Research Paper ■

Integrating SNOMED CT into the UMLS: An Exploration of Different Views of Synonymy and Quality of Editing

KIN WAH FUNG, MD, MS, MA, WILLIAM T. HOLE, MD, STUART J. NELSON, MD, SURESH SRINIVASAN, MS, TAMMY POWELL, MLIS, MS, LAURA ROTH, MLS

Abstract **Objective:** The integration of SNOMED CT into the Unified Medical Language System (UMLS) involved the alignment of two views of synonymy that were different because the two vocabulary systems have different intended purposes and editing principles. The UMLS is organized according to one view of synonymy, but its structure also represents all the individual views of synonymy present in its source vocabularies. Despite progress in knowledge-based automation of development and maintenance of vocabularies, manual curation is still the main method of determining synonymy. The aim of this study was to investigate the quality of human judgment of synonymy.

Design: Sixty pairs of potentially controversial SNOMED CT synonyms were reviewed by 11 domain vocabulary experts (six UMLS editors and five noneditors), and scores were assigned according to the degree of synonymy.

Measurements: The synonymy scores of each subject were compared to the gold standard (the overall mean synonymy score of all subjects) to assess accuracy. Agreement between UMLS editors and noneditors was measured by comparing the mean synonymy scores of editors to noneditors.

Results: Average accuracy was 71% for UMLS editors and 75% for noneditors (difference not statistically significant). Mean scores of editors and noneditors showed significant positive correlation (Spearman's rank correlation coefficient 0.654, two-tailed $p < 0.01$) with a concurrence rate of 75% and an interrater agreement kappa of 0.43.

Conclusion: The accuracy in the judgment of synonymy was comparable for UMLS editors and nonediting domain experts. There was reasonable agreement between the two groups.

■ **J Am Med Inform Assoc.** 2005;12:486–494. DOI 10.1197/jamia.M1767.

Concept-based organization has been named as one of the desirable features of modern biomedical terminologies.[1] The ability to process a unit of meaning (concept) independently from the names used to describe it (variably called terms, strings, descriptions, etc., in different vocabulary systems; they will be referred to as concept names in this article) makes it easier to represent polysemy (one name with multiple meanings) and polyonomy (multiple names for one concept), which are common occurrences in biomedical nomenclature. More importantly, a concept can remain stable even when its names change over time.

Synonymy is a central notion in the organization of a concept-based terminology. Names that have the same meaning (i.e., synonyms) are encompassed by the same concept. In linguistics, synonymy can be defined as follows: X and Y are synonyms if any sentence S1 containing X is equivalent to another sentence S2, which is identical to S1 except that X is replaced by Y.[2] More simply put, X and Y are synonyms if they can be used interchangeably in all circumstances. For instance, "celiac disease" and "gluten enteropathy" are synonyms as the two sentences (and all other sentences differing only by the two concept names): "Celiac disease is a chronic familial disorder associated with sensitivity to dietary gluten" and "Gluten enteropathy is a chronic familial disorder associated with sensitivity to dietary gluten" are equivalent in meaning.[3] In biomedicine, common examples of synonymy include Anglo-Saxon names and their Latinate equivalents (e.g., "kidney stone" and "renal calculus"), abbreviations or acronyms (e.g., "AIDS" and "Acquired Immunodeficiency Syndrome"), and eponyms (e.g., "Wilson's disease" and "hepaticolenticular degeneration").

However, synonymy in practice is fuzzier. Meaning may have multiple aspects. In the search for synonyms, it is rare to find two names having identical meanings. It is more common to find names that overlap closely in meaning but are not

Affiliations of the authors: National Library of Medicine, Bethesda, MD (KWF, WTH, SJN, TP); and MSD Inc., Vienna, VA (SS, LR).

Supported in part by an appointment to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

The authors thank Betsy Humphreys, Olivier Bodenreider, and James Cimino for their advice and suggestions in the preparation of the manuscript. They also thank Olivier Bodenreider, James Cimino, Alexander Yu, and the UMLS clinical editors for completing the synonymy questionnaire.

Correspondence and reprints: Kin Wah Fung, MD, Building 38A, Room 9N904, MS54, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894; e-mail: <kwfung@nlm.nih.gov>.

Received for publication: 12/10/04; accepted for publication: 02/15/05.

equivalent in all situations (plesionymy). Thus, between true synonymy and nonsynonymy, there may be "relative synonyms," which even though they do not satisfy the strictest definition of true synonymy, are close enough in meaning to be considered "practically synonymous" in certain circumstances. In controlled vocabularies, relative synonyms may be treated as if they named exactly the same concept.

The rationale for doing this may be different for different vocabularies. For example, in SNOMED Clinical Terms (SNOMED CT),[4] "hernia" is considered a synonym of "hernia of abdominal cavity." In one sense, "hernia" is a broader concept than "hernia of abdominal cavity" because it could also refer to herniation in other parts of the body (e.g., cerebral herniation). However, in most clinical situations (the intended use cases for SNOMED CT), "hernia" refers to "hernia of the abdominal cavity." Therefore, listing the two names as synonyms is understandable and facilitates the process of clinical data recording.

