Research Paper ■

Auditing the Unified Medical Language System with Semantic Methods

JAMES J. CIMINO

Abstract **Objective:** The National Library of Medicine's (NLM) Unified Medical Language System (UMLS) includes a Metathesaurus (Meta), which is a compilation of medical terms drawn from over 30 controlled vocabularies, and a Semantic Net, which contains the semantic types used to categorize Meta concepts and the semantic relations to connect them. Meta has been constructed through lexical matching techniques and human review. The purpose of this study was to audit the Meta using semantic techniques to identify possible inconsistencies.

Methods: Five different techniques were applied: (1) detection of ambiguity in Meta concepts with two or more semantic types, (2) detection of interchangeable keyword synonyms, (3) detection of redundant pairs of Meta concepts (using lexical matching combined with keyword synonyms), (4) detection of inconsistent parent–child relationships in Meta (based on the semantic type information), and (5) discovery of pairs of semantic types for which relations could be added to the Semantic Net, based on ''other'' relationships between Meta concepts.

Results: Of 57,592 concepts with multiple semantic types, 1817 (3.2%) were judged to be ambiguous. Keyword analysis showed 7121 pairs of interchangeable words. Using the keyword pairs, 5031 pairs of potentially redundant concepts were suggested, of which 3274 (65.1%) were judged to actually be redundant. Review of the 100,586 parent–child relationships revealed 544 (0.54%) that were incorrect. Review of the 219,664 ''Other'' relationships suggested 1299 places in the Semantic Net where relations between pairs of semantic types could be added.

Conclusion: Semantic techniques, alone or in combination, can be used to audit the UMLS to detect inconsistencies that are not detectable through lexical techniques alone. Use of these methods to augment the UMLS maintenance process will lead to improvement in the UMLS.

■ **JAMIA.** 1998;5:41–51.

The National Library of Medicine's (NLM) Unified Medical Language System (UMLS) consists of a set of Knowledge Sources that provide information about medical terminologies.[1] The largest of the Knowledge Sources is the Metathesaurus (Meta),[2] which is a compilation of medical terms drawn from over 30 controlled vocabularies.[3] Meta is more than a simple concordance of terms: Its developers strive to provide a concept-oriented organization in which synonymous terms from disparate source vocabularies map to the same concepts. Meta was constructed, and continues to be maintained, using lexical matching techniques (that is, using similarities in term names to identify possible synonyms), followed by human review.[4] Since its original version in 1990, Meta has grown from 208,559 term names (mapped to 64,123 concepts) to 740,170 term names (mapped to 331,756 concepts) in 1997.

Automated manipulation of conceptual information is inherently difficult when methods rely solely on lexical input. Although the lexical methods used in constructing Meta are among the most sophisticated in existence, they are necessarily limited by the inconsis-

Affiliation of the author: Department of Medical Informatics, Columbia University College of Physicians and Surgeons, New York, NY.

This work was supported by contracts for High-Performance Computing and Communication (N01-LM-4-3513) and Electronic Medical Records (LM05857) from the National Library of Medicine.

Correspondence and reprint requests: James J. Cimino, Associate Professor of Medical Informatics in Medicine, Atchley Pavilion Room 1310, 161 Fort Washington Avenue, New York, NY 10032. e-mail: ⟨James.Cimino@columbia.edu⟩.

Received for publication: 11/27/96; accepted for publication: 8/14/97.

tencies of natural language. The purpose of human review is to identify and correct errors (of commission and omission) produced by the automatic lexical tools. This review process is also imperfect, however, because of human limitations (such as memory capacity and inattention) and variation among reviewers, compounded by the huge volume of information needing review in Meta. Comprehensive, high-quality human review of the entire Meta is probably not achievable with the resources available. However, limited human review might be made more effective if automated techniques can help identify problem areas on which reviewers' attention can be focused.

Since terms in Meta are combined into concepts based on their meanings, it is appropriate to consider the use of semantic methods for identifying term synonymy. A variety of semantic methods for vocabulary mapping[5–7] and construction[8–13] have been described and applied to the vocabularies of multi-purpose clinical information systems.[14–16] These methods have generally relied on explicit inter-concept semantic relationships. For example, suppose disease terms are explicitly defined by linking them to terms representing their anatomic sites and etiologies. The terms "Ornithosis" and "Psittacosis" could then be recognized as synonymous, despite their lexical dissimilarity, since both would be linked to "Lung" and "Chlamydia Psitticae."

Meta does not contain such explicit, definitional concept interrelationships. There is, however, a great deal of subtle, indirect semantic information that can be teased out of Meta. For example, each concept in Meta is identified as having one or more semantic types taken from the UMLS Semantic Net (the second UMLS Knowledge Source).[17] The Semantic Net, in turn, provides information about the potential semantic relations that may exist between Meta concepts. Such information can be used to help understand the meanings of concepts in Meta through formal application of methods to detect possible inconsistencies. This paper describes several such methods, presents the results of applying these methods to the 1995 UMLS, and discusses the potential for semantic approaches to augment the maintenance process of Meta.

