Original Investigations JAMIA

Research Paper ■

Knowledge-based Approaches to the Maintenance of a Large Controlled Medical Terminology

JAMES J. CIMINO, MD, PAUL D. CLAYTON, PHD, GEORGE HRIPCSAK, MD, STEPHEN B. JOHNSON, PHD

Abstract **Objective:** Develop a knowledge-based representation for a controlled terminology of clinical information to facilitate creation, maintenance, and use of the terminology.

Design: The Medical Entities Dictionary (MED) is a semantic network, based on the Unified Medical Language System (UMLS), with a directed acyclic graph to represent multiple hierarchies. Terms from four hospital systems (laboratory, electrocardiography, medical records coding, and pharmacy) were added as nodes in the network. Additional knowledge about terms, added as semantic links, was used to assist in integration, harmonization, and automated classification of disparate terminologies.

Results: The MED contains 32,767 terms and is in active clinical use. Automated classification was successfully applied to terms for laboratory specimens, laboratory tests, and medications. One benefit of the approach has been the automated inclusion of medications into multiple pharmacologic and allergenic classes that were not present in the pharmacy system. Another benefit has been the reduction of maintenance efforts by 90%.

Conclusion: The MED is a hybrid of terminology and knowledge. It provides domain coverage, synonymy, consistency of views, explicit relationships, and multiple classification while preventing redundancy, ambiguity (homonymy) and misclassification.

■ **J Am Med Informatics Assoc.** 1994;1:35–50.

A controlled medical terminology* is a fundamental requirement in a range of medical informatics applications, including hospital departmental systems, patient record systems, expert systems, and medical literature databases. The construction of small terminologies for narrow domains is usually straightforward. Problems arise when larger domains are considered. Despite several national and international efforts to develop comprehensive controlled medical terminologies, none has resulted in a well-accepted product suitable for use in clinical infor-

Affiliation of the authors: Columbia University, New York, NY.

Supported under the National Library of Medicine's Integrated Advanced Information Management Systems (IAIMS) project and by the IBM Corporation. Drs. Cimino and Johnson are supported in part through a Unified Medical Language System contract with the National Library of Medicine.

Correspondence and reprints: James J. Cimino, MD, Center for Medical Informatics, Atchley Pavilion, Room 1310, Columbia–Presbyterian Medical Center, New York, NY 10032.

Received for publication: 4/15/93; accepted for publication: 6/09/93.

*The terminology in this document adheres to the guidelines established by the International Standards Organization[1] and the American Society for Testing and Materials.[2] For example, this document discusses work on *terminology* ["set of terms representing the system of concepts of a particular subject field"] as opposed to *vocabulary* ["terminological dictionary containing the terminology of a subject field or of related subject fields (admitted term: glossary)"]. Occasional license is taken with the strict technical definitions to improve readability.

mation systems. Creating a controlled terminology requires that an author address such issues as what to include, what classification scheme to use, and what level of granularity to provide. Such decisions can be made arbitrarily, but they must be applied consistently if others are to understand and use the terminology. We have proposed that a sophisticated, knowledge-based representation will facilitate the development of more advanced terminology management techniques.[3] This paper describes our experiences with that approach.

Background

Problems Encountered in Terminologic Work

Suppose that a terminology is to include terms for various forms of pneumonia. Suppose also that, for design considerations, each disease term in the terminology is to be placed in a particular class. A generic concept corresponding to that class might be identified by the term "pneumonia." Finally, suppose that some contributor to the terminology wishes to include a term for a particular pneumonia, such as "staphylococcal pneumonia." However, instead of adding the new term as a member of the generic class pneumonia, the contributor decides to place it with diseases with like etiology (e.g., "staphylococcal diseases"). Such inconsistent classification would probably cause confusion for users of the terminology. Similarly, the inclusion of the same disease concept as two different terms with two different names presents redundancy that can be impossible for the user to detect. Many large terminologies, such as the *International Classification of Diseases, Ninth Edition, with Clinical Modifications* (ICD9-CM),[4] the *Systematized Nomenclature of Medicine* (SNOMED),[5] the Medical Subject Headings (MeSH),[6] and the *Current Medical Information and Terminology* (CMIT),[7] are replete with examples of such inconsistencies.[3]

The origins of specific inconsistencies are difficult to trace; however, the methodology by which terminologies are created and maintained could be a contributing factor. As the number of terms in a terminology reaches into the thousands, it exceeds the capacity of individual human memory, making it nearly impossible for an author to recall such essentials as whether a term already exists (under another name) or how similar terms are classified. When a large controlled terminology is created by a committee, the author's problem is compounded by being unaware of decisions made by other authors. Computer-based tools may be able to address these problems.

The simplest tool for terminology authoring and maintenance is one that can perform lexical searches of the terminology. For example, if an author is considering the addition of "Wilson's disease" to a terminology, a search for the exact phrase will find the term if it exists. With stemming, a search for "Wilson" will find the term whether it exists as "Wilson's disease," "Wilson disease," or "Disease, Wilson's." This will work for lexical variants but will fail if the search is performed using a synonym, such as "hepatolenticular degeneration." However, if the terminology includes additional information, such as disease classification, then more sophisticated searching can be performed. For example, a search for "copper" might find "disorders of copper metabolism." The terminology author could then retrieve all terms in that class to see whether Wilson's disease appeared there in one of its various forms.

The Unified Medical Language System™

In 1987, the National Library of Medicine's Unified Medical Language System (UMLS) project established the design for what was later to be a set of knowledge sources to facilitate the use of disparate medical terminologies for accessing a variety of medical information sources.[8] The early design included a Semantic Network of interrelated semantic classes and a Metathesaurus of concepts and names, also interrelated, with each concept assigned to one or more classes in the Network. (Although the models and contents of the UMLS have evolved, the basic description of the design remains valid.[9]) Work by UMLS contractors has explored a variety of aspects of the use of semantic networks. A semantic network is a notation for representing conceptual entities and links between them, allowing the storage of factual knowledge (as opposed to procedural knowledge) that can be intensional (describing the entities themselves) and extensional (describing how entities are related to other entities).[10] The early UMLS work explored the use of such schema for representing controlled medical vocabularies including, for example, graphic browsing[11] and intervocabulary translation.[12]

Formal Approach to Computer-assisted Terminologic Modeling

In 1988, the Center for Medical Informatics at Columbia University began to develop a model for the controlled medical terminology to be used in the clinical computing environment of the Columbia–Presbyterian Medical Center (CPMC). The system, which has continued to evolve since that time,[13] accumulates patient data from ancillary departmental systems,[14] stores the data into a central patient da-

tabase,[15] provides access to on-line medical information sources,[13] and provides automated decision support.[16]