Another example can be found in the controlled vocabulary used for the Physician Data Query Online System (PDQ)[5] developed by the U.S. National Cancer Institute. In this vocabulary, "pain therapy" is considered a synonym of "cancer pain management." Semantically, "cancer pain management" is a subtype of "pain therapy" and not its synonym. However, in the context of PDQ, which is primarily restricted to cancer-related information, it is not unreasonable to treat them as synonymous.

Background

Integration of SNOMED CT into the Unified Medical Language System

The lack of a standard clinical vocabulary has long been identified as a major impediment to more widespread deployment of electronic medical records.[6] To encourage the use of common medical terminology within the U.S. health information systems, the Department of Health and Human Services reached an agreement with the College of American Pathologists in 2003 to make SNOMED CT available to U.S. users at no cost within the National Library of Medicine's Unified Medical Language System (UMLS). With more than 350,000 concepts, 950,000 English descriptions (concept names) and 1,300,000 relationships, SNOMED CT is the largest single vocabulary ever integrated into the UMLS.

Differing Views of Synonymy

In integrating a source vocabulary into the UMLS, a major part of the editing work is to organize the concepts and concept names in the source vocabulary in relation to the existing UMLS concept structure. This is a two-step process. First, the incoming concepts and concept names are aligned with existing UMLS content algorithmically based on normalized string matching and other nonlexical information in the source. This is followed by manual review by UMLS editors. When the source vocabulary is already concept-based like SNOMED CT and its content overlaps substantially with existing content in the UMLS, this is a process of aligning two views of synonymy: the source view and the UMLS view, which may not be in total agreement. In the case of the integration of SNOMED CT, there was concurrence between the SNOMED CT and UMLS views of synonymy for the majority of concepts (86% of all SNOMED CT concepts involved). Concurrence means that all the concept names that SNOMED CT considered as synonymous were put into the same UMLS concept while concept names that were not synonymous in SNOMED CT were put into different UMLS concepts. In the remaining 14% of SNOMED CT concepts, the two views of synonymy were different; either concepts that were not synonymous in SNOMED CT were put together in the same UMLS concept or names that were considered synonymous in SNOMED CT were split across different UMLS concepts. These two types of difference are discussed in more detail.

Synonymy in the Unified Medical Language System View But Not in the SNOMED CT View

Cases in which two or more SNOMED CT concepts were merged into the same UMLS concept represented the bulk of the disagreement between the SNOMED CT and the UMLS views of synonymy, involving 13.4% of SNOMED CT concepts. Merging of two or more SNOMED CT concepts in the UMLS occurred for a variety of reasons, and some common ones are described below. It is worth mentioning that the principles behind the merging are not specific to SNOMED CT but are also used in the editing of other source vocabularies in the UMLS.

In SNOMED CT, because of the strict separation between hierarchies (no single concept can belong to two hierarchies) and the use of Description Logic,[7,8] some concepts that were very close in meaning were considered distinct concepts. For example, "stab wound (morphologic abnormality)" and "stab wound (disorder)" were two discrete concepts belonging to different hierarchies ("body structure" and "clinical finding" hierarchies respectively). In SNOMED CT's Description Logic, the concept "stab wound (morphologic abnormality)" formed part of the definition of "stab wound (disorder)" by the relationship "associated morphology." While the two concepts can be differentiated on a theoretical level, in most clinical situations, this distinction is probably neither necessary nor helpful. For those reasons, the National Library of Medicine (NLM) decided to merge them into the same UMLS concept. Similar merges occurred across other SNOMED CT hierarchies. A summary of the most common types of such merges, their frequencies, and specific examples is given in Table 1. All the counts and examples in this article were based on the 2004AA release of UMLS, which contained all the active content of the January 31, 2004, release of SNOMED CT.

Another reason for merging SNOMED CT concepts was an exceptionally fine level of granularity in some SNOMED CT concepts. For example, there were 16 SNOMED CT concepts for "Sodium chloride 0.9% injection solution" as a product, differing only in their package and sizes (Table 2). This level of granularity was considered to be beyond the useful level of distinction for the UMLS concept structure, and all 16 SNOMED CT concepts were merged into the same UMLS concept. This is analogous to the way specific drug products are treated within the UMLS. The National Drug Codes (NDCs) that refer to such products (e.g., the 100-count bottle of a specific brand of aspirin) are included as attributes of the standard (RxNorm) name of the clinical drug they contain.

A third cause of merging involved SNOMED CT concepts containing the "NOS" (Not Otherwise Specified) qualifier.