Methods and Results

This paper focuses on five different methods for using semantic information to audit the UMLS: (1) using semantic types to detect ambiguity, (2) using synonymy to compile keyword synonyms, (3) using semantic types and keyword synonyms to detect redundancy, (4) using semantic types to detect inconsistent parent–child relationships, and (5) using nonhierarchic ("other") relationships to suggest new semantic relations for the Semantic Net. I have implemented each method using simple string-parsing and string-matching algorithms written in MUMPS (Datatree, Waltham, MA), running on a desktop 90-MHz IBM Pentium PC. I then applied each method to the January 1995 version of the UMLS Knowledge Sources, obtained on CD-ROM from the NLM.

Detection of Ambiguity in Meta Using Semantic Types

As previously noted, the NLM assigns one or more semantic types to each concept in Meta. There are 133 semantic types in the Semantic Net, arranged in a hierarchy. These types are not necessarily mutually exclusive; in fact, many might be considered "multi-axial," making it almost mandatory for concepts to have several assigned types. For example, 14 semantic types appear in the hierarchy under the type Chemical Viewed Structurally and another 12 appear under the type Chemical Viewed Functionally. Chemical concepts in Meta often have one type assigned from each group (e.g., "Glucose" is assigned the type Carbohydrates from the structural group and Biologically Active Substance from the functional group). Many types seem to be mutually exclusive, however. For example, one would expect that any concept with the semantic type Organ or Tissue Function would not also have the semantic type Diagnostic Procedure. In an early version of Meta, however, the concept "Cardiac Output" had both types assigned to it. The NLM recognized this as an ambiguity in that two meanings had been ascribed to a single Meta concept. In subsequent versions, this ambiguity was resolved by assigning "Cardiac Output" a single semantic type (Diagnostic Procedure).

Of the 222,927 concepts in the 1995 Meta, 57,592 are assigned two or more semantic types. An examination of these concepts shows that there are 813 pairings, or semantic types that occur together. For 278 pairs, there is only one representative Meta concept. For example, the pairing of semantic types Embryonic Structure and Food occurs only in the concept "Egg Yolk." At the other extreme, the pairing of semantic types Organic Chemical and Pharmacologic Substance occurs in 26,003 Meta concepts.

I reviewed the 813 pairs of semantic types together with the concepts associated with each pairing and the definitions of each semantic type. I subjectively considered 660 pairs as not mutually exclusive. For example, since the concept "Egg Yolk" represents an embryonic structure that is also a food, the semantic types assigned to it seem appropriate and therefore

Table 1 ■

Examples of Mutually Exclusive Semantic Types (T Codes) and Their Assigned Meta Concepts (C Codes)

T002 Plant and T047 Disease or Syndrome
C0035510: Toxicodendron
T003 Alga and T009 Invertebrate
C0008155: Chlamydomonas
C0015154: Euglena
C0015155: Euglena gracilis
T009 Invertebrate and T047 Disease or Syndrome
C0002642: Amoeba
C0030499: Parasitic Diseases
C0241738: WORM
T020 Acquired Abnormality and T061 Therapeutic or Preventive Procedure
C0003853: Arteriovenous Anastomosis
C0009410: Colostomy
C0016200: Flaps
T024 Tissue and T060 Diagnostic Procedure
C0150887: Synovial Biopsy
T031 Body Substance and T059 Laboratory Procedure
C0017036: SERUM GAMMA-GLUTAMYLTRANSFERASE TESTS
C0017515: GGTP
C0036772: SERUM ACID PHOSPHATASE TESTS
T047 Disease or Syndrome and T109 Organic Chemical
C0032624: Polyvinyl Chloride
T101 Patient or Disabled Group and T184 Sign or Symptom
C0001917: Albino
T106 Element or Ion and T119 Lipid
C0050279: 9-tellurium Te 123m heptadecanoic acid
C0058501: ditrastic
C0060100: fatty ozonides
T108 Inorganic Compound and T109 Organic Chemical
C0050435: Acetard
C0050890: adsorgan
C0051166: Alka-Seltzer

allowable. Similarly, many organic chemicals are certainly pharmacologic substances. I judged the remaining 153 pairings to be mutually exclusive, based on the concepts assigned these pairs. These pairs are assigned to 1817 Meta concepts, representing 3.2% of the concepts that have multiple types and 0.82% of all Meta concepts. Examples of these pairs (Table 1) suggest that some are cases of incorrect semantic type assignments, whereas others are cases of ambiguity.

Compilation of Keyword Synonyms

The lexical process used to map synonymous terms to the same concepts in Meta matches character strings that are exact or similar using a variety of techniques. For example, by ignoring punctuation and word order, a lexical process will match the terms "Congestive Heart Failure" and "Heart Failure, Congestive." It might not, however, recognize that the term "Cardiac Failure, Congestive" should also match, since "cardiac" and "heart" are lexically dissimilar. In this paper, such words are referred to as keyword synonyms, with the recognition that they may not always be true synonyms in the usual sense. Instead, they are defined as words used interchangeably in synonymous strings. If such words can be identified when strings are known to be synonymous, the detection of these words in other strings may allow detection of synonymy that escapes the usual lexical matching process. In the above example, the words "cardiac" and "heart" are not always interchangeable, but may be helpful for enhancing the lexical matching process, which will, in any case, be subject to human review.