In our planning, we envisioned several roles for the terminology. First, it would have to be capable of translating information to and from local hospital systems, such as those in the clinical laboratory and pharmacy. This would allow patient data to be transferred, interpreted, and coded for storage in the central database and then reconstructed as necessary (for example, for reporting results). Second, the terminology would have to contain information needed to map the local terminologies to standard terminologies, such as ICD9-CM and the UMLS. This capability would provide for information transfer to outside systems, for purposes such as literature retrieval[17] or government reporting requirements. Third, the terminology would have to allow integration of disparate terminologies to serve common functions, such as automated decision support. For example, if a decision support system queries the clinical database to determine whether electrocardiography has been performed, there must be a way to map the decision support query term to corresponding terms in several cardiology systems or in ICD9-CM. With these tasks in mind, we established the terminology design criteria shown in Table 1.[1,18]

Given this set of criteria and the awareness that a single central clinical terminology was likely to be quite large, it was clear from the outset that computer-based terminology maintenance tools would be needed. It appeared that sophisticated tools could be developed only if the terminology representation itself was sufficiently sophisticated. Significant research had advanced the notion that controlled medical terminologies require a deeper representation than the traditional tree structure (with or without synonyms). In particular, the inclusion of structured knowledge about terminology terms had been proposed by a number of workers for such purposes as patient data representation,[19] computer-assisted indexing of the medical literature,[20] interterminology translation,[12,21] and automated decision support.[22,23] The idea that the representational scheme used for recording controlled terminologies could be useful for tasks involved in the construction of the terminology has also been put forward[24] and computer-based medical knowledge has been demonstrated to

Table 1 ■

Criteria for Controlled Medical Terminology

Criterion	Description
Domain completeness	Must not restrict terminology size through presuppositions about ultimate dimensions (e.g., no preset coding system that restricts depth or breadth of the hierarchy)
Nonredundancy	A mechanism must exist that can help prevent multiple terms for the same concept from being added to the terminology as unique concepts
Synonymy	Support multiple nonunique names for concepts
Nonvagueness	Concepts in the terminology must be complete in meaning (e.g., "ventricle" is not usually considered a fully described concept, nor does it represent some generic class of anatomic terms, i.e., it means neither "heart ventricle" nor "brain ventricle" when taken out of context)
Nonambiguity	Concepts must have exactly one meaning and, where a common term has two or more associated meanings (homonymy), they must be disambiguated into distinct concepts (e.g., "Paget disease" must be split into "Paget disease of the bone" and "Paget disease of the breast")
Multiple classification	Must not restrict terminology such that a concept is prevented from being assigned to as many classes as required (e.g., "viral pneumonia" can be in classes "pneumonia" and "viral diseases")
Consistency of views	Concepts in multiple classes have the same appearance in each context (e.g., corticosteroid as hormone or anti-inflammatory agent has the same attributes and descendant concepts)
Explicit relationships	Meanings of inter-concept relationships must be clear (e.g., relationship between staphylococcal pneumonia and pneumonia is differentiated from relationship between staphylococcal pneumonia and staphylococcus, where the former is a class relation and the latter is an etiologic relation)

be useful for vocabulary maintenance.[25] In addition to UMLS investigators, other researchers have proposed using semantics and semantic networks for vocabulary representation.[26–28]

Based on our collective experiences with medical terminologies and the work of others, we chose to explore a knowledge-based representation for the controlled medical terminology to be used in the CPMC clinical system. In the following sections of this paper, we describe our methods for developing a representational scheme and the results of our attempts to build and maintain a terminology that adheres to the aforementioned design criteria.

Methods

Terminology Model

Our terminology, called the Medical Entities Dictionary (MED),[3] uses a semantic network model that includes a classification hierarchy. Each concept in the terminology is assigned a unique integer as an identifier (the MED Code) and a unique name (the MED Name). One concept, called "Medical Entity," serves as the topmost node in the classification. All other concepts are nodes in this graph, as immediate descendants of at least one other node. These parent–child relationships between nodes correspond to the classification of the concepts. Each concept may have several parents; however, these relationships are acyclic—that is, a concept may not be its own descendant. Thus, the MED hierarchy is defined by a directed acyclic graph.

Concepts in the network have named attributes that may or may not have values. An attribute is "introduced" at a single node in the graph and is inherited by all nodes descended from that node. For example, the attribute "MED Code" (the unique identifier) is introduced in the node for the concept "Medical Entity" and is inherited by all nodes in the MED. *Literal attributes* may have numeric or string values, such as "MED Code," which takes an integer, or "MED Name," which takes a character string. *Semantic attributes* may take one or more references to MED concepts. Each semantic attribute has a specification (called the *domain*) for restricting the class of concepts that can be values. For example, "Has Part" (introduced in "Medical Entity") may refer to any valid MED concept, while "Substance Measured" (introduced in "Diagnostic Procedure") may refer only to MED concepts in the class "Measurable Substance." Figure 1 shows the MED structural model with sample concepts.

Each concept node in the MED graph can be viewed as a frame with slots, in which some of the slots have values. Each node can also be viewed as having links to nodes other than parent–child nodes through the semantic attributes. When such an attribute includes a reference to another MED concept, the reference is a *semantic link*. A semantic attribute in the MED is always paired with second, inverse, attribute which is introduced into the graph at the node that corresponds to the domain of the first attribute. For example, the inverse of "Has Part" is "Part Of." Both attributes have "Medical Entity" as their domain and are introduced in that concept. "Substance Measured" is paired with the attribute "Measured By." The latter attribute is introduced in the node "Measurable Substance" and has the domain "Diagnostic Procedure."

A hierarchical data structure was designed to represent this network of frames, using PC-MUMPS (Datatree, Waltham, MA). The MUMPS structure of the MED consists of an array of elements, indexed by MED Code, with each element corresponding to a MED concept. Each element is a data structure with fields, or slots, corresponding to the attributes assigned to the MED Concept. Slots hold delimited lists of values, either MED Codes or strings, corresponding to attribute values. Hierarchical structure is provided by the slots "Subclass" and "Subclass Of."

The MED was initially populated with the 131 semantic classes of the second version of the UMLS knowledge sources[29] (except where noted, all mention of the UMLS in this paper refers to the second edition), plus additional classes that were needed to classify concepts or introduce attributes (for example, "Measurable Entity" is not a UMLS semantic type). Terms drawn from various clinical terminologies could be included in the MED as descendants of these high-level semantic classes. Semantic links were added as appropriate to represent information known about the concepts. The hierarchical information and the semantic links constitute the knowledge that was applied to the task of MED construction and maintenance.

Each hospital ancillary clinical system that provides coded data to the central database must have its terminology represented in the MED, in order to allow translation of clinical data into MED Codes. To represent each ancillary system, three steps are involved in MED maintenance: *modeling* of information, *addition* of terms, and *maintenance*. The remainder of this section describes the tools developed and methods applied to accomplish these tasks for the ancillary-system terminologies integrated into the MED thus far.

