Research Paper ■

Portability Issues for a Structured Clinical Vocabulary: Mapping from Yale to the Columbia Medical Entities Dictionary

JOSEPH L. KANNRY, MD, LAWRENCE WRIGHT, MA, MARK SHIFMAN, MD, PHD, SCOT SILVERSTEIN, MD, PERRY L. MILLER, MD, PHD

Abstract **Objective:** To examine the issues involved in mapping an existing structured controlled vocabulary, the Medical Entities Dictionary (MED) developed at Columbia University, to an institutional vocabulary, the laboratory and pharmacy vocabularies of the Yale New Haven Medical Center.

Design: 200 Yale pharmacy terms and 200 Yale laboratory terms were randomly selected from database files containing all of the Yale laboratory and pharmacy terms. These 400 terms were then mapped to the MED in three phases: mapping terms, mapping relationships between terms, and mapping attributes that modify terms.

Results: 73% of the Yale pharmacy terms mapped to MED terms. 49% of the Yale laboratory terms mapped to MED terms. After certain obsolete and otherwise inappropriate laboratory terms were eliminated, the latter rate improved to 59%. 23% of the unmatched Yale laboratory terms failed to match because of differences in granularity with MED terms. The Yale and MED pharmacy terms share 12 of 30 distinct attributes. The Yale and MED laboratory terms share 14 of 23 distinct attributes.

Conclusion: The mapping of an institutional vocabulary to a structured controlled vocabulary requires that the mapping be performed at the level of terms, relationships, and attributes. The mapping process revealed the importance of standardization of local vocabulary subsets, standardization of attribute representation, and term granularity.

■ **JAMIA.** 1996;3:66–78.

Controlled vocabularies are commonly used in biomedical and other scientific domains. In its simplest form, a controlled vocabulary is a selected set of specialized terms that facilitates precise communication by eliminating ambiguity. A set of terms, however, is insufficient to represent biomedical phenomena in a robust fashion. One also must be able to *relate* the terms to one another (with a set of relationships) and to *qualify* the terms (with a set of attributes). In this paper, we use the term "structured" vocabulary to mean a controlled vocabulary that has been augmented with such relationships and attributes.

In a structured vocabulary, relationships between terms can be hierarchical or non-hierarchical. In a hierarchical relationship, a specific term (for example, *serum sodium test*) is a "child" of a broader parent term (such as *intravascular chemistry test*) and inherits

Affiliations of the authors: Yale University School of Medicine, New Haven, CT.

Supported in part by NIH Contract N01 LM13537 and NIH Grants T15 LM 07056 and G08 LM05563 from the National Library of Medicine and by an equipment grant from the Sun Microsystems Corporation.

Correspondence and reprints: Joseph L. Kannry, MD, Mount Sinai School of Medicine, P.O. Box 1087, 1 Gustave L. Levy Plaza, New York, NY 10029. e-mail: joseph_kannry@smtplink.mssm.edu

Received for publication: 6/19/95; accepted for publication: 9/13/95.

attributes (characteristics) from the parent term. For example, the attribute *normal value range* would be inherited by the term *sodium test* if its parent, *intravascular chemistry test*, had this attribute. Non-hierarchical relationships are commonly called semantic relationships. Semantic relationships can be a form of knowledge representation[1,2] focusing on the real-world interaction between the terms. As an example of a semantic relationship, the semantic relation **measured by** defines the relationship between the terms *sodium* and *sodium test*.

A controlled vocabulary enables precise access to important elements of patient data such as medications, procedures, test, costs, and hospital resources required for care. Applications accessing the data can perform queries using unambiguous terms. For example, a query using a controlled vocabulary term such as *glucose* will produce desired results because the meaning of the term has been standardized. In a structured vocabulary, a query may also request the names of all tests that **measure** *glucose*. Since the relation *glucose test* **measures** *glucose* is stored in the structured vocabulary, the query will retrieve correct responses with varying names such as *intravascular glucose test* and *serum glucose test*. An undesired term such as *glucose-6-dehydrogenase deficiency test* (which does not **measure** glucose) will not be retrieved, even though it lexically contains "glucose."

To allow a structured vocabulary to be used in conjunction with a given institution's data, a "mapping" must be established between the elements of that vocabulary (terms, relationships, and attributes) and the vocabulary used in the target institution's database. *Mapping* is the process of identifying corresponding elements in the vocabularies being examined. One approach for mapping an institutional vocabulary to a structured vocabulary involves the following steps. First, terms in the two vocabularies are matched lexically. Second, any relationships in the institutional vocabulary are matched to relationships in the structured vocabulary. Third, the attributes of lexically similar terms are matched.

In this paper, we examine the issues involved in mapping an institutional vocabulary [the laboratory and pharmacy vocabularies currently used at the Yale New Haven Medical Center (YNHMC)] to an existing structured vocabulary [the Medical Entities Dictionary (MED) developed at Columbia University]. In the process we have attempted to identify issues involved in performing such a mapping. We also identify features of a structured vocabulary that could facilitate such a vocabulary mapping process.

Background

Mapping between Clinical Vocabularies

In 1986, the National Library of Medicine initiated the Unified Medical Language System (UMLS) Project.[3] One of the goals of the project was to develop tools to facilitate the process of mapping between different controlled vocabularies.[4,5] Various approaches to mapping two different vocabularies were considered.

Sherertz et al.[5] proposed that a straightforward lexical mapping approach might be attempted as an initial mapping step. They defined lexical mapping as matching on an "exact word by word equivalence of phrase." They demonstrated the approach by lexically mapping 834 disease descriptions [created at the University of Southern California at San Francisco (UCSF)] to Medical Subject Headings (MeSH)[6] terms. They successfully mapped 47.8% of the UCSF disease descriptions to MeSH terms. Then, specific attributes of the UCSF terms such as etiology, treatment, laboratories, and signs were mapped to MeSH, with 48.7% of the attributes successfully mapped to MeSH.

Another approach to mapping was proposed by Evans et al.,[7] whose goal was to identify problems inherent in mapping terms between two vocabularies. The vocabularies they chose were MeSH and the Systematized Nomenclature of Medicine (SNOMED II).[8] Their proposed solution to mapping problems was the construction of "frames." A frame is a self-contained "unit of knowledge representation"[9] that contains a term and its attributes. The attributes describe semantic and hierarchical relationships between terms.