The process for detecting such keyword synonyms is relatively straightforward: Given a pair of strings known to be synonymous, remove all the identical words, or known keyword synonyms, and list the remaining words as potentially synonymous. After excluding foreign language terms and lexical variants based on word order, I performed this comparison on each pair of strings associated with the same Meta concept (a total of 100,458 comparisons). The process isolated 15,900 word pairs, and from these I selected 9650 pairs that I believed would be helpful for subsequent string comparisons. When pairs had one or both words in common with other pairs, I merged them into sets of interchangeable keywords, resulting in 8087 sets (ranging from 7121 pairs up to one set with 14 words). Some examples of these sets are shown in Table 2.

Detection of Redundancy among Meta Concepts

The lexical matching processes used by the NLM (developed by Lexical Technologies, Inc., Alameda, CA) are among the most sophisticated in existence. I chose a simpler approach which, while almost certainly less powerful, sought to determine whether the use of keyword synonyms could detect matching terms that might have escaped the standard process. In this approach, a program created a word index for Meta strings, ignoring all punctuation and word case. The program then collapsed the index on the basis of keyword synonym sets (e.g., the index lists for "Kidney," "Kidneys," and "Renal" were merged into a single list for "Kidney"). The program used this index to compare each Meta string with every other concept that had words in common. If all the normalized words in a given string from one concept were found to be associated with a second concept, the program considered the first concept to be "contained in" the second.

If it found that the second concept had a string "contained in" the first, the program considered the two concepts to be potential matches. The Appendix shows an example of this process.

This method, a mixture of semantic techniques (using keyword synonyms) and lexical ones (using string matching), is extremely sensitive to matching concepts with similar words. The results, however, are prone to have poor specificity. The program therefore includes an additional semantic technique: filtering potential matches on the basis of compatible semantic types. The criteria for compatibility are the same as those used to detect ambiguous concepts with multiple semantic types. For example, the filter will consider two concepts a match if they have semantic type Disease or Syndrome and mutually satisfy the "contained in" criterion, but if one concept has semantic type Disease or Syndrome and the other has semantic type Organism, the filter will exclude the match.*

*This filter is merely consistent with the test for ambiguity previously described: If such concepts were deemed synonymous and merged into one, a subsequent search for ambiguity would flag the merged concept as having incompatible semantic types.

The matching process involved 9,294,483 pair-wise string comparisons and identified 701,493 unidirectional string matches, of which 5,031 were bidirectional (mutual) matches. I reviewed these pairs to determine whether the method could be used to detect redundancy among Meta concepts. In the process, I found recurring reasons for true and false positives that suggested possible ways to improve the technique.

Overall, the keyword synonyms appeared to be useful. In many cases, the method identified pairs of single word concepts that consisted of keyword synonyms (such as "Dyspepsia" and "Indigestion," "Fever" and "Hyperpyrexia," "Furunculosis" and "Boils," "Oceans" and "Seas," "Ornithosis" and "Psittacosis," "Sports" and "Athletics," and "Montmorrillonite" and "Montmorillonite"). In other cases, pairs of multiple-word strings that differed by only a keyword were discovered (such as "Anthrax, Pulmonary" and "Lung, Anthrax," "Blurred Vision" and "Vision Blurring," "Duodenal Ulcer" and "Duodenal Ulceration," "Food Habits" and "Dietary Habits," "Herpes Genitalis" and "Herpes Genital," "Myocardial Infarction" and "Heart Infarction," "Pulmonary Alveolar Proteinosis" and "Lung, Alveolar Protei-

Table 2 ■

Examples of Interchangeable Keywords Found Through Examination of Synonymous Terms in Meta*

3	III	Three			
Abstinence	Sobriety	Temperance			
Acute	Sudden				
Adhesive	Glue	Paste			
Adrenaline	Epinephrine				
Alba	Caucasian	Caucasoid	Pale	White	
Arachnid	Arachnida	Spider	Spiderlike		
Arrhythmia	Arrythmia	Dysrhythmia	Dysrrhythmia		
Atomic	Nuclear				
Azotemia	Uremia	Uremic			
Bean	Legume	Leguminosae			
Belching	Burping	Eructation			
Bilharzia	Schistosoma				
Brain	Cephalic	Cereb	Cerebral	Cerebrum	
Cardiac	Cor	Coronary	Heart	Myocardial	Myocardium
Cerumen	Earwax	Wax			
Contagiosum	Contagious	Infected	Infection	Infectious	Infective
Erythema	Red	Erythematous	Redness	Rubor	
Essential	Primary				
Familial	Family	Genetic	Hereditary	Inherited	
Febrile	Fever	Hyperpyrexia	Hyperthermia	Pyrexia	Pyrexial
Ferric	Iron				
Firearm	Gun	Handgun	Pistol		
Freidreich	Freidriech	Friedreich	Friedriech		
Gangrene	Gangrenous	Necroses	Necrosis	Necrotic	Rot
Hive	Urticaria	Welt			
Kidney	Renal				
Marine	Ocean	Sea			
Mice	Mouse				
Pharangeal	Pharyngeal	Pharynges	Pharynx	Throat	
Radiogr	Radiographic	Xray	Radiography		

*Most of these sets include plural forms, which have been omitted here.

nosis,'' ''Stomach Contents'' and ''Gastric Contents,'' ''Bicillin-3'' and ''Bicillin III,'' ''Ribonuclease B'' and ''RNase B,'' ''T4 Sulfate'' and ''Thyroxine Sulfate,'' and ''Ill Feeling'' and ''Feeling Unwell''). In some cases, matches were identified through the use of multiple keyword synonyms (such as ''Liver, Benign Tumor'' and ''Hepatic Neoplasm Benign,'' ''Heart Rupture, Post-Infarction'' and ''Myocardial Rupture (Post Infarct),'' ''Renal Infarct'' and ''Kidney, Infarction,'' ''Postural Proteinuria'' and ''Albuminuria, Orthostatic,'' and ''Ischemia, Bowel'' and ''Intestinal Ischemia'').