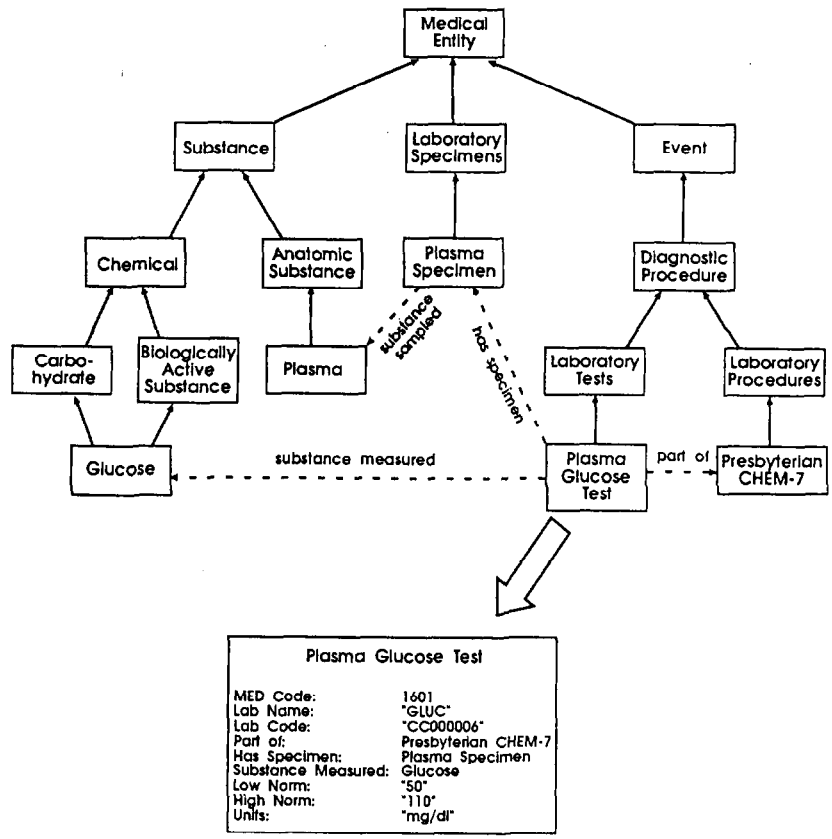

Figure 1 Sample representation of concepts in the Medical Entities Dictionary (MED). Boxes represent concepts in the MED's semantic network, solid arrows represent generic relationships (that is, hierarchical "is-a" relationships), and broken arrows represent nonhierarchical (but nonetheless semantic) relationships. High-level terms are drawn from the UMLS Semantic Network, while lower-level terms are from ancillary systems or the UMLS Metathesaurus (see text). The node for "Plasma Glucose Test" has been expanded (white arrow) to show the associated frame, with its literal and semantic attributes.

MED Editing Tools

A MUMPS-based vocabulary browser[30] was adapted to serve as the MED Editor. The program provides a windows-and-mouse, point-and-click environment with two windows. One window displays information about a single MED concept, organized as a frame with slots and values. The second window displays the ancestors of the concept in a hierarchical fashion. Editor functions include *browsing* (viewing concept frames and traversing the semantic net through point-and-click or string look-up), *concept creation* (placing new concepts in the hierarchy), *attribute creation* (creating and filling slots for concepts), *concept modification*, and *automated classification* (discussed later).

Modeling Candidate Terminologies in the MED

In order to model the terminology in an ancillary system, the following information had to be compiled: 1) the general collections, or groups of terms (e.g., all laboratory test terms or all drug terms); 2) general interactions between term groups (i.e., the ways in which specific terms from one group interact with those in another group); 3) other typical attributes of terms in each group; 4) classification (if any) of terms within the general groups; 5) the members of each group; 6) specific relationships between terms in the same or different groups; 7) additional information about each term; and 8) a clear understanding of what this information means to the ancillary system and to potential users of it in the central system (such as health care workers reviewing it, or the decision support system reasoning with it).

A model of the classes, attributes, concepts, and semantic links was formulated for four ancillary systems: laboratory, electrocardiography, medical records (diagnostic and procedure coding), and pharmacy. For example, laboratory terminology consists of four classes of terms: procedures (or test panels), tests (components of procedures), specimens, and results. Procedures are related to tests and specimens; tests are related to procedures, specimens, and results; specimens are related to procedures and tests; and results are related to tests. Inter-term relations were added to represent the meanings of terms, to resolve vagueness and ambiguity, to detect redundancy, and to assist in the assignment of concepts to appropriate multiple classes. For example, laboratory tests can be defined by the substances they measure, while specimens can be defined by the body part, body fluid, or other material sampled to perform a test. Literal attributes included procedure codes, test codes, specimen codes, normal ranges, and test units.

Adding Candidate Terminologies to the MED

The information collected in the modeling step was used to add concepts to the MED by the following method: 1) concepts were created in the MED, as descendants of appropriate UMLS semantic types, to correspond to each term group; 2) interactions between term groups were represented as semantic attributes of the term group concepts and the other attributes were represented as literal-valued attributes; 3) if terms in the group were organized into a classification scheme, the classes were added as concepts that descended from the term group concept and, where appropriate, each other; 4) the terms themselves were added as concepts descended from the term group concepts or, if present, the classification concepts; 5) relationships between specific terms were added as semantic links between corresponding MED concepts and additional information was added to literal-valued attributes. The first two steps were carried out manually using MED Editor functions. The remaining steps were performed with editor functions in batch-processing mode with term files obtained from the ancillary systems.

Making Implicit Knowledge Explicit

The addition of knowledge describing the ancillary terminologies was also carried out in a series of steps. First, the type of information to be represented was determined and a semantic relationship was defined that could be used to express that information. Second, a class in the MED was selected (or created, if necessary) to contain the concepts to be linked to the terms. Third, the specific concepts for use in the semantic links were collected and added to appropriate classes in the MED. Fourth, the semantic links between concepts were created in the MED. With laboratory tests, for example, it was determined that laboratory tests measure some substances. Therefore, a new class, called "measurable substances," was created, to subsume UMLS semantic types, such as "chemical" and "cell." The class "laboratory diagnostic test" was linked to the "measurable substances" class by a "substance measured" link. Specific concepts (such as glucose) were drawn from the UMLS and were added to the MED and linked to the appropriate terms (such as serum glucose test).

The first two steps were accomplished through manual interaction with the MED Editor. The third step, collection of specific terms, was accomplished by browsing the UMLS Metathesaurus for appropriate terms. For laboratory tests, this was done by searching the Metathesaurus for concepts (of the appropriate semantic types, such as chemicals and cells) that had one or more words in common with the test name. These terms were then added to the MED as subclasses of their corresponding UMLS semantic types. Finally, the specific links were added through a batch process.

Automated Classification of Terminologies

Classes provided within the ancillary systems were inadequate for the MED hierarchy, both for the multiple-classification criterion and for use in clinical applications. A *subclassification* function was added to the MED Editor to create new classes of concepts and ensure that the appropriate concepts were included in those classes. Adapted from previous experiments with automated classification,[31] the subclassification function uses knowledge associated with each concept, either derived from the ancillary system or added during the incorporation of the concept into the MED. It also detects natural groupings within members of a class. The program requires two pieces of information: the class of terms on which to operate and the attribute by which to form subclasses. It identifies an attribute value that can be used to separate members of the class into those that have the value and those that do not. It then proposes to create a new class in which to include the former. For example, the subclassification function was applied to the laboratory tests class and was instructed to classify based on the attribute "substance measured." During the program's review of measured substances for laboratory tests, it encountered a test with attribute value "glucose." The program then discovered that other tests shared this attribute value and suggested a new subclass laboratory test, linked to glucose and subsuming the appropriate tests.