Masarie et al.[10] approached the issue of mapping multiple vocabularies by attempting to identify a common set of concepts represented by the terms from four vocabularies and then mapping terms in each vocabulary to these concepts. Thus, these concepts served as a type of "interlingua." For example, the central concept underlying the three symptom terms *heartburn*, *pleuritic pain*, and *angina* (which could come from three different vocabularies) is chest pain. Therefore, the three terms could be mapped to *chest pain*. The use of concepts as an interlingua reduces the number of term comparisons. Like Evans et al., Masarie et al. used a frame-based approach to mapping. They constructed frames and placed one term and its attributes inside each frame. These attributes described term relationships and hierarchies. Using this approach, they were able to map terms from four large controlled vocabularies: Quick Medical Reference (QMR), HELP's PTXT, DXplain, and MeSH.

To make the process manageable, they limited mapping to terms representing signs and symptoms. A significant challenge was the creation and maintenance of the frames.

Cimino and colleagues[4,11] also used a frame-based mapping approach. Like Masarie et al. and Evans et al., they constructed frames containing terms and their attributes. They then placed links between the frames. One type of link was a semantic relationship. The nature of the semantic relationship was specified by a term's attribute. For example, the term *sodium* has the attribute **measured by** and the "attribute value" *sodium test*. Since *sodium test* is also a term, the semantic relationship **measured by** now links the two terms *sodium* and *sodium test*.

By allowing "attribute values" to be terms, a complex web or network of semantic relationships between terms develops. The semantic network facilitated automated mapping by permitting a mapping algorithm to utilize the attributes to look for similarities between two vocabularies. The overall success rate of automated mapping of ICD9-CM[12] terms to MeSH terms was 25 of 56, or 44.6%. The semantic network had the potential to facilitate and automate mapping and the maintenance of mapped terms.[4] The semantic network became one of the cornerstones for the MED, since it facilitates adding terms and maintaining term lists.[2]

Overview of the MED and the Yale Vocabulary

The MED

The MED from Columbia University is a data dictionary that contains several linguistic elements, including 1) terms derived from multiple controlled vocabularies, 2) hierarchical and non-hierarchical term relationships, and 3) term attributes.

Terms. One of the goals set by the designers of the MED was "domain completeness."[2,11] Since no controlled vocabulary adequately covers the entire domain of medicine, an attempt to move in the direction of domain completeness was made by populating the MED with terms derived from multiple vocabulary sources, including the UMLS Metathesaurus, ICD9-CM, and the institutional vocabulary from four Columbia Presbyterian Medical Center (CPMC) systems (laboratory, electrocardiography, medical record coding, and pharmacy).[2]

Relationships. Structure in the MED is represented by 54 hierarchical and non-hierarchical relationships. The broadest terms (higher-level concepts) are at the top of a hierarchical "relationship tree" and include such terms as *chemical, diagnostic test,* and *cardiovascular drugs*. More focused descendant terms (lower-level concepts) such as *serum sodium tests, beta-HCG tests,* and *digoxin preparations* are found along branches of the tree. The "leaves" of the tree are the most focused, and in this paper are called "instances." Instances are specific variations or implementations of a term. For example, *presbyterian serum sodium test, cpmc plasma sodium test,* and *allen pavilion serum sodium test* are instances of the term *serum sodium test*.

As described above, the structure of *hierarchical* relationships resembles a tree and its branches. *Non-hierarchical* relationships are links between terms in different branches. Non-hierarchical relationships in the MED are referred to as semantic relationships. Cimino et al. required explicit specification of such relationships in the MED.[2] This was accomplished by the use of term attributes (called "slots").

Attributes. There are two types of attributes in the MED, non-literal attributes and literal attributes. Non-literal attributes specify the type of relationship between terms (such as **measures**), and point to the second term, such as *sodium*. Literal attributes specify characteristics that are not relationships. The item pointed to by the literal attribute is not another term but is instead a value. For example, *serum glucose test* has the literal attributes **normal value** and **units**, and the values of the literal attribute might be "70–120" and "mg/dL."

Attributes are inherited through hierarchical parent–child relationships in the MED. Figure 1 shows a directed acyclic graph that depicts the "family tree" of hierarchical inheritance for a typical instance, *digoxin .125 mg po*. Boxes represent terms and the lines between boxes indicate hierarchical parent–child relationships. Each higher-level term in Figure 1 is a parent that can give rise to many children, which are not shown because these children are not hierarchically related to *digoxin .125 mg po*.

The Yale Laboratory and Pharmacy Vocabularies

In contrast to the MED, the Yale laboratory and pharmacy vocabularies are institutional vocabularies with relatively little superimposed structure.

Terms. The Yale laboratory and pharmacy vocabularies consist of 1,854 and 3,771 terms, respectively, and are derived from the respective clinical systems of YNHMC. The vocabulary in the YNHMC pharmacy system, however, as well that of the MED, is derived from a standard vocabulary of the pharmaceutical industry. The set of terms that make up this vocabulary is listed in the AHFS (American Hospital Formulary System) Drug Information book, which