As described above, the matching method did not require strings to be mutually inclusive. The finding that all the words of one string in one concept were found in any of the strings in a second concept, and vice versa, was sufficient. As a result, some matches could not be discovered by simple pair-wise string comparison. For example, all the words in the concept ''Adrenal Hypercorticism'' were found either in the preferred name for a second concept, ''Adrenal Gland Hyperfunction,'' or in its synonym ''Hypercorticism.'' This synonym of the second concept was, of course, completely contained in the preferred name of the first concept.

Finally, the insensitivity of the method to word order and punctuation (with or without keyword synonyms) permitted the discovery of many redundant concepts such as ''Abnormal Pap Smear'' and ''Pap Smear Abnormal.'' Table 3 lists some additional examples. Over all, after reviewing the 5031 pairs of proposed matches, I found 3274 (65.1%) pairs that I believed to represent redundant concepts, representing 2.94% of the concepts in Meta. I considered the remaining 1757 pairs to represent false positives, which appeared to be due to five different causes.

Table 3 ■

Examples of True Positive Redundancy Among Meta Concepts, Based on Word Order and Punctuation Differences

C0000760: ABNORMAL PAP SMEAR C0240660: PAP SMEAR ABNORMAL
C0002965: Angina, Unstable C0235466: ANGINA UNSTABLE
C0016481: Food poisoning, bacterial C0178496: bacterial food poisoning
C0018572: Hand, Foot and Mouth Disease C0238150: HAND-FOOT-AND-MOUTH DISEASE
C0030321: Panic disorder, without agoraphobia C0236794: Panic Disorder Without Agoraphobia
C0051848: angiotensin I, Sar(1)- C0089853: 1-Sar-angiotensin I
C0061996: guar gum C0120507: gum guar
C0062157: heat stable toxin (E coli) C0115249: E coli heat stable toxin
C0065828: measles, mumps, rubella vaccine C0244995: mumps-measles-rubella vaccine
C0111956: D&C Yellow No 10 C0112178: D.C. Yellow No. 10
C0158646: Cleft palate and cleft lip C0221728: PALATE AND LIP CLEFT
C0176975: LYSIS OF PERITONEAL ADHESIONS C0198664: Lysis of adhesions of peritoneum
C0206167: Grants, Peer Review C0206168: Peer Review, Grants
C0212986: 4-hydroxy-7,8,11,12,15,7′,8′,11′,12′,15′-decahydro-beta,psi-carotene C0212988: 4-hydroxy-7,7′,8,8′,11,11′,12,12′,15,15′-decahydro-beta,psi-carotene
C0216838: Na(+)-H(+) exchanger-2 C0216839: Na-H exchanger 2
C0232495: Lower abdominal pain C0235298: ABDOMINAL PAIN LOWER

False-positive Matches

One reason for the occurrence of false-positive matches was the fact that keyword synonyms are not truly synonymous in all contexts. The most glaring example of this weakness of the approach was with the keyword synonyms for ''cancer,'' which include ''cancers,'' ''carcinoma,'' ''carcinomas,'' ''malignancy,'' ''malignancies,'' ''neoplasm,'' ''neoplasms,'' ''neoplastic,'' ''neopl,'' ''neopls,'' ''tumor,'' ''tumors,'' and ''tumour.''† For example, while the process detected ''Larynx Neoplasm Malignant'' and ''Laryngeal Cancer'' (an appropriate match, in my judgment), it also matched ''Esophageal Neoplasms'' and ''Oesophageal Carcinoma'' (an inappropriate match because there are noncarcinomatous esophageal neoplasms). In fact, of the 204 matched pairs that contained one of these keywords, I judged only 51 (25%) to be true positives. This suggests that some of the keyword synonym sets may be less useful than others, or perhaps useful only with certain semantic types. In this case, the ''cancer'' keyword synonym set, which was generated automatically, might be more useful if divided manually

†Obviously, these words are not freely interchangeable, since neoplasms and tumors may or may not be malignant. Since these words were used interchangeably in many Meta concepts, however, they were retained as keyword synonyms.

into two sets (''cancer,'' ''cancers,'' ''carcinoma,'' ''carcinomas,'' ''malignancy,'' and ''malignancies'' in one set,‡ and ''neoplasm,'' ''neoplasms,'' ''neoplastic,'' ''neopl,'' ''neopls,'' ''tumor,'' ''tumors,'' and ''tumour'' in the other).

Another common keyword-synonym pattern that yielded false-positive matches was based on singular–plural forms. Many of the concepts that matched were singular and plural chemical concepts that were, in reality, instances and classes, respectively. For example, the process matched ''Acetic Acid'' and ''Acetic Acids,'' where the former concept is a specific compound and the latter a class of compounds that includes the former concept as well as others (such as ''Nitroacetic Acid''). Further analysis of such occurrences may show that singular–plural forms of keyword synonyms should not be used in particular cases, as with chemicals or when a parent–child relationship exists in Meta for the two matched concepts.

