the concepts of intersection semantic types, actually which are more likely to have an erroneous categorization. For this purpose, we have developed an efficient auditing technique based on the pure intersections of meta-semantic types of the metaschema. Our divide and conquer approach treats small and large pure intersections differently. The review of the concepts of small pure intersections led to the recognition of different types of wrong assignments. The results of analysis for the pure intersections containing 1 to 10 concepts are presented. On the other hand, the combinations of all pure intersections containing more than 10 concepts are reviewed to check their semantic soundness. The list of semantically suspicious pure intersections containing more than 10 concepts is presented and all their concepts are reviewed. The results confirm our two hypotheses, which are the basis for our auditing technique showing its efficiency.

Due to our divide and conquer approach, only a limited number of concepts are actually reviewed. A meaningful portion of them were found to have erroneous categorization. Hence our technique provides an effective auditing,

In this paper, we do not review intersections of semantic types associated with the same meta-semantic type. More errors are expected there but their likelihood is lower than in this paper. Still it is recommended to review these intersections, but we do not consider it in this work.

7 Acknowledgment

We would like to thank James Cimino for his important feedback on our analysis results. An anonymous reviewer presented interesting remarks and insights, which helped to enrich this paper.

References

- [1] O. Bodenreider. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. In *Proc. '2001 AMIA Symposium*, pages 57–61, 2001.
- [2] O. Bodenreider. An object-oriented model for representing semantic locality in the umls. *Proc. Medinfo. 2001*, 10(1):161–165, 2001.
- [3] O. Bodenreider and A. T. McCray. Exploring semantic groups through visual approaches. *J. Biomed Inform*, 2003.
- [4] Z. Chen, Y. Perl, M. Halper, J. Geller, and H. Gu. Partitioning the UMLS Semantic Network. *IEEE Trans. Information Technology in Biomedicine*, 6(2):102–108, June 2002.
- [5] J. J. Cimino. Auditing the unified medical language system with semantic methods. *JAMIA*, 5(1):41–51, 1998.
- [6] J. J. Cimino. Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS Metathesaurus. In *Proc. '2001 AMIA Symposium*, pages 120–124, 2001.
- [7] L. Goldstein, D. Schneider, and M. Siegel. *Finite Mathematics and Its Applications*. Prentice Hall, 2003.
- [8] H. Gu, Y. Perl, J. Geller, M. Halper, L. Liu, and J. J. Cimino. Representing the UMLS as an OODB: Modeling issues and advantages. *JAMIA*, 7(1):66–80, Jan./Feb. 2000. Selected for reprint in Haux R, Kulikowski C, eds.: Yearbook of Medical Informatics, International Medical Informatics Association, Rotterdam, 2001: 271-285.
- [9] W. T. Hole and S. Srinivasin. Discovering missed synonymy in a large concept-oriented Metathesaurus. *JAMIA*, 7:354–358, 2000.

- [10] B. L. Humphreys and D. A. B. Lindberg. Building the Unified Medical Language System. In *Proc. Thirteenth Annual Symposium on Computer Applications in Medical Care*, pages 475–480, Washington, DC, Nov. 1989.
- [11] B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett. The Unified Medical Language System: An informatics research collaboration. *JAMIA*, 5(1):1–11, 1998.
- [12] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32:281–291, 1993.
- [13] A. T. McCray. UMLS semantic network. In *Proceedings of the Thirteenth Annual SCAMC*, pages 503–507, 1989.
- [14] A. T. McCray. Representing biomedical knowledge in the UMLS semantic network. In N. C. Broering, editor, *High-Performance Medical Libraries: Advances in Information Management for the Virtual Era*, pages 45–55. Meckler, Westport, CT, 1993.
- [15] A. T. McCray, A. Burgun, and O. Bodenreider. Aggregating UMLS semantic types for reducing conceptual complexity. In *Proceedings of the Medinfo 2001*, pages 171–175, London, UK, September 2001.
- [16] A. T. McCray and W. T. Hole. The scope and structure of the first version of the UMLS semantic network. In *Proceedings of the Fourteenth Annual SCAMC*, pages 126–130, 1990.
- [17] Y. Peng, M. Halper, Y. Perl, and J. Geller. Auditing the UMLS for redundant classifications. In *Proc. 2002 AMIA Annual Symposium*, pages 612–616, San Antonio, TX, Nov. 2002.
- [18] Y. Perl, Z. Chen, M. Halper, J. Geller, L. Zhang, and Y. Peng. The cohesive metaschema: a higher-level abstraction of the UMLS semantic network. *Journal of Biomedical Informatics*, 35(3):194–212, June 2003.

- [19] P. L. Schuyler, W. T. Hole, M. S. Tuttle, and D. D. Sherertz. The UMLS metathesaurus: Representing different views of biomedical concepts. *Bull Med Libr Assoc*, 81(2):217–222, 1993.
- [20] M. S. Tuttle, D. D. Sherertz, N. E. Olson, M. S. Erlbaum, W. D. Sperzel, L. F. Fuller, and S. J. Nelson. Using meta-1 the first version of the UMLS metathesaurus. In *Proceedings of the Fourteenth Annual SCAMC*, pages 131–135, 1990.
- [21] US Dept. of Health and Human Services, NIH, National Library of Medicine. *Unified Medical Language System*, 2001.