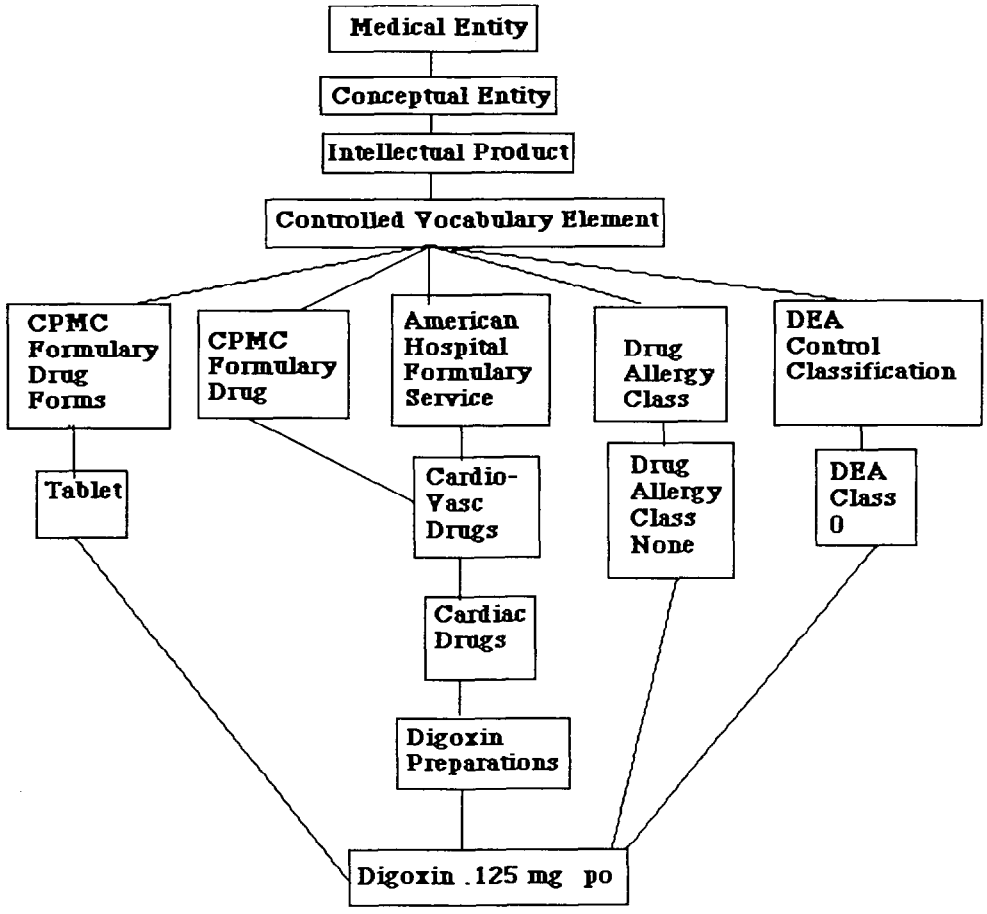

Figure 1 Family tree of hierarchical inheritance for a typical Medical Entities Dictionary (MED) pharmacy term *digoxin .125 mg po*. The boxes represent terms and the lines between the boxes are hierarchical parent–child relationships. Only terms with or ancestors of terms with hierarchical relationships to *digoxin .125 mg po* are shown. Since attributes are inherited hierarchically, attributes are inherited by *digoxin .125 mg po* from all terms shown in this figure. CPMC = Columbia Presbyterian Medical Center; DEA = Drug Enforcement Agency.

contains information about term names and attributes such as dose, dosage, and route. Unique to the AHFS book is the AHFS Classification System, which organizes drugs into therapeutic categories (such as nervous system agents, cardiovascular agents). The AHFS classification is used as a term attribute in both the MED and the Yale vocabularies.

Relationships. Only limited hierarchical relationships are present in Yale's vocabularies. These hierarchical relationships exist between groupings of terms in the laboratory vocabulary. For example, *serum sodium, serum potassium, serum chloride,* and *serum bicarbonate* are four separate laboratory terms that, as a group, are related hierarchically to (i.e., are **part of**) the term *electrolytes panel.* The electrolytes panel allows four tests to be ordered and reported together. As another example, *stool potassium, weight,* and *total in sample* are three separate laboratory terms that are **part of** *stool potassium panel.* Unlike *electrolytes panel,* however, none of the three terms for *stool potassium panel* is used separately since there is no separate test named "weight" or "total in sample." The stool potassium grouping permits the laboratory system to transmit three related results to the hospital system as one entity, in the form of a result panel. Nonhierarchical (semantic) relationships do not exist for Yale terms.

Attributes. The Yale laboratory and pharmacy vocabularies both have literal attributes. For example, *serum sodium* has the attributes *low value, upper value,* and *units. Digoxin* has the attributes *full drug name* and *dose.* The Yale vocabularies do not have nonliteral attributes. Although, as discussed above, the Yale laboratory vocabulary does have a few hierarchical relationships, it does not explicitly represent them by use of non-literal attributes or other means.

This is in contrast to the MED, which explicitly represents hierarchical relationships by using non-literal attributes.

Methods

200 Yale pharmacy terms and 200 Yale laboratory terms were randomly selected from database files containing all of the Yale laboratory and pharmacy terms. These 400 terms were then mapped to the MED, as described below. To facilitate the mapping process, the MED as well as the Yale terms were stored as a series of relational tables in a Sybase database running on a Sun SPARCstation.

Mapping the Yale vocabulary to the MED vocabulary was conducted in three phases. These phases were:

1. mapping terms,
2. mapping relationships between terms, and
3. mapping attributes that modify terms.

Mapping Terms

Lexically equivalent term names were identified by using SQL (Structured Query Language) to query the Yale laboratory and pharmacy vocabulary tables and MED vocabulary tables. Each record in the Yale and MED vocabulary tables contained a term and term attributes. Lexical matches between terms were identified by using case-insensitive substring searches supported by SQL. The results of these matches were then reviewed manually and annotated using a text editor. Anticipating that pharmacy matches might be straightforward, dosing information and route information (in addition to term name) were also retrieved for each lexical match to see whether drug terms could be matched not only by name but by route, dosage, and dose. If drug terms could be matched in such great detail, they could be ported to the MED with minimal modification.

Preliminary attempts to identify lexically equivalent laboratory terms were complicated by greater term name variability. To reduce laboratory term name variability, we performed modifications on Yale laboratory term names. Prior to modification, each term had one entry listed for term name in the Yale laboratory term table. After modification, each term had several entries in the table. Three types of modification were done to Yale terms: string reduction, synonym addition, and phrase breakdown. String reduction (to a lesser degree than the other two), synonym addition, and phrase breakdown have been successfully employed in previous lexical matching efforts.[5,13]

String reduction is the process of reducing each word (a "white space delimited" alphanumeric string[5]) to the least number of unique characters required to perform meaningful lexical matches. The term *lymphocytes* may be reduced to *lympho*, or the term *timed* to *time*, for example. The advantage is increased recall. The disadvantage is reduced precision. Other than removing suffixes, our modification was subjective and not based on an algorithm.

Synonym addition was performed to compensate for *highly variable* or *uncommon term* usage. The assessment of term usage was subjectively determined. A substantial number of term names were considered to be neither uncommon nor highly variable. A Yale laboratory name was considered to be highly variable if it was believed to be frequently expressed in several different forms at other institutions. For example, the Yale laboratory term "ALT" can be correctly expressed at other institutions as "alanine transferase," "SGPT," or "serum glutamate pyruvate transferase." Uncommon usage was defined as a Yale laboratory name that was believed to be rarely used at other institutions. For example, the Yale laboratory name "anti-globulin-direct test" is a rarely used name for "direct Coombs test." After synonym addition was performed, each single term then had several term names associated with it. This process facilitated the automated lexical matching by accounting for differences in nomenclature.

In phrase breakdown, complex terms are broken down into component words. Each individual word becomes a new term. For example, the term *indirect coombs* became two terms, *indirect* and *coombs*. This process facilitates automated lexical matching because it overcomes such problems as reversed order and word separation, and also provides a clearer basis for partial matching.