The program was designed to detect similarities not just based on exact attribute values but also by using classes of values. Figure 2 shows the automated classification of viral hepatitis antibody tests. In Panel A, the original arrangement is shown: all tests were the immediate descendants of the concept "Laboratory Test." Also shown in Panel A are the links between each test and the substance it measures (i.e., antibodies to hepatitis viruses). The subclassification program detected that, while no two of these tests measured the same substance, some did measure substances from the same class. The results of this process can be seen in Panel B: two new classes, "Anti-Hepatitis A Antibody Tests" and "Anti-Hepatitis B Antibody Tests" were created to subsume the IgM and IgG tests for each virus. When the function was applied again to the class of laboratory tests, the class "Viral Hepatitis Antibody Tests" was created to

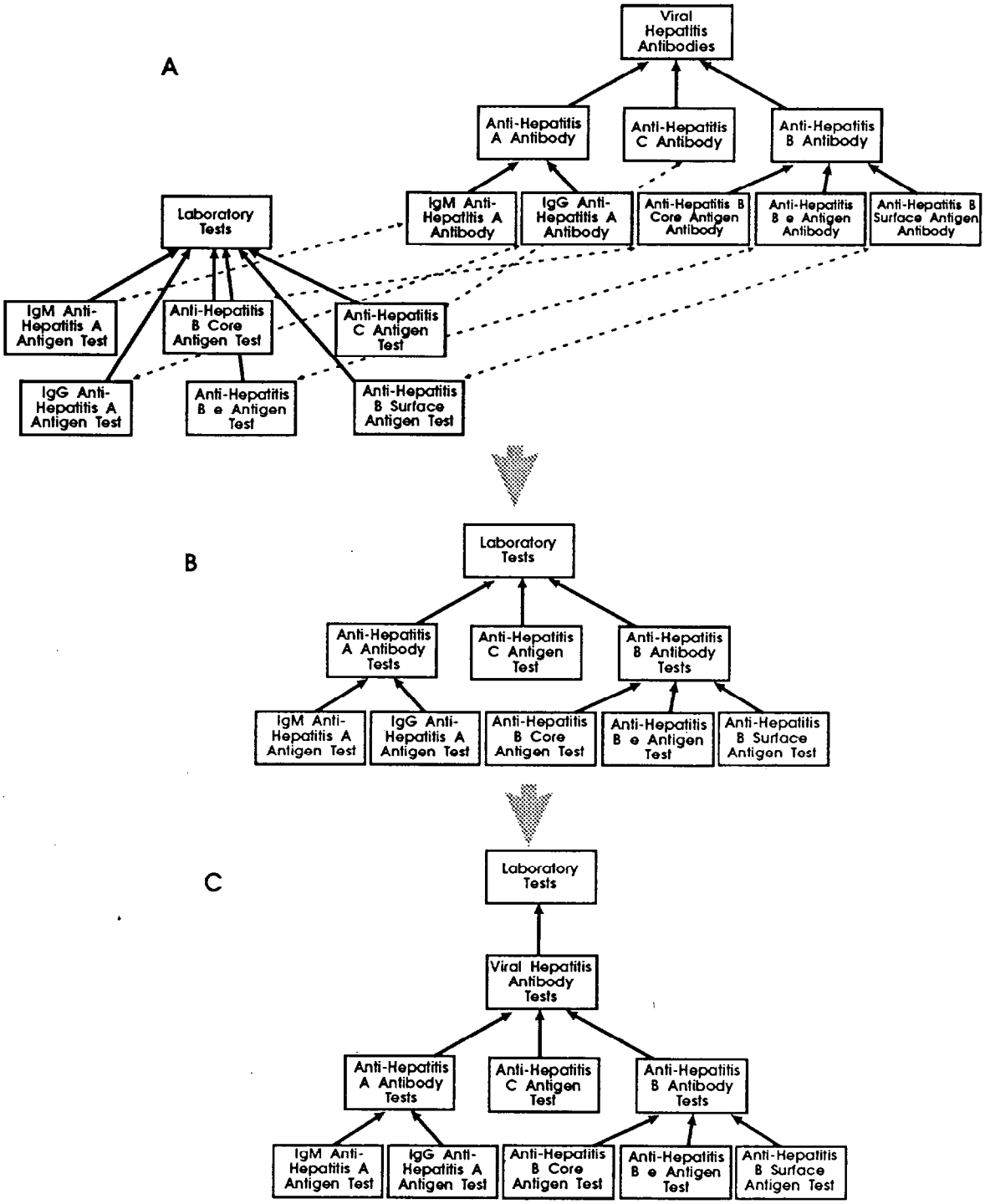

Figure 2 Automated classification using semantic information. In *A*, a subset of hepatitis tests is shown, imported from the laboratory system into the Medical Entities Dictionary. Solid arrows show initial, two-layer hierarchy. Also shown are chemical terms imported from the UMLS. Broken arrows represent semantic links between tests and the chemicals they measure. In *B*, information about measured substances was used to group tests into intermediate classes. The new classes are linked automatically to appropriate chemical concepts (not shown). In *C*, further subclassification resulted in grouping the intermediate classes into a single common class.

subsume the two new classes and one additional concept, as shown in Panel C. The subclassification program was applied repeatedly to each new class of ancillary system terms until no new class could be detected. Previous work describes the algorithm in detail.[31]

Most of the ancillary terminologies added to the MED undergo periodic updates. Maintaining their currency within the MED requires not only adding new terms and information about those terms, but also determining where those terms should be placed in the MED hierarchy. Furthermore, when a pre-existing term in the ancillary terminology is changed, not only must the changes be made in the MED, but it must be determined whether the changed information might cause the corresponding MED concept to be reclassified. Most of the update work can be carried out using the MED Editor functions already described. In order to identify the appropriate class for a new term or for a term whose classification information has changed, an *autoclassification* function was developed.