Table 1 ■ Five Most Common Types of Merges Across SNOMED CT Hierarchies in the Unified Medical Language System

SNOMED CT Hierarchies Involved	Examples of Merged SNOMED CT Concepts	No. of Merges
"Product" and "substance"	"Antacid (product)" and "antacid (substance)"	2957
"Clinical finding" and "body structure"	"Stab wound (disorder)" and "Stab wound (morphologic abnormality)"	978
"Clinical finding" and "context-dependent category"	"Eye symptoms (finding)" and "eye symptom findings (context-dependent category)"	784
"Clinical finding" and "observable entity"	"Antenatal screening finding (finding)" and "antenatal screening finding (observable entity)"	142
"Procedure" and "qualifier value"	"Vascular surgery procedure (procedure)" and "vascular surgery (qualifier value)"	79

One example was "multiple cranial nerve palsies NOS (disorder)," a SNOMED CT concept distinct from the concept "multiple cranial nerve palsies (disorder)." Most of these concepts have a concept status of "limited" in SNOMED CT, meaning that they are of limited clinical value because they are based on a classification concept or an administrative definition. However, they are still valid for current use and considered active.[9] UMLS editing policy, based on the reason that the addition of "NOS" does not convey any additional information about a concept, is that there is no difference in meaning between a concept (e.g., "diabetes mellitus") and its NOS counterpart ("diabetes mellitus, NOS"). As a result, they were merged into the same UMLS concept.

Finally, there were some cases of missed synonymy (two equivalent concepts existing as distinct concepts) in SNOMED CT that were discovered in the editing process of UMLS and therefore merged.[10] One example was "abnormal ECG (finding)" and "ECG abnormal (finding)" existing as two SNOMED CT concepts. Unlike the previous types of merges, these merges represented unintentional differences between SNOMED CT and UMLS. A list of such cases was sent to the College of American Pathologists for review. In the above example, the concept "ECG abnormal (finding)" was demoted to a status of "duplicate" and made inactive in the July 31, 2004 release of SNOMED CT. Other authors have also reported on cases of missed synonymy within SNOMED CT detected by ontology-based methods.[11]

Synonymy in the SNOMED CT View But Not in the Unified Medical Language System View

There were also cases in which the concept names of one SNOMED CT concept were placed in multiple UMLS concepts. During the integration of a vocabulary into the UMLS, human editors evaluate all the concept names of a source concept to see whether they are close enough in meaning to be included in the same UMLS concept. There are cases in which the difference in meaning is substantial enough for a particular concept name to be split out and put into another UMLS concept.[3] For example, in SNOMED CT, "motor vehicle accident" was considered a synonym of "motor vehicle accident (victim)." In the UMLS, the concept name "motor vehicle accident" was split out and put into another existing

Table 2 ■ Sixteen SNOMED CT Concepts for "Sodium Chloride 0.9% Injection Solution"

SNOMED CT ConceptId	Fully specified name
400642004	Sodium chloride 0.9% injection solution 100-mL vial (product)
400274004	Sodium chloride 0.9% injection solution 10-mL vial (product)
400814003	Sodium chloride 0.9% injection solution 2.5-mL vial (product)
400854001	Sodium chloride 0.9% injection solution 20-mL vial (product)
400808005	Sodium chloride 0.9% injection solution 25-mL vial (product)
400474002	Sodium chloride 0.9% injection solution 2-mL vial (product)
400752009	Sodium chloride 0.9% injection solution 30-mL vial (product)
400703005	Sodium chloride 0.9% injection solution 3-mL vial (product)
400307009	Sodium chloride 0.9% injection solution 50-mL vial (product)
400787009	Sodium chloride 0.9% injection solution 5-mL vial (product)
400322004	Sodium chloride 0.9% injection solution 6-mL vial (product)
351453001	Sodium chloride 0.9% injection solution ampule (product)
406330009	Sodium chloride 0.9% injection solution 10-mL ampule (product)
406328007	Sodium chloride 0.9% injection solution 2-mL ampule (product)
406331008	Sodium chloride 0.9% injection solution 20-mL ampule (product)
406329004	Sodium chloride 0.9% injection solution 5-mL ampule (product)

UMLS concept "traffic accidents." Altogether 3040 SNOMED CT synonyms were split from 1972 SNOMED CT concepts (0.7% of total number of SNOMED CT concepts). Some of these cases may be unintentional differences between UMLS and SNOMED CT due to previously undetected nonsynonymous SNOMED CT synonyms.[12]

Representation of Different Views of Synonymy in the Unified Medical Language System

One of the purposes of the UMLS is to act as the bridge between multiple biomedical vocabularies.[13] To do this, the UMLS must be able to present a unifying view (the UMLS concept view) through which contents of different vocabularies can be linked together.[14] However, this does not imply that the UMLS concept view is the only "correct" view of synonymy. An important UMLS goal is that, in the incorporation of a source vocabulary, there should be no information loss. Every bit of important information in the source should be retrievable from the UMLS (source transparency), even though the representation (e.g., file and data structure) is different from that in the source. The new Rich Release Format of the Metathesaurus greatly enhances its ability to achieve source transparency.[15] The representation of information at the atomic level allows source information to be expressed more clearly and accurately. An atom in UMLS is a unit of meaning from a source vocabulary. A UMLS concept is made up of one or (usually) more atoms.