After synonym addition and phrase breakdown, each original term is associated with many related words, resulting in a larger number of possible lexical matches, which must then be reviewed manually. A "weighted" matching system was used to help deal with this problem. Each word was classified according to its relation to the original term. One of four descriptors was used: name/synonym, modifier, abbreviation, and site. Each type of term relation was assigned a number of points, based on a subjective estimate of its importance. Name/synonym matches were given 10 points. Modifier matches (nonspecific words in a term name such as "indirect" or "cell" or "timed") received 5 points. Site matches, which were matches for the site from which a sample was taken (e.g., "blood" or "tissue"), were given 3 points. Abbrevi-

ation matches received 2 points. If a MED term name lexically matched more than one word associated with a Yale term, the points for each were added together. The matches were then listed in descending order of total point value, with the assumption that the most useful matches would have the highest combined point total. The matches were then manually reviewed for correctness. If a Yale term could not be matched to a MED term after the manual review, an extensive manual search of MED terms was conducted in an attempt to find a match. The final compilation of matched and unmatched terms was reviewed by a domain expert for suggestions on further manual searching strategies.

Mapping Relationships

Since the Yale vocabulary had only a modest amount of relationship structure, it was not possible to map one set of complex relationships to another in a robust fashion. However, each MED term has a meaning defined by its relationships. The MED relationships were therefore used to assist in the lexical mapping process. In this way, the MED relationships were mapped conceptually to the meaning of the Yale terms.

Comparing term meanings is important because an apparent lexical match does not imply semantic equivalence between the terms. Ascertaining semantic equivalence is relatively easy and requires only a cursory look at term relationships if there is only one lexical match, the term names are identical, and the term names are well defined. Determining semantic equivalence is more problematic with a one-to-many lexical match, which occurs when a term from one vocabulary lexically matches to multiple terms from another vocabulary. For example, the term *serum sodium* can have multiple matches to terms in another vocabulary such as *chem-7 sodium, whole blood sodium,* and *plasma sodium.* Additionally, determining semantic equivalence can be problematic when lexically matched term names are similar but have different modifiers such as *serum sodium* and *stat serum sodium.* To identify semantic equivalence, we examined the term's meaning as characterized by a term's hierarchical and non-hierarchical relationships to other terms.

Since one Yale term frequently matched many MED terms, determining semantic equivalence often meant examining MED term relationships in detail. Relationships in the MED are specified by non-literal attributes and thus can be obtained from queries of the MED term tables. For hierarchical relationships, a term's ancestors were identified by tracing all parent–child relationships until the top of the hierarchy was reached. For non-hierarchical relationships, we determined relationship type (i.e., **measured by**, **sampled by**) and the target term of the relationship. After examination of these relationships, if the lexically paired terms were semantically equivalent, then the mapping was considered to be appropriate.

Mapping Attributes

Though terms may be lexically and semantically equivalent, terms may not share attributes and thus may have different characteristics (e.g., a laboratory test may have different normal values). To complete our mapping effort, representative semantically equivalent term pairs (such as *blood sodium* and *allen plasma sodium ion measurement, blood potassium* and *allen plasma sodium ion measurement, furosemide (lasix) 40 mg tab* and *furosemide 40 mg tab,* and *digoxin .125 mg po* and *digoxin .125 mg po*) were selected for attribute comparison. Attribute lists were compiled for each term in the term pair. Each attribute list was divided into literal and non-literal attributes.

To ensure that the attribute lists derived from the selected MED terms were representative, term ancestry was examined. Since attributes in the MED are inherited hierarchically, terms that share ancestors will inherit the same attributes. For example, if all laboratory terms share the same ancestor terms, then all laboratory terms will have the same attributes. It would no longer be necessary to examine each laboratory term individually to derive a complete list of laboratory term attributes. Through a series of queries to the MED tables, we identified shared ancestor terms for all laboratory as well as pharmacy terms.

Results

Pharmacy Terms

Results of Lexical Mapping

Of the original 200 Yale pharmacy terms, 168 terms (84%) lexically matched MED pharmacy terms. We placed the lexical matches of these 168 terms into five categories: identical, closely similar, moderately similar, minimally similar, and unspecifically similar. The percentage of Yale pharmacy terms that lexically matched MED terms in each of the four categories is shown in Table 1. A match was considered "identical" when the Yale pharmacy term had the same *name, dose, dosage,* and *route* as did the MED pharmacy term. For example, Yale's *captopril (capoten) 100 mg tab* and the MED's *cpmc drug:capoten tab 100 mg*

Table 1 ■

Pharmacy Lexical Match Breakdown*

Match Categorization	Percentage of Lexical Matches
Identical	41%
Closely similar	15%
Moderately similar	9%
Minimally similar	14%
Unspecifically similar	22%

*One hundred sixty-eight of the original 200 Yale pharmacy terms lexically matched Medical Entities Dictionary (MED) terms. These lexical matches can be placed into the five categories of similarity shown.

would be identical. Sixty-eight of the 168 (41%) lexical matches were in the identical category. Matches were considered "closely similar" when the Yale pharmacy term had the same name but differed in the value of one attribute (such as *dose*, *dosage*, or *route*); 25 of the 168 (15%) lexical matches were in the closely similar category. For example, Yale's *triamcinolone tab 2 mg* (tab implies po) and the MED's *cpmc drug: triamcinolone 4 mg tab* would be closely similar because they differ only in dose. When the Yale pharmacy term had the same lexical name as did a MED pharmacy term but they had two different attributes, the match was considered moderately similar; 15 of the 168 (9%) Yale matches were in the moderately similar category. For example, Yale's *albuterol (proventil) oral soln .5 mg/1.25 ml* and the MED's *cpmc drug: albuterol inh sol 0.5% 20 ml* would be an example of moderately similar because the *doses* and *routes* of the two terms are dissimilar. When the Yale pharmacy term had only the same name as did a MED pharmacy term, but the two had no other attribute in common, the match was considered minimally similar. These matches generally involved a match between a narrowly focused Yale term and a broader MED term. For example, Yale's *urea (carmol 20) cream 20% 90 gm*, a medication, lexically matched to the broader MED preparation class *urea preparation*. Matches involving these terms accounted for 23 of the 168 (14%).