Many false-positive matches occurred because the mutual-inclusion process matches strings to concepts rather than matching strings to strings. For example, the preferred concept name ''ADTN'' is contained in the preferred concept name ''5,6-ADTN''—a one-way match. When the latter string is compared to the former concept, the ''ADTN'' is found to be included in the preferred name and the ''5'' and ''6'' are found in the former concept's synonym, ''6-amino-5,6,7,8-tetrahydro-2,3-naphthalenediol''—a two-way match. In reality, however, the two concepts are not synonymous. This was a common occurrence in matching chemical names; however, the method was significantly sensitive to make it a useful, though inexact, approach.

Another problem occurred because the matching was done on a word-by-word basis with no consideration to repeated words, so the preferred concepts ''AMP-activated protein kinase'' and ''AMP-activated protein kinase kinase'' were considered mutually inclusive. Fortunately, this was an uncommon occurrence, and in some cases (such as ''Larynx Neoplasm Malignant'' and ''Laryngeal Cancer'') the match was in fact a true positive.

Finally, word-order insensitivity was often a problem. Some matches were successfully filtered because of differences in the semantic types (such as ''Nursing Home'' and ''Home Nursing,'' and ''Renal Pelvis'' and ''Pelvic Kidney''). With chemical names, however, the technique was less specific, since the meanings of many chemical names are order-sensitive (e.g., ''Pro-Phe-Arg-CH2-Cl'' and ''Phe-Pro-Arg-CH2-Cl''). Again, this process produced enough true positives, even among chemicals (e.g., ''Angiotensin I, Sar(1)-'' and ''1-Sar-angiotensin I''), to still be useful.

‡Even the inclusion of ''carcinoma'' and ''carcinomas'' may be too permissive, since they refer to specific, epithelial forms of malignancy.

Table 4 ■

Incorrect Relationships in Pairs of Parent–Child Terms in Meta*

C0001973: Alcoholism {Mental or Behavioral Dysfunction}
C0019187: Hepatitis, Alcoholic {Disease or Syndrome}
C0004611: Bacteria ⟨1⟩ {Bacterium}
C0038027: Spores {Organism}
C0005538: Biomedical and Dental Materials {Biomedical or Dental Material}
C0007245: Cariogenic Agents {Chemical Viewed Functionally}
C0008532: Christianity {Idea or Concept}
C0242823: Saints {Conceptual Entity}
C0021213: Indicators and Reagents {Indicator or Reagent}
C0011740: Detergents {Chemical Viewed Functionally}
C0021521: Inorganic Chemicals {Inorganic Chemical}
C0017110: Gases {Chemical Viewed Structurally}
C0021948: Invertebrates {Invertebrate}
C0003064: Animals, Laboratory {Animal}
C0021948: Invertebrates {Invertebrate}
C0008485: Chordata {Animal}
C0025351: Mental Disorders {Mental or Behavioral Dysfunction}
C0021116: Impotence {Disease or Syndrome}
C0028214: Nitrous Acid {Inorganic Compound}
C0028137: Nitrites {Chemical Viewed Structurally}

*The semantic type for each term appears in braces.

Detection of Inconsistent Parent–Child Relationships in Meta

As previously mentioned, Meta does not contain a rich set of explicit, definitional semantic relations among its concepts. It does, however, contain the semantic relationships ''broader–narrower,'' ''parent–child,'' and ''other.'' The parent–child relations are drawn from hierarchic information occurring among terms in the source vocabularies. For example, if Term A is the parent of Term B in some source vocabulary, then the concepts to which these terms are assigned in Meta will have a parent–child relationship as well. If these parent–child relationships can be considered ''is-a'' hierarchic relationships, it follows that the semantic types of the children concepts should be the same as, or subtypes of, the semantic types of the par-

ent concepts. For example, if concept A has the semantic type Chemical, then its child, Concept B, should have the semantic type Chemical or a subordinate semantic type, such as Organic Chemical.

From the 100,586 parent–child relationships in Meta, there were 544 pairs (0.54%) of concepts for which the semantic type of the child was neither the same as, nor a more specific form of, the semantic type of the parent. One possible explanation was that the Semantic Net was missing "is-a" relationships between the semantic types in question. After examining each pair, however, I concluded that none of the pairs suggested such additions to the Semantic Network. Instead, I found that 22 pairs represented incorrect parent–child relationships in Meta. For example, Meta lists the concept "Inorganic Chemicals," with the semantic type Inorganic Chemical, as a parent of the concept "Gases," with the semantic type Chemical Viewed Structurally (see Table 4). In the remaining 522 pairs, I judged the semantic type assigned to one or both concepts to be incorrect. As shown in Table 5, a common problem was that the semantic type assigned to the child concept was less specific than the type assigned to the parent. For example, "Electrodes" has semantic type Medical Device, yet its child "Electrodes, Implanted" has the less specific semantic type Manufactured Object. In this case, I suggest that "Electrodes, Implanted" should also be assigned the semantic type Medical Device.