Part of source transparency is the preservation of the source's view of synonymy. The SNOMED CT view of synonymy can be retrieved from the UMLS Rich Release Format. In the UMLS view, synonyms are grouped under a UMLS concept identified by a single Concept Unique Identifier (CUI). In other words, all atoms having the same CUI are synonymous in the UMLS view. For example, the two UMLS atoms representing the SNOMED CT concepts "stab wound (disorder)" and "stab wound (morphologic abnormality)" were merged into the same UMLS concept; thus, they shared the same CUI. However, these two atoms had different Source Concept Unique Identifiers (SCUI, corresponding to ConceptIds in SNOMED CT), revealing the fact that they were different concepts in the SNOMED CT view. The source-asserted relationship between these two SNOMED CT concepts was also preserved and represented by an atom level relationship in the UMLS. On the other hand, when SNOMED CT–asserted synonyms were split in the UMLS (e.g., the SNOMED CT synonyms "hernia" and "hernia of abdominal cavity" were put into two different concepts in the UMLS), the two atoms representing the split synonyms would have different CUIs in the UMLS. However, all SNOMED CT atoms from the same SNOMED CT concept would have the same SCUI, signifying that they were synonymous in the SNOMED CT view.

The Study on Human Determination of Synonymy

True synonymy, relative synonymy, and nonsynonymy can be seen as a continuum. Where synonymy ends and nonsynonymy begins is often fuzzy and context dependent. For example, a patient suffers from residual muscle weakness of his left arm after recovering from a stroke. This finding is recorded as "incomplete paralysis of the left arm" in the patient's record. Can we say, in general, that "muscle weakness" is synonymous with "incomplete paralysis"? It is likely that some health care professionals would agree, while others would disagree. In the determination of synonymy, there is often an element of subjective judgment. In the UMLS, human review plays a major role in the editing process. How accurately are these potentially subjective judgments of synonymy being made by individual editors? To what extent do UMLS editors agree with other domain experts? These are important questions with direct relevance to the quality of the UMLS. To answer these questions, we carried out a study on human determination of synonymy.

Study Design and Methodology

In March 2004, at the completion of the integration of SNOMED CT into the UMLS, pairs of SNOMED CT fully specified names (i.e., concept names with full specification of the hierarchy that the concept belongs to [e.g., "pneumonia (disorder)"]) and their synonyms were selected. These pairs of SNOMED CT–asserted synonymous concept names came from two pools. Half of them came from a pool in which there was concordance between the SNOMED CT and UMLS views of synonymy (i.e., the fully specified name and its synonym both ended up in the same UMLS concept). The remaining half came from a second pool in which there was discordance of the two views (i.e., the fully specified name and its synonym ended up in different UMLS concepts). The pairs were listed in no particular order, and the subjects of the study were not aware of the pool from which a particular pair of names originated. The sample of concept names was chosen in a random fashion with the following exclusions:

1. Straightforward synonymy, e.g., "myocardial infarction (disorder)" and "myocardial infarct"
2. Acronyms, abbreviations, and eponyms, e.g., "benign prostatic hyperplasia (disorder)" and "BPH"
3. Nonclinical concepts, e.g., geographical locations
4. Nonhuman concepts, e.g., hypoglycemia of piglets
5. Specialized concepts outside the scope of general medicine, e.g., neurosurgical procedures (this was to avoid the need for the subjects to do extensive lookup in reference sources)

The subjects of the study were all clinical terminology specialists, with slightly more than half of them serving as UMLS editors. Each subject was asked to rate each pair of terms (concept names) according to how synonymous that he or she thought the terms were, using a scale of 1 (strongly disagree) to 5 (strongly agree), with 3 being neutral. A sample of the terms used is shown in Table 3. The subjects were free to use any reference material that they thought necessary. The UMLS editors were not allowed to see UMLS concept reports normally available to them in the UMLS editing environment, lest they be biased by the UMLS view of synonymy developed in earlier work. In real UMLS editing, editors sometimes need to see the concept reports to understand the meaning of a particular name in a source vocabulary, particularly when the face validity of that name is questionable (e.g., "Prostate" in International Classification of Diseases 9CM may actually mean prostate neoplasms). However, this phenomenon was unlikely to be present because one member of a pair of names was a fully specified name, for which there should be no ambiguity in its meaning.

Statistical analysis was done by the statistical program package SPSS 12.0 for Windows (SPSS Inc., Chicago, IL).

Table 3 ■ A Sample of the Synonymy Study Questionnaire

Please determine if Term1 is synonymous with Term2 and choose from 1 to 5: 1. Strongly disagree 2. Slightly disagree 3. Neutral 4. Slightly agree 5. Strongly agree

Term1	Term2	Your Answer
Tobacco dependence syndrome (disorder)	Tobacco abuse	
Fourth nerve palsy (disorder)	Trochlear nerve disease	
Lung inflation by intermittent compression of reservoir bag (regimen/therapy)	Hand bagging	
Congenital disease (disorder)	Fetal developmental abnormality	
Arthroscopic surgical procedures on knee (procedure)	Arthroscopic knee procedures	
Malignant tumor of anorectal junction (disorder)	Malignant tumor of anorectum	
Atrophic vulva (disorder)	Atrophic vulvitis	
Muscle weakness (finding)	Incomplete paralysis	
Premature beats (disorder)	Ectopics	
Terminal insomnia (disorder)	Early waking	

Results

A total of 11 clinical terminology specialists (six UMLS editors and five noneditors) completed the questionnaire, which contained 60 pairs of concept names.