Finally, some Yale pharmacy terms do not specify values for the attributes *dose* and *dosage* in advance (this allows dose and dosage to be filled in by physicians and nurses upon ordering), and matches involving these terms were considered unspecifically similar. An example of this would be Yale's *aminophylline inj*, which has no specified dose or dosage. Matches involving these terms accounted for 37 of the 168 (22%) lexical matches.

Table 2 ■

Pharmacy Literal Attributes*

Literal Attributes	Yale Pharmacy Terms	MED Pharmacy Terms
Full drug name	√	√
Print name	√	√
Hospital code	√	√
AHFS code	√	√
Dose strength units	√	√
Dose strength number	√	√
Formulary name	N/A	√
Short formulary name	N/A	√
Order entry name	√	N/A
Drug brand name	√	√
Drug generic name	√	√
Drug manufacturer	N/A	√
Drug Rx vs OTC	N/A	√
Drug form code	N/A	√
Drug floor stock	N/A	√
Drug route	√	√
Drug in formulary	N/A	√
Drug volume	N/A	√
Allergy class code	N/A	√
Drug description	√	√
Drug category	√	√
DEA code	√	√
Drug specifier	N/A	√
Drug generic code	√	√
Drug interaction code	N/A	√

*Twenty-five literal attributes were identified for both Yale and Medical Entities Dictionary (MED) pharmacy terms. Twelve of these attributes (boldface) are unique to either Yale or MED terms. N/A = information not available in database; AHFS = American Hospital Formulary Service; Rx vs OTC = prescription versus over-the-counter; DEA = Drug Enforcement Agency.

Results of Semantic Mapping

One hundred forty-five of the 168 lexical matches were semantically equivalent, which resulted in an overall lexical and semantic matching rate of 73%. The 23 lexical matches that were not semantically equivalent were of the minimally similar type.

Semantic equivalence of lexically matched pharmacy terms was easily verified because term names were almost always identical and well defined. Since few questions of semantic equivalence arose, we felt that detailed subjective examination of term relationships was not necessary.

Results of Term Attribute Mapping

In the mapping of term attributes between the Yale pharmacy vocabulary and the MED vocabulary, 30 attributes (25 literal and five non-literal) were identified. We found that a significant number of attributes, literal and non-literal, are present in one vocabulary but not the other. A side-by-side comparison of literal attributes for Yale and MED pharmacy terms appears in Table 2. The names of literal attributes in

Table 2 are a combination of existing Yale and MED attribute names. N/A indicates that an attribute was not present in the online database. Seventeen attributes are listed that are specified for MED terms but are not specified for Yale pharmacy terms. In contrast, only one Yale attribute, *order entry name*, is not present in the MED vocabulary.

The MED attributes listed are common to all pharmacy terms in the MED. They are inherited from five terms (*cpmc formulary drug forms*, *cpmc formulary drug*, *american hospital formulary service*, *drug allergy class*, and *dea control classification*; Fig. 1).

Laboratory Terms

Results of Lexical and Semantic Mapping

Lexical and semantic matching of laboratory terms had an overall success rate of 49%. We did not obtain a separate lexical matching rate because lexical matches for laboratory terms were often meaningless without detailed examination of term relationships. After the domain expert examined the final list of matched and unmatched terms, 33 of the original 200 (17%) Yale laboratory terms were removed from the vocabulary set because they were either inappropriately included, obsolete, or duplicated elsewhere in the randomized 200 terms. Of the removed terms, 27 of 33 were not matched. The removal of these terms improved the overall match rate to 59% (98/167). Terms were considered to be inappropriate for mapping if they were not laboratory tests (such as *blood with special filter* or *pooled platelets*). Additionally, terms were considered to be inappropriate for mapping if they were vague and generic (such as *special blood tests*, *special chemistry tests*, and *special csf tests*). Obsolete terms were laboratory tests that are no longer performed at Yale (such as *teichoic acid antibody*). Finally, duplicates were identical terms that occurred more than once in the list of 200 randomized Yale terms. These duplicated terms were tests that could be ordered either as an individual test or as part of a panel. For example, the Yale term *blood glucose* appears twice in the Yale laboratory vocabulary. It appears once as a freestanding term and again as a component of *glucose tolerance test*, a panel.

Determining semantic equivalence was more difficult for laboratory terms than it was for pharmacy terms. The difficulty was due to a larger number of lexical matches for each Yale laboratory term, and term names from both vocabularies that had ambiguous meanings.

For example, the Yale term *blood potassium* lexically matches to 13 MED terms, which are listed in Table 3. One of the 13 MED terms was *potassium*, which could semantically represent a potassium ion or tests that measure potassium. To determine semantic equivalence, term relationships for *potassium* were examined. *Potassium* is a descendent of *element or ion* and is **measured by** *intravascular potassium test*. The Yale term *blood potassium* measures serum potassium and has attributes such as *low value* or *high value*. Therefore, Yale's *blood potassium* and the MED's *potassium* are not semantically equivalent. A similar examination of term relationships for another of the 13 MED terms, *stat whole blood potassium ion measurement*, proves that it, too, is not semantically equivalent to Yale's *blood potassium*.

Table 3 ■

Lexical Matches for Yale's Blood–Potassium*

Allen whole blood potassium ion measurement
Presbyterian whole blood potassium ion measurement
Stat whole blood potassium ion measurement
Whole blood potassium tests
Allen plasma potassium ion measurement
Chem-7 plasma potassium ion measurement
Intravascular potassium test
New chem-7 plasma potassium ion measurement
Potassium
Presbyterian plasma potassium ion test
Serum potassium ion measurement
Serum potassium ion measurement 2
Serum potassium ion tests

*Thirteen Medical Entities Dictionary (MED) terms lexically match the Yale term *blood–potassium*. Seven of these 13 matches (boldface) are also semantically equivalent to *blood–potassium* as described in the text.

The number of candidate matches can be reduced further by examining the hierarchical relationships of the 11 remaining MED terms (Fig. 2). Four of the 11 terms (*intravascular potassium test*, *whole blood potassium tests*, *chem-7 plasma potassium ion measurement*, and *serum potassium ion measurement*) are higher-level terms. The remaining seven terms (bottom row of Fig. 2) are instances. Laboratory terms will generally correspond to MED instances because both are highly focused terms with values associated with literal attributes. For example, for the Yale term *blood potassium*, the value for *units* is d mmol/L and the value for *low normal* is 3.5; and for the MED instance *serum potassium ion measurement*, the value for *units* is mEq/L and the value for *low normal* is 3.2. Thus, by being able to identify instances through examination of hierarchical relationships, only seven of the 13 MED terms were lexically and semantically equivalent to Yale's *blood potassium*.