Discovery of Potential Semantic Relations for the Semantic Network

As mentioned above, Meta concepts are associated with each other through a variety of relations. One of these, the "other" relation, seems to imply that a semantic relationship exists but is as yet unnamed. The semantic types in the UMLS Semantic Net are related to each other through explicit semantic links. It is reasonable, therefore, to infer that an "other" link found in Meta may actually be an example of a semantic relationship found in the UMLS Semantic Net. For example, if the concept "Borrelia Burgdorferi" is related to the concept "Lyme Disease," one might infer that the former causes the latter, since the Semantic Net shows us that concepts of type "Bacterium" can be related to concepts of type "Disease or Syndrome" via the causes relationship. (Of course, one might also conclude that "Borrelia Burgdorferi" is affected by "Lyme Disease" since, in the Semantic Net, Bacterium (like other Organisms) may be affected by a Disease or Syndrome.) Sometimes, however, an "other" relation between two concepts in Meta is not explained by a relationship between the semantic types assigned to the two concepts. This suggests that the relations found in Meta can be used to enhance the Semantic Network.

Meta contains 219,664 "other" relations between concepts. Examination of the semantic types of these concepts shows 92,487 cases in which the relation is not explained by the Semantic Net (using the semantic relationships listed in the SRSTRE2 file from the UMLS CD-ROM). When viewed as examples of possible relations for the Semantic Net, these cases represent 1299 different type–type relationships. For some 308 relationships, only one example could be found in Meta. For example, there is no relationship between the type Mental or Behavioral Dysfunction and the type Activity; however, there is one "other" relation between concepts of these types: "Agraphia" and "Handwriting." This example suggests that a semantic relationship, such as Mental or Behavioral Dysfunction–affects–Activity should be added to the Semantic Net.

At the other extreme, the process found 6018 cases of "other" relations between concepts of semantic types Disease or Syndrome and Quantitative Concept. For example, Meta lists an "other" relationship between

Table 5 ■

Incorrect Semantic Types in Pairs of Parent–Child Terms in Meta*

C0002808: Anatomy ⟨1⟩ {Biomedical Occupation or Discipline}
C0002812: Anatomy, Regional {Occupation or Discipline}
C0004245: Atrioventricular Block {Disease or Syndrome}
C0085614: First degree AV block {Pathologic Function}
C0005528: Biological Transport {Organism Function}
C0005529: Biological Transport, Active {Physiologic Function}
C0008031: Chest Pain {Sign or Symptom}
C0235718: CHEST PAIN PRECORDIAL {Finding}
C0011331: Dental Care {Therapeutic or Preventive Procedure}
C0206196: Dental Care for Chronically Ill {Health Care Activity}
C0013473: Eating Disorders {Mental or Behavioral Dysfunction}
C0031873: Pica {Disease or Syndrome}
C0013812: Electrodes {Medical Device}
C0013814: Electrodes, Implanted {Manufactured Object}
C0026213 Miscellaneous Drugs and Agents {Chemical Viewed Functionally}
C0023688: Ligands {Chemical}
C0150903: Laboratory Findings: Urine {Laboratory or Test Result}
C0151430: Granular casts in urine {Finding}
C0161860: OPERATIONS ON CORNEA {Therapeutic or Preventive Procedure}
C0161861: Repair of cornea, NOS {Health Care Activity}

*The semantic type for each term appears in braces.

"Corneal Diseases" and "Mortality ⟨2⟩." This suggests a relationship such as Disease or Syndrome–is measured by–Quantitative Concept.

In some cases, the pairing was the result of an incorrect assignment of a semantic type to a concept. For example, the concept "Industrial Microbiology" has an "other" relation to the concept "Genetic Engineering." For some reason, "Industrial Microbiology" has been assigned the semantic type Organism, serving as an example of the pairing of the types Organism and Molecular Biology Research Technique. Since the semantic type assignment seems wrong, the implication of the "other" relation for the Semantic Net is unclear.

It is beyond the scope of this current work to determine which of the 92,487 unexplained "other" relationships are due to mistakes in Meta and which are actually suggesting additions to the Semantic Net. However, a semantic view of the "other" relationships would suggest that they (a) have some meaning and (b) this meaning is not explained by the Semantic Net. Therefore, my analysis simply suggests 1299 locations in the Semantic Net where relations could be added and provides examples of each.

Discussion

This study makes use of a combination of automated lexical and semantic methods for detecting problems in Meta (ambiguity and redundancy) and the Semantic Net (inconsistencies and missing relations). The approach is a modest one, considering that the lexical methods are simpler than those used by the UMLS developers; the semantic methods are imprecise (since they are only as precise as the UMLS semantic information); and the manual methods are obviously operator-dependent, making them difficult to reproduce in a consistent manner. However, I believe that their ability to suggest alterations in the UMLS is evident from the examples shown in the Tables.

Strict lexical methods (ignoring punctuation and word order) were often sufficient for detecting Meta concept redundancy. Sensitivity was improved by the addition of semantic information—namely, the use of keyword synonyms. Specificity was also improved by the use of semantic information to filter potential matches that were of clearly different meanings, based on their semantic types. This combined lexico-semantic approach is similar to previous morpho-semantic methods,[18,19] although I used semantic information at the word level rather than the morpheme level (e.g., "-itis" = "inflammation"). This approach differs, however, from statistical decomposition methods, which require large sets of examples to detect significant evidence of potential relations.[20]

Other techniques relied solely on semantic information. To detect ambiguity, I simply retrieved all concepts with multiple semantic types. To detect problems with parent–child links, I compared the semantic types of the parent to those of the child, looking for discrepancies. To identify possible gaps in the Semantic Net, I compared "other" relationships in Meta with existing semantic relations in the Semantic Net. In each case, the automated method used semantic information to produce a data set, which was then subjected to manual review.