Range of synonymy scores for individual questions

For each question, each subject could assign a synonymy score from 1 to 5. The range of actual scores being given (i.e., the difference between the highest and lowest scores) reflected the magnitude of division of opinions. In the more straightforward cases, the range would be low, as most answers were concentrated toward one end of the scale. In the most controversial cases, the range would be 4, meaning that at least one subject rated the names as highly synonymous (score of 5) while at least one other subject strongly disagreed (score of 1). The distribution of the range of scores is shown in Figure 1. The average of the ranges of scores for all questions was 3.4. There was an obvious skew toward higher values, with 52 (87%) of 60 questions having ranges of 3 or above. This result confirmed that most of the questions exhibited some degree of controversy, one of the design criteria. The most straightforward cases of synonymy had been excluded in the creation of the questionnaire.

Pattern of Scoring of Individual Subjects

Two distinct patterns of scoring could be observed. The first was a bell-shaped pattern, with the highest number of answers being neutral answers. This pattern was observed in two subjects, both of them UMLS editors. The other pattern was the bimodal pattern with very few neutral answers. This pattern was observed in nine subjects. Typical examples of the two patterns of response are shown in Figure 2. Altogether, neutral answers constituted only 8% of all the answers. This showed that for the majority of the questions, the subjects did have an opinion in favor of either synonymy or nonsynonymy.

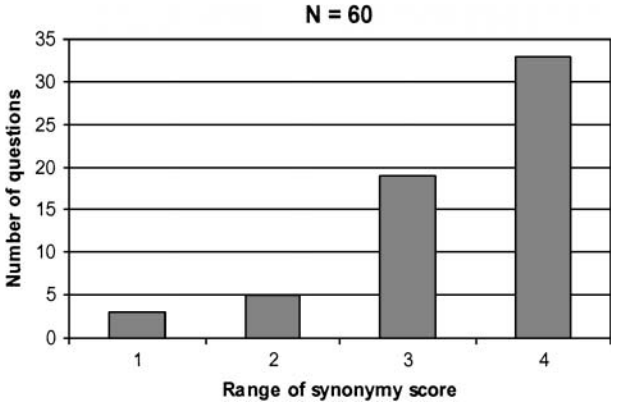

Figure 1. Distribution of the range of synonymy score for each question.

Consistency between Subjects

One of the aims of this study was to examine the accuracy of synonymy judgment by individual UMLS editors. To do this, we needed a gold standard with which to compare. We decided to use the pooled average synonymy score of all subjects as our gold standard for the degree of synonymy of each pair of terms. Before we could pool the scores of different subjects, we had to show that the scores from different subjects were internally consistent, i.e., that they were measuring the same attribute in the objects. The statistic that we used was Cronbach's alpha. Essentially, Cronbach's alpha is the proportion of the total observed variability that is explained by true score differences (the true difference between the objects) rather than interobserver differences. Cronbach's alpha is commonly used in the determination of the internal consistency or reliability of measurement scales in evaluation studies. A high alpha implies high internal consistency. In our study, the overall value of alpha for all subjects was 0.824. In general, an alpha of above 0.7 is considered to be adequate for comparison studies.[16]

We also needed to know the internal consistency among editors and noneditors separately, as we would be using the group average scores to assess the degree of agreement between the two groups. When calculated as separate groups, the alpha for the six editors was 0.735, which was slightly higher than that of the five noneditors (0.717). However, a higher number of subjects typically inflates the magnitude of alpha, and this effect can be corrected for by the Spearman-Brown prophecy formula.[16] According to this formula, the projected value of alpha for noneditors would become 0.752 had there been six noneditors instead of five. In any case, whether we were looking at the individual groups or all subjects as a whole, the degree of internal consistency was reasonably high.

Accuracy of Individual Unified Medical Language System Editors' judgment

An overall mean synonymy score was calculated for each pair of terms using scores from all 11 subjects. These average scores were collapsed into three synonymy categories: nonsynonymous, neutral, and synonymous (for average scores less