Of the 69 terms that remained unmatched, 23% (16/69) of the Yale terms were unmatched due to differ-

Table 4 ■

Laboratory Attributes*

	Yale	MED
Literal attributes		
Name	√	√
Laboratory test names	√	√
Laboratory test code	√	√
Low normal value	√	√
High normal value	√	√
Female low normal value	√	√
Female high normal value	√	√
Male low normal value	√	√
Male high normal value	√	√
Unit	√	√
Test number	√	√
Minimum age	√	N/A
Maximum age	√	N/A
Start date	√	N/A
End date	√	N/A
Can appear as order	√	N/A
Can appear as result	√	N/A
Non-literal attributes		
Part of	√†	√
Specimen of	√†	√
Substance measured by	√†	√
Result type	N/A	√
Child of	N/A	√
Descendent of	N/A	√

*Twenty-three literal and non-literal attributes were identified for both Yale and Medical Entities Dictionary (MED) laboratory terms. Nine of these attributes (boldface) are unique to either Yale or MED terms. N/A = information not available in database.
†Relationships exist though not explicitly stated.

ences in "term granularity." (The remaining Yale terms were unmatched because no corresponding MED term could be identified.) Term granularity refers to the level of detail captured by a term. For example, the Yale term *rbc morphology* represents a set of terms (a panel) describing red blood cell shapes such as *poikilocytosis* and *anisocytosis*. There is no corresponding MED term. However, the concept and terms for measurement must exist at CPMC because they are standard components of any complete blood count test. Another example is the Yale term *gram-positive rods*, which describes the appearance of bacteria on a culture slide. There is no equivalent MED term. Instead, MED terms describe bacterial appearance not only by shape, i.e., gram-positive rods, but by number and size of bacteria, with terms such as *moderate large gram-positive rods*, and *moderate small gram-positive rods*.

Results of Attribute Mapping

A comparison of 23 (17 literal and six non-literal) Yale and MED laboratory attributes can be seen in Table 4. For attribute names, a combination of existing attribute names for both Yale and MED laboratory terms was used. N/A again indicates that the attribute was not present (but does not imply that the information needed to create and assign a value to an attribute does not exist).

When laboratory attribute mapping is compared with pharmacy attribute mapping, there are noticeable differences. Though there are fewer total laboratory attributes to map (23 laboratory attributes compared with 30 for pharmacy), the percentage of shared attributes is higher [i.e., 14/23 (61%)] for laboratory than it is for pharmacy [i.e., 12/30 (40%)]. Finally, unlike Yale pharmacy terms, Yale laboratory terms can possess non-literal attributes because they possess hierarchical relationships.

The MED attributes listed are common to all laboratory terms in the MED. These attributes are inherited from one laboratory term (*laboratory diagnostic procedure*). Figure 2 illustrates this point graphically but does so only for "chemistry" terms. Other hierarchical relationships not shown in Figure 2 can demonstrate that *laboratory diagnostic procedure* is the "common ancestor" of all laboratory terms.

Discussion

In mapping Yale's pharmacy and laboratory vocabulary to the MED, several issues arose, which are likely to occur in other such efforts.

Degree of Standardization in Subsets of a Local Vocabulary

Certain subsets of a local vocabulary (e.g., pharmacy) may be more amenable to mapping to a structured vocabulary than others (e.g., laboratory). For example, mapping Yale pharmacy terms to the MED is easier because the pharmacy vocabulary is based on the standardized vocabulary of the pharmaceutical industry. We define a standardized vocabulary as a vocabulary with term nomenclature, classification, usage, and maintenance agreed upon across multiple institutions, sites, or authorities. Such a vocabulary is usually a controlled vocabulary, but not necessarily so.[10,14–16] Since term names are consistently applied, automated lexical matching requires a relatively small amount of manual verification. Forty-one percent of the Yale pharmacy terms mapped to an identical MED term at the level of an instance (the most precisely defined level of term possible in the MED). In addition, some of the ease and success of pharmacy term mapping may have been due to similar drug dispensation practices at the two institutions.

We had initially expected that the meaning of terms from a standardized vocabulary would be clear, and that as a result semantic equivalence would be easy to determine. This was true with the pharmacy vocabulary. An additional finding was that semantic relationships of pharmacy terms in the MED were applicable to Yale terms without modification. Further, many of the MED literal attributes and the values assigned to them are also applicable to Yale terms without modification. Thus, the MED's *allergy class code 00* applies to both the MED's *digoxin .125 mg po* and Yale's *digoxin .125 mg po*. Given that term names and meanings are standardized, it is not surprising that matched terms would have identical relationships and values for shared characteristics. It is also not surprising that differences in literal attributes between matched terms could be characterized as mostly administrative (i.e., institution-specific). For example, the MED attributes *drug in formulary* and *drug floor stock* are administrative and not specified for Yale pharmacy terms.

Laboratory vocabulary was mapped with more difficulty because nomenclature and meaning are nonstandardized.[17–19] AMIA's Board of Directors noted that a standard "is sorely needed."[18] Thus, Yale and MED laboratory terms were often institutional terms that were chosen to meet unique local needs and situations. An alternative approach would be to use a standardized set of names and codes as an inter-