While the automated techniques are easily reproduced, the manual steps I have employed are subjective. Despite the qualitative aspect of these processes, however, performing them in an objective manner is possible. For example, the NLM could decide that certain pairs of semantic types (e.g., Disease or Syndrome and Organic Chemical) are mutually exclusive and that no concept in Meta should have both assigned. Their decisions about which pairs are or are not exclusive will be subjective and open to debate, but once these pairs are determined, they can be applied consistently. Similarly, the analysis of parent–child and semantic relations in Meta can be performed objectively, in much the same way as has previously been done in examining NLM's Medical Subject Headings (MeSH) to study hierarchic[21] and nonhierarchic[22] relationships.

Much of the work in building the UMLS consists of manual review of imprecise lexical matching. The semantic methods I have described are also imprecise, but I believe the data sets they produce will be complimentary to the current methods. Readers can judge for themselves whether there seems to be value in the data sets, based on the examples in the Tables. It would seem worthwhile for the NLM to consider reviewing such data sets if, for example, a set of 5031 matches reveals 3274 instances of redundancy.

The results of this study should not be considered a criticism, or even an evaluation, of the UMLS. A thorough study might show the UMLS in a very favorable light. For example, I found only 1817 examples of ambiguity in almost 58,000 places where such ambiguity might occur. This represents an occurrence rate of 3.2%—a small number. If it represents the only ambiguity in Meta, then the true occurrence rate is smaller still, at 0.82%. Similarly, I found 3274 examples of redundancy. This is 65.1% of the concept pairs I examined, but it is only 0.0035% of the pairs my computer examined. There are 222,927 concepts in the

1995 version of Meta, so there are almost 50 billion concept pairs that have the potential for redundancy. My findings, therefore, represent a paltry 0.000007% of these pairs.

The creation, review, and maintenance of the UMLS is a huge effort, requiring sophisticated automated processes and extensive manual review. The review process is extremely resource-intensive,[23] and NLM support for this activity is not unlimited. Therefore, any methods that can focus reviewers' attention on potentially troublesome or inconsistent UMLS content will help maximize the effectiveness of this resource. The sensitivity of the methods described in this paper is unknown. However, I believe that they are sensitive enough to find at least some potential problems worthy of review, and I believe that their specificity (and therefore the relevance of the results) is acceptable.

Conclusion

The UMLS will never be "finished," nor will it ever be "perfect." The maintenance process will continue and will strive toward perfection. It is already a labor-intensive process, and the developers apply automated lexical techniques where possible. My study shows that methods using readily available semantic information have the potential to support the process through semi-automated auditing techniques.

The author thanks Gai Elhanan and Carol Bean for their helpful comments in preparing the manuscript.

References ■

1. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993;32(4): 281–91.
2. Tuttle MS, Olson NE, Campbell KE, Sherertz DD, Nelson SJ, Cole WG. Formal properties of the Metathesaurus. In: Ozbolt JG (ed). Proc 18th Annu Symp Comput App Med Care. New York: McGraw-Hill, 1994;145–9.
3. National Library of Medicine. UMLS Knowledge Sources, Experimental Edition. Bethesda, MD: National Library of Medicine, 1995 (updated annually).
4. Tuttle MS, Sherertz D, Erlbaum M, Olson N, Nelson SJ. Implementing Meta-1: the first version of the UMLS Metathesaurus. In: Kingsland LC (ed). Proc 13th Annu Symp Comput App Med Care, Washington, DC, November 1989. New York: IEEE Computer Society Press, 1989;483–7.
5. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. In: Proceedings of the American Association for Medical Systems and Informatics Congress, May 10, 1989. 113–117. Reprinted in MD Comput. 1990;7(2):104–9.
6. Masarie FE Jr, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24(4):379–400.
7. Rocha RA, Rocha BH, Huff SM. Automated translation between medical vocabularies using a frame-based interlingua. In Clayton PD (ed). Proc 15th Annu Symp Comput App Med Care. New York: McGraw-Hill, 1991;690–4.
8. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD. Designing an introspective, controlled medical vocabulary. In: Kingsland LC (ed). Proc 13th Annu Symp Comput App Med Care. New York: IEEE Computer Society Press, 1989;513–8.
9. Rector AL, Nowan WA, Kay S. Foundations for an electronic medical record. Methods Inf Med. 1991;30(3):179–86.
10. Cambell KE, Musen MA. Representation of clinical data using SNOMED III and conceptual graphs. In: Frisse ME (ed). Proc 16th Annu Symp Comput App Med Care. New York: McGraw-Hill, 1992;354–8.
11. Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a controlled medical vocabulary server: the VOSER project. Comp Biomed Res. 1994;27(6):472–507.
12. Cambell KE, Musen MA. A logical foundation for representation of clinical data. JAMIA. 1994;1(3):218–32.
13. Schoop M, Schoop D, Bernauer J. A classification manager for compositional concept systems exemplarily shown by the AO/ASIF classification of fractures of long bones. In: Kaihara S, Greenes RA (eds). Proc MEDINFO 95. Edmonton, AL, Canada: Healthcare Computing and Communications Canada, 1995;1655.
14. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB: Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA. 1994;1(1):35–50.
15. Rector AL, Nowlan WA. The GALEN Project. Comput Methods Programs Biomed. 1994;45(1–2):75–8.
16. Cimino JJ, Johnson SB, Hripcsak G, Hill CL, Clayton PD. Managing vocabulary for a centralized clinical system. In: Kaihara S, Greenes RA (eds). Proc MEDINFO 95. Edmonton, AL, Canada: Healthcare Computing and Communications Canada, 1995;117–20.
17. McCray A: The UMLS Semantic Network. In: Kingsland LC (ed). Proc 13th Annu Symp Comput App Med Care. New York: IEEE Computer Society Press, 1989;503–7.
18. Dujols P, Aubas P, Baylon C, Gremy F. Morpho-semantic analysis and translation of medical compound terms. Methods Inf Med. 1991;30(1):30–5.
19. Lovis C, Michel PA, Baud R, Scherrer JR. Word segmentation processing: a way to exponentially extend medical dictionaries. In: Kaihara S, Greenes RA (eds). Proc MEDINFO 95. Edmonton, AL, Canada: Healthcare Computing and Communications Canada, 1995;28–32.
20. Chute CG, Yang Y, Evans DA. Latent semantic indexing of medical diagnoses using UMLS semantic structures. In Clayton PD (ed). Proc 15th Annu Symp Comput App Med Care. New York: McGraw-Hill, 1991;185–9.
21. Hollander D, Greenes RA. Identification of semantic relations between child and parent MeSH terms in the MeSH tree structures: preliminary report on a first-pass analysis. Report to the National Library of Medicine Unified Medical Language System Project, Decision Systems Group, Brigham and Women's Hospital, Boston, 1988.
22. Bean CA. Analysis of non-hierarchical associative relationships among Medical Subject Headings (MeSH): anatomical and related terminology. In Green R (ed). Knowledge Organization and Change: Proc 4th Int ISKO Conf. Frankfurt am Main, Germany: The International Society of Knowledge Organization, 1996;80–6.
23. Tuttle MS, Suarez-Munist ON, Olson NE, et al. Merging terminologies. In Kaihara S, Greenes RA (eds). Proc MEDINFO 95. Edmonton, AL, Canada: Healthcare Computing and Communications Canada, 1995;162–6.