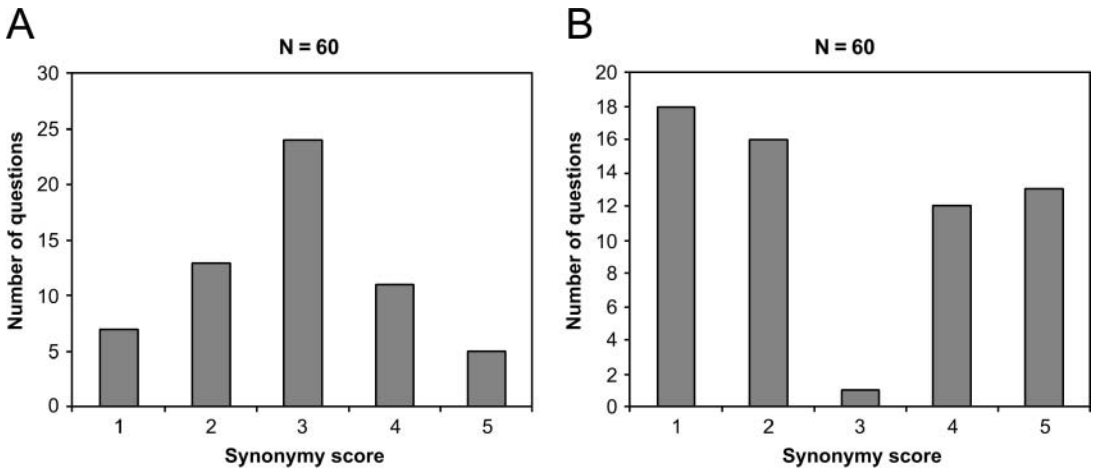

Figure 2. Examples of the two distinct patterns of response. **A:** Bell-shaped (observed in two subjects). **B:** Bimodal (observed in nine subjects). $N = 60$.

than, equal to, and greater than 3). This was the gold standard to which each subject's judgment was compared. The use of a uniform gold standard simplified the analysis, but the fact that each subject's scores also formed part of the gold standard might yield slightly higher concurrence rates compared to the alternative method of comparing each subject's scores with the average scores of all other subjects except him- or herself. Each UMLS editor's scores were similarly collapsed into the same three categories for each pair of terms. A cross-tabulation of the overall mean synonymy category assignment against the individual editor's synonymy category assignment is shown in Table 4. The accuracy of each editor was assessed by the degree of concurrence between the two synonymy category assignments, which was calculated as follows. Take Editor-1 as an example; the two synonymy category assignments agreed in 35 cases (both nonsynonymous: 18, both neutral: 1, both synonymous: 16). In eight cases (middle cells of the four edges of the 3×3 table: 6, 1, 1, 0), either one of the two synonymy category assignments was neutral. We considered half of these cases as concurrences (theoretically there would have been a 50% chance of the two assignments agreeing if a neutral category was not allowed). Therefore, the total degree of concurrence was 39 of 60 cases, an accuracy of 65%. The accuracy of editors ranged from 65% to 77.5%, and the average accuracy was 71%. Similar analysis was done for noneditors, and the result is shown in Table 5. The accuracy of nonediting domain experts ranged from 68% to 82.5%, with an average of 75%. The difference between the accuracy of editors and noneditors was not statistically significant (independent samples t-test, t = 1.267, two-tailed p = 0.237).

Agreement between Editors and Noneditors

The second aim of this study was to investigate the extent to which the judgments of UMLS editors agreed with other domain experts. A high degree of agreement would give support to the validity of the UMLS editing process. The extent of editors and noneditors agreement was estimated by three different methods.

First, the average synonymy score for each question was calculated for editors and noneditors separately. In the scatterplot of the average scores of the editors against noneditors (Fig. 3), there was a definite pattern of positive correlation, meaning that a higher editor score generally corresponded with a higher noneditor score. Statistically, this impression was confirmed. The Spearman's rank correlation coefficient was 0.654, which was significant at the 0.01 level (two tailed).

Second, the average scores were collapsed into three synonymy categories as above. The results are shown in Table 6. Among the 60 cases, editors and noneditors concurred in their synonymy categories in 40 cases (both nonsynonymous: 30, both synonymous: 10). In 10 cases, either one of the synonymy category assignments was neutral. As before, we considered half of these cases as concurrences, giving an overall concurrence rate of 45 (75%) of 60 cases.

Finally, the kappa statistic can be used to quantify interrater agreement. Kappa is a measurement of the degree of observed agreement above that which can be explained by chance alone.[17] To simplify calculation, Table 6 was reduced to a 2×2 table by eliminating the neutral score cells and splitting their contents equally into adjacent cells (i.e., the first row

Table 4 ■ Accuracy of Individual Editors, Using the Overall Mean Synonymy Score as the Gold Standard

	Editor-1			Editor-2			Editor-3			Editor-4			Editor-5			Editor-6		
Overall Mean Score	NS	N	S	NS	N	S	NS	N	S	NS	N	S	NS	N	S	NS	N	S
NS	18	6	15	30	1	8	34	0	5	24	2	13	19	14	6	30	7	2
N	0	1	1	0	0	2	2	0	0	0	0	2	0	0	2	1	0	1
S	2	1	16	4	0	15	9	0	10	4	2	13	1	10	8	6	5	8
Accuracy	65%			77.5%			75%			67%			67%			75%		

NS = nonsynonymous; N = neutral; S = synonymous.

Table 5 ■ Accuracy of Individual Noneditors, Using the Overall Mean Synonymy Score as the Gold Standard

	Noneditor-1			Noneditor-2			Noneditor-3			Noneditor-4			Noneditor-5		
Overall Mean Score	NS	N	S	NS	N	S	NS	N	S	NS	N	S	NS	N	S
NS	32	1	6	30	0	9	35	1	3	22	2	15	26	0	13
N	2	0	0	0	0	2	2	0	0	1	0	1	1	0	1
S	3	0	16	5	0	14	8	0	11	2	0	17	3	0	16
Accuracy		82.5%			75%			79%			68%			72%	

NS = nonsynonymous; N = neutral; S = synonymous.