The autoclassification program attempts to identify the most appropriate location for a given term in a given classification hierarchy. The algorithm compares the new term with subclasses in the selected class to see whether it can be "pushed" beneath one or more of them. The comparison is based on the semantic slot values; in each case, the value for the new term must be equal to or a subclass of the value in the subclass under consideration. For example, if the concept is a test that has glucose as the "substance measured" (e.g., "glucose tolerance test") and the class has "chemical" as the "substance measured" (i.e., the class is "chemical tests"), then the test would be placed in the class since the MED includes the knowledge that glucose is a chemical. The process can be repeated to reclassify the test into the "carbohydrate tests" class and the "glucose tests" class. This function was used for applying changes received from the laboratory and pharmacy systems. Expressed as pseudocode, the algorithm is:

```
FOR each subclass of the class, where the subclass is not the new term
  FOR each slot in subclass
   FOR each value in slot for the subclass
    FOR each value in the slot for the new concept
     IF new concept's value equals the subclass's value
     OR new concept's value is a subclass of the subclass's value
     THEN succeed
    IF ANY value in slot for the new concept succeeds, THEN subclass slot value succeeds
   IF ALL values for the slot in the subclass succeed, THEN the subclass slot succeeds
  IF ALL slots for subclass succeed, THEN propose the new term be placed under the subclass
```

Results

Thus far, four terminologies used by ancillary systems have been incorporated into the MED. The laboratory terminology was provided by the CPMC's locally developed system. The electrocardiography terminology is that used by a commercial system (Marquette Electronics, Milwaukee, WI). ICD9-CM terminology, which is used by several hospital systems, was also included in the MED. We make use of an enhanced version, which includes synonyms and "index" terms with some reorganization to make it more appropriate for use in clinical applications.[32] Pharmacy terms were obtained from the Digimedics system (Mediware, Scotts Valley, California), which includes the American Hospital Formulary Service (AHFS) codes.[33]

Terminologies Modeled in the MED

Using the modeling approach described above, each of the four terminologies was found to include several general term groups; the amounts of interaction between and within groups were variable. The six groups in the laboratory terminology were procedures, tests, specimens, sensitivity panels, sensitivity tests, and results; they included six intergroup interactions (i.e., procedure has part test, procedure has specimen, test has specimen, test has result, sensitivity panel has part sensitivity test, sensitivity test has result) but no interaction among terms within the groups. Literal attributes included names, codes, units, and normal-range values. The electrocardiography terminology had a similar, although simpler, model, i.e., two groups (procedures and tests) and two intergroup relations (procedure has part test, test has result). The terms in four different ICD9 groups were diseases, procedures, health status factors, and external causes of injury; they were found not to interact with each other at all; however, each group had interactions among its terms in the form of hierarchical relationships. Literal attributes for ICD9

Table 2 ■

Addition of Clinical Terms to the Medical Entities Dictionary

	Laboratory	Electrocardiography	ICD9*	Pharmacy
No. of groups in terminology	6	2	4	5
Example	Specimens	12-lead ECG components	Procedure	Medication
No. of terms in all groups	1,488†	8	25,510	2,451‡
Example	PB lab serum	P-R interval	Spinal tap	Aspirin tablet
Interterm relations	1,577	7	25,508	6,019
Example	PB lab serum SAMPLES serum	P-R interval PART OF 12-lead ECG	Spinal tap SUBCLASS OF operative neurologic procedure	Aspirin tablet HAS COMPONENT aspirin
Literal attributes	26	1	4	22
Example	Specimen code "PB001"	Units	ICD9 Code "03.31"	Unit dose "325 MG"
No. of values	8,572	7	63,503	44,348
(bytes)	(91,915)	(44)	(1,518,439)	(234,405)

*International Classification of Diseases, Ninth Edition.
†724 of these terms were laboratory tests; 177 were laboratory specimens.
‡2,114 of these terms were medications.

terms included the name, the ICD9 codes (preferred and index), and synonyms. In the pharmacy terminology, the terms in one group (medications) could be classified by terms in each of the other four groups [AHFS classes, allergy classes, drug forms (tablet, liquid, etc.), and Drug Enforcement Administration (DEA) classes]. None of the other groups showed any interaction, and of the five groups, only the AHFS terms interacted with each other (as hierarchical relations). Literal attributes included information about the unit dose, manufacturer, generic and brand name, price, etc. Table 2 shows the sizes of the terminologies, with the numbers of terms in the term groups, the numbers of relationships between pairs of terms, the amounts of data represented as literal information, and examples of each.

Making Implicit Knowledge Explicit

Thus far, additional knowledge, in the form of semantic links to other MED concepts, has been added to three groups of terms: laboratory specimens, laboratory tests, and medications. Table 3 shows these groups ("Class 1"), the names of semantic links created to express this information, the total number of semantic links created to express the implicit knowledge, the class of terms used as attribute values for each link ("Class 2"), the number of terms added to the MED for each class, and the number derived from the UMLS. The 177 laboratory specimens were linked to 97 "sampleable entities" (including 82 anatomic substances and 15 external substances such as fomites, air, ice, food, instruments, and medica-

Table 3 ■

Implicit Knowledge Added to the Medical Entities Dictionary

	Laboratory	Laboratory	Pharmacy
Class 1	Specimens	Tests	Medications
Number	177	724	2,114
Semantic link between classes	Substance sampled	Substance measured	Pharmaceutical component
Links added	177	693	2,235
Class 2	Sampleable entities*	Measurable entities†	Chemical
Number	97	309	636
Terms in UMLS	77	238	635

*Includes the UMLS classes Anatomic Structure, Body Location or Region, Body Space or Junction, Body Substance, Food, Manufactured Object, and Natural Substance.
†Includes the UMLS classes Body Substance, Cell, Chemical, Organism, and Quantitative Concept.

```
LABORATORY TEST
. CELL TEST
. PHYSICAL PROPERTY TEST
. CHEMISTRY TEST
. . URINE CHEMISTRY TEST
. . SWEAT CHEMISTRY TEST
. . ANTIBODY TEST
. . . BLOOD TOXOPLASMA GONDII ANTIBODY TESTS
. . . . TOXOPLASMA GONDII ANTIBODY TITER MEASUREMENT
. . . . BLOOD TOXOPLASMA GONDII IGG ANTIBODY TESTS
. . . . . BLOOD TOXOPLASMA GONDII IGG ANTIBODY TEST
. . . . . BLOOD TOXOPLASMA GONDII IGG ANTIBODY MEASUREMENT
. . . . BLOOD TOXOPLASMA GONDII IGM ANTIBODY TESTS
. . . . . BLOOD TOXOPLASMA GONDII IGM ANTIBODY MEASUREMENT
. . . . . BLOOD TOXOPLASMA GONDII IGM ANTIBODIES MEASUREMENT
. . INTRAVASCULAR CHEMISTRY TEST
. . . INTRAVASCULAR CALCIUM TEST
. . . . PLASMA CALCIUM TESTS
. . . . . IONIZED CALCIUM MEASUREMENT
. . . . . PRESBYTERIAN PLASMA CALCIUM MEASUREMENT
. . . . WHOLE BLOOD CALCIUM TESTS
. . . . . STAT WHOLE BLOOD CALCIUM ION MEASUREMENT
. . . . . PRESBYTERIAN WHOLE BLOOD CALCIUM ION MEASUREMENT
. . . . . ALLEN WHOLE BLOOD CALCIUM ION MEASUREMENT
. . . . SERUM CALCIUM TESTS
. . . . . SERUM CALCIUM MEASUREMENT
. . . . . SPECIAL CHEMISTRY CALCIUM MEASUREMENT
. . . . . ALLEN SERUM CALCIUM MEASUREMENT
. . . SERUM CHEMISTRY TEST
. . . . SERUM CALCIUM TESTS *
. . . . SERUM ANTICONVULSANT TEST
. . . . . SERUM VALPROIC ACID MEASUREMENT
. . . . . SERUM PHENOBARBITAL MEASUREMENT
. . . . . SERUM PHENYTOIN TEST
. . . . . SERUM PHENYTOIN MEASUREMENT
. . . . . FREE PHENYTOIN LEVEL MEASUREMENT
. . . PLASMA CHEMISTRY TEST
. . . . PLASMA CALCIUM TESTS *
. . . WHOLE BLOOD CHEMISTRY TEST
. . . . WHOLE BLOOD CALCIUM TESTS *
. . . . BLOOD TOXOPLASMA GONDII ANTIBODY TESTS *
```