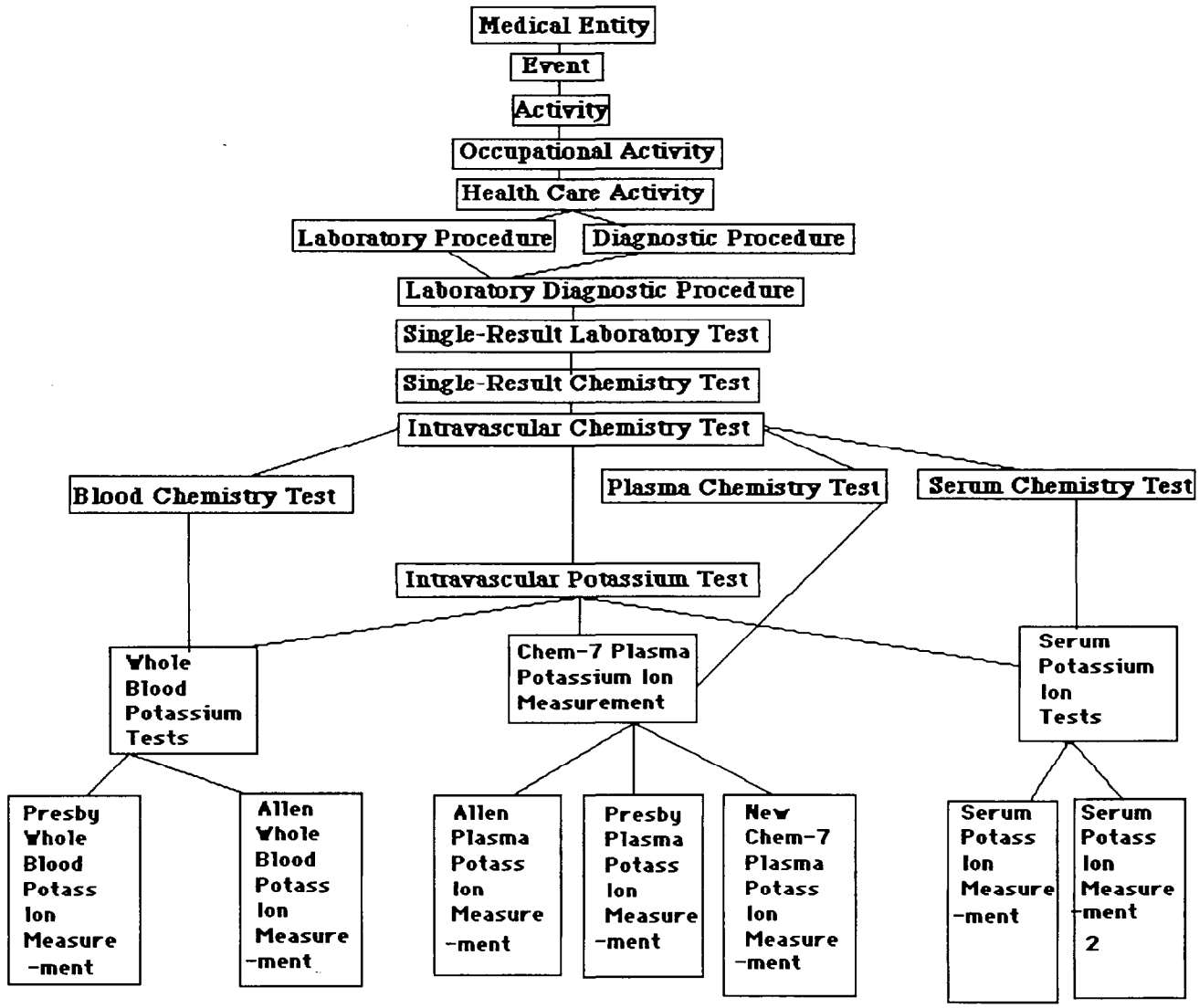

Figure 2 Family tree of hierarchical inheritance for Medical Entities Dictionary (MED) potassium terms. The boxes represent terms and the lines between the boxes are hierarchical parent–child relationships. The bottom level of terms are instances.

lingua for mapping laboratory terms. For example, Yale and MED laboratory terms could be mapped to this standardized set and then mapped to each other. The advantage of this approach is that once terms are mapped to this interlingua, it becomes relatively easy to map laboratory terms from other institutions. A candidate set of names and codes was developed by the LOINC (Laboratory Observation Identifier Names and Codes) Consortium, which consists of the Regenstrief Institute and 11 other institutions.[19] The more than 6,000 identifier names and codes, derived from multiple sources, were designed to serve as a standard to which institutions could map their laboratory terms. Currently LOINC does not deal with panels or orders, nor does it have a scheme for standardizing differences in reporting normal results.

Given that laboratory term names and meanings are not standardized, it is not surprising that matched terms would not share all relationships and values for shared characteristics. In contrast to the administrative differences between pharmacy attributes, the differences between laboratory attributes are mostly clinical. For example, the Yale literal laboratory attributes *low age*, *high age*, *start date*, and *stop date* are clinical and not specified for MED laboratory terms.

Representation of Attributes for Vocabulary Terms

The attempt to map from Yale's laboratory vocabulary to the MED illustrated a number of the problems that arise because of different representations of attributes for vocabulary terms.

An example of the type of problem that occurs is seen in the representation of information about normal range of a laboratory test, and the various qualifiers that apply to such a normal range. In the Yale vocabulary, some tests have a single normal range. For other tests, however, the normal range was dependent upon the patient's age. For example, there might be a pediatric normal range and an adult normal range. In addition, a test's normal range might change at a specified date, for example, when new equipment is installed in the laboratory to perform the test.

1. In the files describing the Yale laboratory tests, this information is stored in tables that include textual fields describing each test. For example, the term *hematocrit-blood* has ten different normal ranges stored in records. Two of the normal ranges specify normal values for adult male and female patients. The remaining eight normal ranges specify normal values for pediatric patients of different ages.
2. In the MED, there are few pediatric tests, and these are entered as separate entities. Thus, there are separate terms for adult bilirubin tests and for pediatric bilirubin tests. In addition, when the normal values of a test change, then a new MED entry is created to represent the new variant of the test.

To allow ready mapping from one clinical vocabulary to another, neither of these approaches is satisfactory. If the information is stored in part as free text, as at Yale, then it cannot be accessed in an organized way. On the other hand, if different linguistic entities are created to represent all the permutations of, for example, age ranges and occasional changes in the range of normal values, then one is not capturing the test as a single linguistic term, which in turn makes it difficult to map among vocabularies.

For these reasons, a standardized model is needed to capture literal attribute information for each term. A simple but powerful model would be an association list consisting of name–value pairs. Using this model, the normal values for a given laboratory test might be represented by a series of association lists, such as:

[high_normal 44, low_normal 22,
age_range_in_years 0–12,
date_started 6/1/83, date_stopped 9/3/94]

One could associate as many such lists with a test as necessary to describe the different qualifiers that apply to that test, and the number of times the information has been modified. The set of names used in the association list would need to be refined by examining a broader spectrum of terms than those looked at in the present study. If the information describing a test's literal attributes were represented in such a standardized fashion, there would be a number of potential advantages.

1. It would be more straightforward to map any institution's test to another, or to a controlled, clinical vocabulary. Institution-specific information about the test could be stored in a standardized way.
2. A single term could describe a number of possible variants of a test. Only if a test changed in a truly fundamental way would a new term need to be created.
3. A standardized description of a term's literal attributes would also facilitate the comparison of clinical logic and practice patterns across different

institutions, since the information about what laboratory value was considered normal or abnormal, for example, at any given time would be associated with the laboratory test in an organized, standardized fashion.