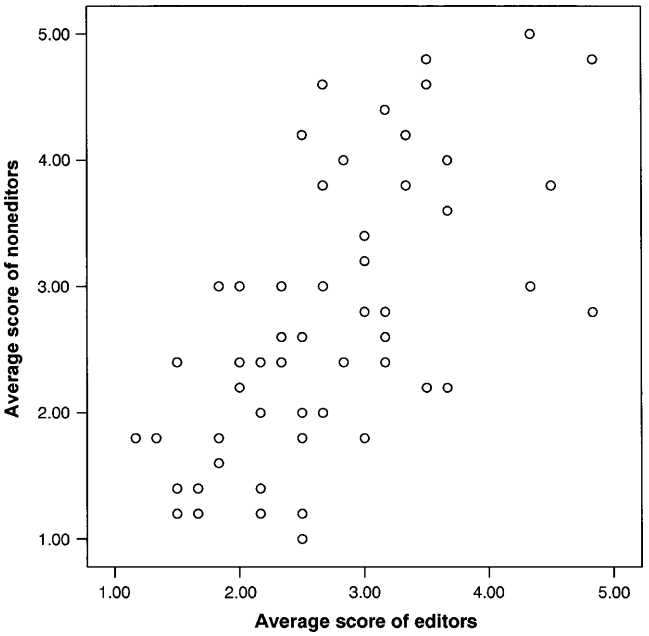

Figure 3. Scatterplot of average scores of editors and noneditors.

Table 6 ■ Agreement between the Synonymy Category Assigned by Editors and Noneditors

Noneditors' Synonymy Category	Editors' Synonymy Category		
	Nonsynonymous	Neutral	Synonymous
Nonsynonymous	30	2	6
Neutral	4	0	2
Synonymous	4	2	10

Table 7 ■ Effect of Matching Name in Another UMLS Concept on the Odds of Splitting in Nonsynonymous Cases (N = 38)

	Has Exact Matching Name in Another UMLS Concept	
Result of Editing	Yes	No
Split	17	2
Not split	3	16
Odds of being split (no. split/no. not split)	5.67	0.125

UMLS = Unified Medical Language System.

became (33, 8), the third row became (7, 12), and the middle row and middle column were eliminated). Based on the simplified table, kappa was calculated to be 0.43 (standard error = 0.129, z = 3.34, one-tailed p value <0.0004), showing a significantly better interrater agreement than that due to chance alone. In general, kappa values above 0.40 can be taken to represent fair to good agreement beyond chance.[17]

Discussion

Human Editing in Biomedical Vocabularies

Cruse[2] pictured synonymy as a series of concentric circles, the center being the point of semantic identity. The further away from the center, the larger is the semantic difference. In a concept-based vocabulary, the bigger the circles of synonymy, the larger will be the size (semantic span) of its concepts. Why then do we not make the circles so small that all concepts will be as narrowly and clearly defined as possible? The reason is, as Blois[18,19] suggested, that in the biomedical domain (unlike in disciplines such as physics or chemistry), we often need concepts with sufficient semantic span to capture the meanings that we need in normal professional discourse. How big should these circles be? There is no single "correct" concept size for all vocabularies. The requirements of a vocabulary for clinical data capture will be quite different from another created primarily for statistical reporting or indexing of biomedical literature.

Despite significant progress in knowledge-based automation in their development and maintenance,[20–22] manual curation is still an essential part in the editing of most, if not all, controlled biomedical vocabularies. Human expert judgment is often regarded as the gold standard in many studies involving terminologies. However, the quality of human judgment in vocabulary editing has been called into question.[23] Given the somewhat fuzzy boundary between synonymy and nonsynonymy, can we expect editors to make accurate decisions regarding synonymy between concepts? As far as we are aware, there has been no published research addressing this issue. Our study shows that UMLS editors are accurate in their decisions in about 71% of the cases. This figure alone is not particularly impressive because random guessing alone would give an accuracy of 50%. However, in real-life editing, the accuracy is expected to be higher because simple and straightforward cases of synonymy were deliberately excluded in this study. The accuracy of UMLS editors is not significantly different from nonediting domain experts (75%). Moreover, the fairly good degree of agreement between editors and noneditors is also encouraging and suggests that the criteria used by the UMLS editors for determining synonymy are generally in agreement with that of other

domain experts. These findings justify the choice made to allow human editing as the fine-tuning finishing touch after the algorithmic processing in the production of the UMLS.

Unique Features of the Unified Medical Language System Editing Environment

There are significant differences between the nature of the editing processes for the UMLS and for individual vocabularies. The UMLS is not an independent vocabulary. It links multiple vocabularies. The main task in UMLS editing is to determine accurately the meaning of a source vocabulary and to integrate its contents with other vocabularies already in the UMLS according to synonymy.