Figure 3 Results of automated classification of laboratory terms in the Medical Entities Dictionary (MED). Terms in **bold** correspond to actual laboratory test terms. Tests in the laboratory system, as parts of procedures, usually have names such as "CA" and "GLUC." The names shown here are the canonical names created to represent them in the MED. Terms (except for "Laboratory Tests") listed in normal font correspond to a few of the classes created during the automated classification process. A sample of specific tests is listed for each class when it first appears in the hierarchy. When a class has two or more positions in the hierarchy, the term appears with an asterisk (*) and the subordinate tests are omitted.

tions). Twenty of these terms could not be found in the UMLS, including some body surfaces (e.g, "rectal surface" and "lip lesion") and certain body fluids ("serum," "lochia," and "pancreatic cyst fluid"). The 724 laboratory tests were linked to 309 "measurable entities," which were, to a great extent, chemicals. Seventy-one terms could not be found in the UMLS, including 66 antibodies, three antigens, and two specific forms of alpha-amylase (pancreatic and salivary). The 2,114 medications from the pharmacy terminology were linked to chemicals. Only one of these (magnesium salicylate) could not be found in the 1991 version of the UMLS. Of note, that chemical has subsequently been added to the 1992 version.

Automated Classification of Terminologies

The subclassification algorithm was applied first to the laboratory specimen terms to group 170 of the 177 terms into 23 classes, based on their sampled substances (blood, serum, urine, etc.). The subclassification algorithm was then applied to the 724 laboratory tests, which were grouped into 251 classes based on substance measured and class of specimen (e.g., blood glucose tests, serum potassium tests. Figure 3 shows part of the resulting hierarchy for laboratory tests. Of the 2,112 medications, 2,037 were classified, based on their ingredients, into 636 "preparation" classes. The new preparation classes could then be placed in classes based on the classes of their children (the medications). For example, the new aspirin medications class was placed in the AHFS class for nonsteroidal anti-inflammatory agents and in the nonsteroidal anti-inflammatory agents' allergy class.

The automatic addition of preparation classes to parent classes provided increased functionality of the MED. Three examples illustrate the result of this process. First, there were 115 cases where medications included multiple pharmaceutic components. As a result, these medications were included in multiple preparation classes, placing them in multiple AHFS classes. For example, "Empirin with Codeine #3 Tab" was placed in the "Aspirin Preparations" and "Codeine Preparations" classes. Since these classes were under "Nonsteroidal Anti-inflammatory Agents" and "Opiate Agonists," respectively, the drug is now under both classes.

Second, the AHFS has a class called "Anticonvulsants" (AHFS code 281200), which has under it the class "Barbiturate Anticonvulsants" (AHFS code 281204). There is also a class "Anxiolytics, Sedatives, and Hypnotics" (AHFS code 282400), which has under it the class "Barbiturate Preparations" (AHFS code 282404). The drug "Phenobarbital Sodium 130 mg/ml" (AHFS code 282404) is classified under "Barbiturate Preparations" but not under "Barbiturate Anticonvulsant." Conversely, there is a drug "Unit Dose Phenobarbital 100 mg Tablet" (AHFS code 281204) that is classified as a "Barbiturate Anticonvulsant" but not as a "Barbiturate Preparation." Because it has subordinate terms assigned to the two AHFS codes, the "Phenobarbital Preparations" class was added to both AHFS classes. Figure 4 shows part of the resulting hierarchy for these medications.

Third, while most of the aspirin-containing medications were assigned to allergy class "03" ("Allergy Class: Salicylates, NSAIDs, and Pyrazoles"), some (due to an oversight in the pharmacy system) were assigned to allergy class "00" ("Allergy Class: None"). However, because all of these aspirin medications were placed in the aspirin preparation class, the en-

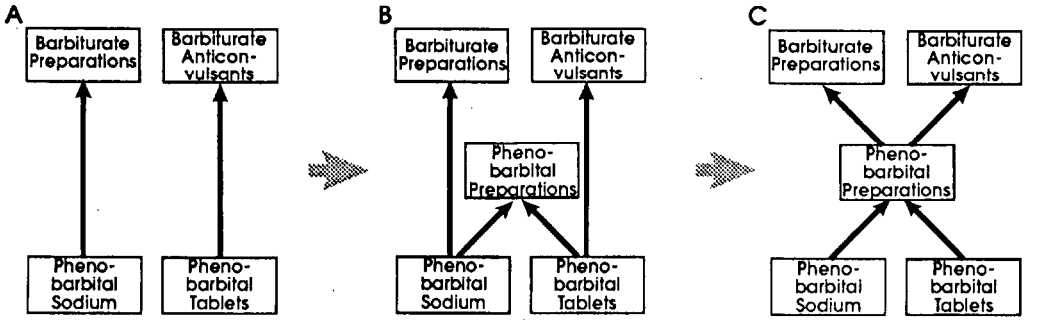

Figure 4 Using preparation classes to classify medications in multiple ways. In *A*, two phenobarbital medications are classified into single, different (although appropriate) AHFS classes. In *B*, a new class has been created to contain all phenobarbital medications. In *C*, the new class is added to all American Hospital Formulary Service classes that its children are in, and redundant links are removed. As a result, the two phenobarbital drugs are now both considered to be in the classes "Barbiturate Preparations" and "Barbiturate Anticonvulsant."

tire class, regardless of individual allergy code assignments, could be placed in "Allergy Class: Salicylates, NSAIDs, and Pyrazoles" (see Figure 5).

MED Maintenance

Since its creation, most of the terminologies added to the MED have required maintenance due to changes in the original terminologies. Only the electrocardiography terminology has required no change. ICD9 changes involved addition of 145 new terms and alteration of 71 term names to reflect changes in the 1992 ICD9 update. These changes were accomplished manually, using the MED Editor tools; however, automated knowledge-based approaches have been used to carry out changes to the laboratory and pharmacy terminologies.