Granularity

Matched terms derived from standardized vocabularies, such as the pharmacy vocabularies of Yale and MED, are similar in term granularity (the level of detail represented by a term). In contrast, matched terms derived from non-standardized vocabularies, such as the laboratory vocabularies of the MED and Yale, may have differences in granularity. These differences can be divided into coarse and fine differences. Differences in granularity create mapping difficulties because matched terms may have somewhat different though overlapping meanings.

Coarse differences arise from differences in concept representation. For example, Yale has the terms *vdrl blood, cord (fetal)* and *coombs direct, cord*. An argument could be made that *vdrl blood, cord* and *type and screen, cord* are simply variants of MED terms and should be mapped to the MED's *blood antitreponemal antibody measurement* and *allen direct coombs test*. However, another argument could be made that the concept being represented by Yale's *vdrl blood, cord* and *coombs direct, cord* is tests on umbilical cord blood, and there is no MED term that represents this concept. Thus, a coarse difference in granularity occurs between terms that share either the concept of vdrl-blood or coombs-direct but do not share the concept of blood tests on cord. As a result of the coarse granularity difference, these terms do not map to each other precisely.

Fine granularity differences involve differences in the number of terms required to represent the smallest possible concepts such as a test or a panel. For example, the concept "gram stain smear result: positive wbc" is represented in the MED by three terms: *many wbc, moderate wbc,* and *few wbc,* while at Yale it is represented by one term, *positive wbc*. This disparity creates mapping difficulty because instead of there being one MED term that captures the meaning of the Yale term there are three MED terms, each of which captures part of the meaning of the one Yale term. Once again, matched pairs of terms cannot be mapped to each other precisely.

Differences in fine granularity would also be expected with test panel names. Laboratory vocabulary is non-standardized, and panels (which group tests) have even more variability. For example, *viral serology* is a Yale panel term and the equivalent MED terms are *viral antibodies* and *blood viral antibody tests*. The MED and Yale panels share 11 identical test terms. However, Yale's *viral serology panel* possesses two tests not represented in either the MED's *viral antibodies* or *blood viral antibodies tests*. Additionally, the MED panels contain some test terms that the Yale panel does not. This again creates difficulty in mapping because meanings are similar but not identical.

Hidden Term Usage

Non-standardization of laboratory vocabulary results in another difficulty, which might be called "hidden term usage." Hidden terms represent hospital laboratory tests but are not listed in either Yale or the MED's laboratory vocabulary because, for example, these tests are done outside the institution in specialized research laboratories. The existence of hidden terms results in unnecessary exhaustive searches since it is hard to distinguish between hidden terms and faulty searches for terms that one has a high expectation of finding. An example is Yale's *western blot, anti-hiv-1-blood*. There is no matching term in the MED because in New York the test is always sent out to the New York City's Department of Health and is therefore not currently represented in the MED.

Appropriateness

Finally, not all mapping difficulties are caused by differences in vocabulary standardization. Both the Yale laboratory and pharmacy vocabularies possess terms that represent concepts inappropriate for matching. Example include clinical trial research terms such as *librium-urine* and *erythromycin/placebo*. It is not clear whether such terms should be rejected as being inappropriate for mapping, or whether they should be considered terms that fail to map and need to be considered for addition to the structured vocabulary.

Conclusion

Institutions with pre-existing systems and data will face the challenge of mapping their local vocabularies to the developing structured vocabularies. Such mapping is performed at the level of terms, relationships, and attributes. This study of the mapping process in the context of a structured vocabulary developed at one institution, and an unstructured vocabulary from another institution, identifies issues that will need to be addressed. Problems relate to a lack of standardization in use of terms, representation of attributes, and granularity of concepts.

References ■

1. Evans DA. Pragmatically-structured, lexical-semantic knowledge bases for unifed medical language systems. Proc Annu Symp Comput Appl Med Care. 1988:169–73.
2. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA. 1994;1:35–50.
3. Humphreys BL, Lindberg DAB. The Unified Medical Language System Project: A Distributed Experiment in Improving Access to Biomedical Information. In: Degoulet KCLP, Plemme TE, Rienhoff O, eds. MEDINFO 92. New York: Elsevier Science Publications, 1992;1496–500.
4. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. MD Comput. 1990;7:104–9.
5. Sherertz DD, Tuttle MS, Blois MS, Eribaum MS. Intervocabulary mapping within the UMLS: the role of lexical matching. Proc Annu Symp Comput Appl Med Care. 1988:201–6.
6. National Library of Medicine. Medical Subject Headings. Bethesda, MD: NLM, 1992.
7. Evans DA, Rothwell DJ, Monarch IA, Lefferts RG, Côté RA. Towards Representations for Medical Concepts. Med Decis Making. 1991;11(suppl):S102–S108.
8. Côté RA, ed. The Systematized Nomenclature of Medicine. Northfield, IL: College of American Pathologists, 1982.
9. Pryor TA, Clayton PD, Haug PJ, Wigertz O. Design of a knowledge driven HIS. Proc Annu Symp Comp Appl Med Care. 1987:116–21.
10. Masarie FE Jr, Miller RA, Bouhaddou O, Guise NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24:379–400.
11. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD. Designing an introspective, multipurpose, controlled medical vocabulary. Proc Annu Symp Comput Appl Med Care. 1989:513–8.
12. United States National Center for Health Statistics, ed. International Classification of Diseases, Ninth Revision, with Clinical Modifications. Washington, DC: National Center for Health Statistics, 1980.
13. Huff SM, Warner HR. A comparison of Meta-1 and HELP terms: implications for clinical data. Proc Annu Symp Comput Appl Med Care. 1990:166–9.
14. Ball MJ, Cohen MF, eds. Aspects of the Computer-based Patient Record. New York: Springer-Verlag, 1992.
15. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Towards a medical-concept representation language. JAMIA. 1994;1:207–17.
16. Rocha RA, Rocha BHSC, Huff SM. Automated translation between medical vocabularies using a frame-based interlingua. Proc Annu Symp Comput Appl Med Care. 1993:690–4.
17. Mendenhall S. The ICSS code: a new development for an old problem. Proc Annu Symp Comput Appl Med Care. 1987:703–9.
18. Board of Directors of the American Medical Informatics Association. Standards for medical identifiers, codes, and messages needed to create an efficient computer-stored medical record. JAMIA. 1994;1:1–7.
19. Regenstrief Institute and LOINC Committee. Laboratory Observation Identifier Name and Codes (LOINC™) Users Guide, version 1.0. Indianapolis, IN: Regenstrief Institute, 1995.