APPENDIX

Detection of Redundancy Through Keyword Synonym Matching

The matching algorithm uses a keyword index of Meta in which each word in each string is indexed, but the index is collapsed to merge the words on the basis of their keyword synonyms. The matching process is then carried out by comparing each string in Meta with every other concept that has at least one word in common. As an example, consider these six concepts from the 1995 Meta:

Concept ID	*Preferred Name*	*Synonym*
C0010762	Cytochrome P-450	Flavoprotein-Linked Monooxygenase
C0022658	Kidney Diseases	Nephropathy
C0111822	Cytochrome P-450 Monooxygenase	(none)
C0152100	Gouty Nephropathy	Gouty Kidney Disease
C0221739	Renal Disorder	(none)
C0238145	Gout, Renal Disease	Uric Acid Nephropathy

Consider, too, the following keyword synonyms, which have been identified through a separate process:

Keyword Synonym	*Preferred Form*
Diseases	Disease
Disorder	Disease
Gouty	Gout
Renal	Kidney

When each string is indexed and the keywords merged on the basis of keyword synonyms, the following table results:

Word	*Concept*
Acid	C0238145
Cytochrome	C0010762
Cytochrome	C0111822
Disease	C0022658
Disease	C0152100
Disease	C0221739
Disease	C0238145
Flavoprotein	C0010762
Gout	C0152100
Gout	C0238145
Kidney	C0022658
Kidney	C0152100
Kidney	C0221739
Kidney	C0238145
Linked	C0010762
Monooxygenase	C0010762
Monooxygenase	C0111822
Nephropathy	C0022658
Nephropathy	C0152100
Nephropathy	C0238145
P	C0010762
P	C0111822
Uric	C0238145
450	C0010762
450	C0111822

Each string is then examined, through the index, to find concepts that contain the words in the string. So, for ex-

ample, the string ''Kidney Diseases'' contains ''Kidney'' and ''Diseases'' and is therefore contained in C0221739 and C0238145. Examining all the strings provides the following matches:

String	*Source Concept*	*Matching Concepts*
Cytochrome P-450	C0010762	C0111822
Cytochrome P-450 Monooxygenase	C0111822	C0010762
Flavoprotein-Linked Monooxygenase	C0010762	(none)
Gout, Renal Disease	C0238145	C0152100
Gouty Kidney Disease	C0152100	C0238145
Gouty Nephropathy	C0152100	C0238145
Kidney Diseases	C0022658	C0221739, C0238145
Nephropathy	C0022658	C0152100, C0238145
Renal Disorder	C0221739	C0022658, C0238145
Uric Acid Nephropathy	C0238145	(none)

Finally, the algorithm examines the list for symmetric matches to suggest redundant Meta concepts:

C0010762 Cytochrome P-450	and	C0111822 Cytochrome P-450 Monooxygenase
C0022658 Kidney Diseases	and	C0221739 Renal Disorder
C0152100 Gouty Nephropathy	and	C0238145 Gout, Renal Disease

Note, however, that while C0022658 Kidney Diseases matches C0238145 Gout, Renal Disease, the reverse is not true, so the two concepts are not mutually inclusive and are not proposed as possible redundancies.