Between the time of the first addition of laboratory terms and this writing, there have been two updates to the laboratory terminology, including two new specimen terms, eight new laboratory procedures, and 40 new tests. The new terms were added to the top-level class ("laboratory specimen," "laboratory procedure," or "laboratory test") and each was linked to the appropriate term for sampled or measured substance. One of us (JJC) spent half an hour manually identifying appropriate classes for the terms in advance. Subsequent performance of the autoclassification program (which was blinded to the manual class assignments) was evaluated. In each case, the new term was placed in the same class by both methods, with the total time required less than 5 minutes. For example, the laboratory personnel defined a new procedure called "Prolactin and/or Growth Hormone"; a concept of this new procedure was added to the laboratory procedures class in the MED. Its specimen "PB Lab Serum" was linked to the term Serum (as a sampled substance) and the system was able to add PB Lab Serum to the "Serum Specimens" class. The two tests in the procedure, "PROLAC" and

A

```
ALLERGY CLASS: SALICYLATES; NSAID; PYRAZOLES  (ALLERGY CODE 3)
   . UD ASCRIPTIN TAB  (ALLERGY CODE 3)
   . ASPIRIN 30 MG EC TAB  (ALLERGY CODE 3)
   . ASPIRIN E.C. 975 MG TAB  (ALLERGY CODE 3)
   . UD ASPIRIN EC 325 MG TAB  (ALLERGY CODE 3)
   . ASPIRIN 30 MG SUPP  (ALLERGY CODE 3)
   . ASPIRIN 60 MG SUPP  (ALLERGY CODE 3)


ALLERGY CLASS: NONE  (ALLERGY CODE 0)
   . ASPIRIN 30 MG TAB  (ALLERGY CODE 0)
   . ASPIRIN 75 MG TAB  (ALLERGY CODE 0)
   . UD ASPIRIN 325 MG TAB  (ALLERGY CODE 0)
   . ASCRIPTIN 325 MG TAB  (ALLERGY CODE 0)
   . ASPIRIN, SUSTAINED 650 MG TAB  (ALLERGY CODE 0)
   . ASPIRIN 120 MG SUPP  (ALLERGY CODE 0)
   . ASPIRIN 60 MG SUPP  (ALLERGY CODE 0)
```

B

```
ALLERGY CLASS: SALICYLATES; NSAID; PYRAZOLES  (ALLERGY CODE 3)
   . ASPIRIN PREPARATIONS
   .  . ASPIRIN 30 MG TAB  (ALLERGY CODE 0)
   .  . ASPIRIN 75 MG TAB  (ALLERGY CODE 0)
   .  . UD ASPIRIN 325 MG TAB  (ALLERGY CODE 0)
   .  . ASCRIPTIN 325 MG TAB  (ALLERGY CODE 0)
   .  . UD ASCRIPTIN TAB  (ALLERGY CODE 3)
   .  . ASPIRIN 30 MG EC TAB  (ALLERGY CODE 3)
   .  . ASPIRIN E.C. 975 MG TAB  (ALLERGY CODE 3)
   .  . UD ASPIRIN EC 325 MG TAB  (ALLERGY CODE 3)
   .  . ASPIRIN, SUSTAINED 650 MG TAB  (ALLERGY CODE 0)
   .  . ASPIRIN 120 MG SUPP  (ALLERGY CODE 0)
   .  . ASPIRIN 30 MG SUPP  (ALLERGY CODE 3)
   .  . ASPIRIN 60 MG SUPP  (ALLERGY CODE 0)
   .  . ASPIRIN 60 MG SUPP  (ALLERGY CODE 3)

ALLERGY CLASS: NONE  (ALLERGY CODE 0)
```

Figure 5 Allergy classification of aspirin medications in the Medical Entities Dictionary. Panel *A* shows the allergy classification of all aspirin-containing medications (excluding aspirin combination medications) based on the original formulary. Panel *B* shows the reclassification that was accomplished by placing all aspirin medications in the "aspirin preparation class" and then placing the new class in the "Salicylates: NSAIDs; Pyrazoles" allergy class. Assignment of medications to the aspirin preparation class was acomplished automatically, based on the generic ingredients of the medications. Assignment of the aspirin preparation class to the proper allergy class was also accomplished automatically by using the allergy codes of the class members.

"GH," were linked to the chemical terms Prolactin and Growth Hormone (as substances measured). This information and the information that the specimen was in the Serum Specimen class allowed the system to add the two new test terms to the "Serum Chemistry Tests" class. There was no previous test in the MED involving prolactin or growth hormone, so the system did not propose creating new classes for these tests.

For maintenance of pharmacy terms, a recent-update file that included 600 new medications was obtained from the pharmacy system. Using generic ingredient information in the medication records, the program correctly classified 310 (52%) of the new medications into preparation classes. The program correctly classified an additional 237 (39%) of the new medications into AHFS classes, using the AHFS codes. Classification of these 547 medication terms took approximately an hour. It is estimated that the same process, if attempted manually, would take over ten hours.

The program failed to classify the remaining 53 medications (9%) because the AHFS code was absent and the generic information was either new (45 medications) or absent (eight medications). For the 290 medications with new generic ingredients, a total of 154 "new" unique ingredients were found. Of these, 113 were truly new and 41 were misspellings of existing generic ingredients.

Discussion

Since the original proposal of our model in 1989, the MED has grown from a design specification to a working controlled terminology. A review of the knowledge-based approach is in order. This section addresses three issues: success in developing a terminology that meets the original design criteria, testing of the hypothesis that such a design would facilitate automated terminology construction and maintenance, and usefulness for a clinical information system.

Domain Completeness, Synonymy, Consistency of Views, Explicit Relationships, and Multiple Classification

No vocabulary can claim complete coverage of the domain of medicine; however, the MED provides complete coverage of its stated domain: terminologies of selected ancillary systems. Domain completeness is allowed in theory, since there is no inherent limitation on the size of the network with regard to number of nodes in the network as a whole, number of nodes in a class, depth of a node in the hierarchy, number of relations in the network, or number of relations involving any one node. In practice, the present MUMPS implementation has sufficient room for growth to include terminologies needed for at least several years.

Synonymy is present in a straightforward manner. The top node in the MED (Medical Entity) has the literal attribute "synonym"; every MED term inherits this attribute, which can be filled with alternate names. For example, the 25,510 ICD9 terms include 7,070 synonyms. Synonyms need not be unique. Thus, "MI" can be a synonym of "myocardial infarction" and "mitral insufficiency."

Several other original design criteria were met easily by our choice of representational scheme (a semantic network of frames with a superimposed directed acyclic graph for classification). Specifically, consistency of views and explicit relationships are inherent features of a semantic network and were therefore achieved, by definition, in the MED. The directed acyclic graph model permits multiple classification. Therefore, this criteria can be met by the MED, so long as editors are willing to provide the information. The knowledge-based approach offers a useful way to provide such information automatically.

Nonvagueness and Nonambiguity

The four ancillary terminologies that have received close attention benefited by a reduction in vagueness and ambiguity. For example, laboratory tests or medications that have vague names can be understood much more readily by examining the semantic links (added manually) and their classes (added automatically). The MED provided for an expansion of the definitions of laboratory tests, in effect making them less vague. For example, the "K" (potassium test) of a Chem-7 (a plasma chemistry panel) and the "K" of a Stat Panel (a whole blood chemistry panel) have the same code (CC000002) in the laboratory system, suggesting that the same test is a component of two panels. However, for technical reasons, the "K" in the Chem-7 has a different normal range than the "K" in a Stat Panel. Thus, the true meaning of "K" (or CC000002) is vague in the laboratory system. Inclusion of the specimen relationship clarifies the meaning.