Each new vocabulary added to the UMLS undergoes an initial analysis to determine whether it is itself concept based, i.e., whether terms labeled as synonyms within that source are in fact generally synonymous and do not include, for example, many entry terms obviously narrower in meaning. If a new source is not concept based, it receives special preprocessing before lexical matching and normal UMLS editing. After any initial preprocessing step, as a general UMLS editing principle, synonymy genuinely asserted by a source vocabulary is respected unless it conflicts with that asserted by other source vocabularies. For example, if a certain source vocabulary X asserts that concept names a and b are synonyms, and if X is the only source that contains a and b, they will usually be put in the same UMLS concept. On the other hand, if there is another source vocabulary Y that says that a and b are not synonymous, then the UMLS editor has to decide between the two conflicting views of synonymy. If the editor thinks that the degree of synonymy between a and b is indeed low, they will be split into separate UMLS concepts. In other words, a questionable degree of synonymy is necessary but not always sufficient for the splitting of source-asserted synonyms in UMLS editing.

With this understanding, we could explain an apparent discrepancy when we correlated the actual outcome of UMLS editing with the assessed degree of synonymy of the concept names used in our study. Intuitively, one would expect a high proportion of those names that were considered nonsynonymous to end up being split into different UMLS concepts. However, among the 38 pairs of names that had low synonymy scores (average editors' synonymy score less than 3), only half of them ended up being split. What happens in UMLS editing is that after algorithmic processing of a source vocabulary, all cases in which matching names exist in different UMLS concepts (signifying potential conflicting views of synonymy) are flagged for attention. Splitting is most likely to occur in those flagged concepts. As can be seen from Table 7, among the 38 nonsynonymous cases, those with matching names in another UMLS concept had much higher odds of being split (5.67) compared to those without (0.125), giving a very high odds ratio of 45.3. As a corollary of this, it can be said that the UMLS view of synonymy is not a totally independent assertion of synonymy. It is a representation of the sum of the views of all its *concept-based* source vocabularies after conflicts are resolved in the editing process.

Conclusion

Integration of SNOMED CT into the UMLS involved the alignment of two views of synonymy in two concept-based vocabulary systems. In the majority of cases, the two views agreed with each other but were different for about 14% of SNOMED CT concepts. The differences were largely reflections of the different organizing principles and purposes of the two vocabulary systems. Both views of synonymy coexisted in the UMLS and were explicitly represented.

Determination of synonymy between concepts is a key process in the creation and maintenance of concept-based terminologies. Despite efforts in automation, human editing is still the main method in determining synonymy. However, the quality of human decisions is often assumed but seldom proven. Our study shows both UMLS editors and domain experts not involved in UMLS editing achieved fair accuracy in their judgment of synonymy in potentially controversial SNOMED CT synonyms. The overall agreement between the two groups was satisfactory.

References ■

1. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37:394–403.
2. Cruse DA. Lexical semantics. Cambridge, UK: Cambridge University Press; 1986.
3. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med. 1995;34:193–201.
4. SNOMED International Web site. Available from: www.snomed.org. Accessed February 7, 2005.
5. Physician Data Query Online System Web site. Available from: www.nci.nih.gov/cancertopics/pdq. Accessed February 7, 2005.
6. Automated medical records: leadership needed to expedite standards development. USGAO-IMTEC-93-17. Washington, DC: U.S. General Accounting Office; 1993.
7. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Inform Assoc. 1994; 1:218–32.
8. Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp. 1997;640–4.
9. SNOMED Clinical Terms Technical Reference Guide. Northfield, IL: College of American Pathologists, 2003.
10. Hole WT, Srinivasan S. Discovering missed synonymy in a large concept-oriented metathesaurus. Proc AMIA Annu Symp. 2000; 354–8.
11. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT(R). Medinfo. 2004;2004:482–6.
12. Nash SK. Nonsynonymous synonyms: correcting and improving SNOMED CT. AMIA Annu Symp Proc. 2003;949.
13. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993;32:281–91.
14. Campbell KE, Oliver DE, Spackman KA, Shortliffe EH. Representing thoughts, words, and things in the UMLS. J Am Med Inform Assoc. 1998;5:421–31.
15. Hole WT, Carlsen B, Tuttle MS, Srinivasan S, Lipow SS, Olson N, et al. Achieving "Source Transparency" in the UMLS metathesaurus. Medinfo. 2004;2004:371–5.
16. Friedman C, Wyatt J. Evaluation methods in medical informatics. New York: Springer; 1997;89–117.
17. Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley & Sons; 2003;598–626.
18. Blois MS. The effect of hierarchy on the encoding of meaning. Presented at the Proceedings of the Tenth Annual Symposium on Computer Applications in Medical Care (SCAMC), 1986.
19. Blois MS. Medicine and the nature of vertical reasoning. N Engl J Med. 1988;318:847–51.

20. Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the Medical Entities Dictionary. J Am Med Inform Assoc. 2000;7:288–97.
21. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. 1994;1:35–50.
22. Rodriguez J, Maojo V, Crespo J, Fernandez I. A concept model for the automatic maintenance of controlled medical vocabularies. Medinfo. 1998;9:618–22.
23. Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. Proc AMIA Annu Symp. 1998;795–9.