A single test that seems to measure two different chemicals suggests ambiguity. Five examples of such ambiguity were found and communicated to the laboratory system personnel, who recoded them to disambiguate "single" tests into multiple, unambiguous tests. For example, a test with the code CC000009

appeared in one procedure ("Serum Inorganic Ions") with the name "MG." In a second procedure ("SMAC"), it appeared with the name "PHOS."

Complete satisfaction of these criteria would require the inclusion of medical knowledge for each term in the MED sufficient to determine that some minimum definition was present (nonvagueness) and that no definition appeared to contain information that might be construed as forming two separate definitions (nonambiguity). This is not the case in the present MED, given that the semantic information is present in only certain classes of terms. For example, adding semantic definitions of all of the ICD9 concepts is beyond the capabilities of our present resources (estimated to be six man-months if definitions take 5 minutes each).

Nonredundancy

It was originally believed that redundancy could be detected by comparing the set of semantic relations of a proposed term with the set of each existing term and, when an identical match was found, suggesting that redundancy was present. There are pragmatic reasons why such an approach is impractical. First, since the MED contains semantic information in only limited domains, detecting, for example, redundant ICD9 index terms would require a significant scaling up of the present vocabulary maintenance effort. Second, the presence of identical semantic descriptions in the MED turns out to be the method used to detect new classes of terms. For example, three tests in the laboratory system measure sodium ion concentration in whole blood. To say that these are redundant and should be represented as a single concept in the MED ignores nuances relevant to the laboratory and, perhaps, the clinician. For example, one of these tests is done as part of a stat panel and may be less reliable than those done by conventional methods. Ideally, this nuance could be captured with a terminology of analyzers and semantic links between tests and analyzers. This level of differentiation remains implicit in the MED at present.

Analysis of result terminologies from the laboratory system occasionally yielded minor coding problems. For example, the culture result "Flavobacterium breve" appeared twice, with codes "RC577" and "RC580." The detection of such problems was not due to our knowledge-based approach but due to the new scrutiny under which the terms were placed. We remain optimistic that automated classification will identify redundant terms when it tries to create new classes for terms with the same meaning. This hypothesis remains unproven at present.

Implications for Terminologic Work

The MED is a hybrid of terminology and knowledge. We believe that the full benefit of computerized medical information systems cannot be realized without such hybridization.[23] Others have advocated similar approaches for the design of computer-based controlled medical terminologies to increase their utility.[19,34,35] The MED extends that utility to include the task of terminology maintenance. The construction and maintenance of terminologies for current clinical systems is, for the most part, carried out by manual means with limited computer-based tools, while the MED model accommodates sophisticated techniques and tools for these tasks.

The availability of these tools, in turn, allows for closer scrutiny of terminologies. For example, the addition of laboratory terms to the MED led to the detection of spelling errors. As each test was added to the MED, an attempt was made to look up its measured substance in the UMLS, using the name of the test to obtain a lexical match. In several cases, the look-up failed, and when the appropriate substance was found in the UMLS through manual look-up, the reason for the failure of the lexical match was found to be a misspelling of the test name.

Such tools also provide for automated maintenance that would be otherwise impossible. For example, in the near future, it will be necessary to modify the MED to reflect daily changes in the pharmacy medication list. If this maintenance task requires close attention by a human editor, there will be delays and, through haste, mistakes (or at least inconsistencies). With an automated approach, such maintenance can proceed rapidly (with at least a tenfold improvement in update time) and consistently, using the automated classification techniques described here.

It can be argued that instead of manually linking concepts to one another to add "knowledge," the time could be better spent by simply working directly on tasks such as multiple classification. We find, however, that while linking a term to one or two other specifically related concepts can be somewhat tedious, it is relatively straightforward and much simpler than attempting to identify manually all the relevant classes for a term. Furthermore, we believe that the knowledge can be reused. For example, a program that displays summaries of laboratory results might retrieve test results based on the measured substances rather than on a list of specific tests.

The work at CPMC to merge controlled medical terminologies is not being done in isolation. Our efforts with local terminologies are supported by efforts at

the National Library of Medicine to merge international vocabularies in the UMLS.[36,37] Concepts in the MED are identified, wherever possible, in the UMLS Metathesaurus. This simplifies the task of accruing new concepts for knowledge representation. For example, the chemical terminology used for describing drug and laboratory concepts was drawn from the UMLS, rather than having to be reinvented. Use of the UMLS also facilitates some of the semantic linking needed in the MED. For example, the ability to use the MED to translate ICD9 to MeSH uses interconcept relations in the Metathesaurus.[38]

Despite the ability to build the MED on the UMLS, there are important differences between the two projects in their present forms. From a modeling point of view, each includes a semantic network with classes and relations and large sets of terms from disparate vocabularies. However, while the UMLS provides information about potential semantic relations between classes and provides some linkages between members of classes (as "Broader," "Narrower," and "Other" relations), the MED includes specific, named relations between specific terms. Both projects attempt to harmonize terms from disparate terminologies. However, the focus of UMLS methods has been on lexical approaches, while the focus of MED methods has been on knowledge-based approaches (although each project makes some use of alternate methods). In addition, the domain of the MED differs from that of the UMLS. By focusing on terminologies in use in CPMC ancillary systems, the MED covers a much smaller range of terms, yet it delves into areas where the UMLS currently provides little coverage.

Despite these differences, the MED is not intended to serve as an alternative to the UMLS. Instead, we view the two projects as complementary, The MED has clearly benefited from the availability of the UMLS.[31,38] Likewise, the UMLS has benefited from MED-related research.[32] As the UMLS content grows to include more clinical domains (such as those provided by the MED) and its model evolves to include more information about inter-term relationships (for example, links between ICD9 terms and other concepts), we envision that the mutual benefits between the two projects will increase.

Implications for Clinical Information Systems

The list of MED criteria was originally viewed as an experimental set of desired features. Yet, the MED is more than an experiment in sophisticated terminology creation and maintenance. It was created with the intent of supporting an advanced, centralized clinical information system. The design criteria are, in fact, as much involved with medical information processing requirements as they are with those of terminology maintenance. As the MED begins to reach critical mass (it currently contains 32,767 concepts), we are beginning to realize the benefits of the model. The MED has been (and continues to be) used to encode real patient data.[39] As of this writing, over six million procedures and 48 million test results have been coded for over 300,000 patients. The MED provides a single, stable coding system for representing disparate patient data from a variety of ancillary systems with changing terminologies. This capability, together with the "value added" knowledge in the MED, has a number of implications for clinical computing at CPMC.

First, the MED provides a central resource for identifying the current terms in ancillary systems. Such information is usually not generally available in a timely way to those who need it.

Second, maintaining the interrelationships between classes in the MED simplifies the clinical database design.[39] Instead of having to represent complex semantics of medical data and change the design to accommodate new information from the ancillary system, it is only necessary to represent that semantics are present and allow the MED to keep track of the specific semantics.

Third, the developers of clinical applications on the central system are relieved of the responsibility of keeping their programs synchronized with changes in ancillary systems. For example, a program to provide summary reports of laboratory tests of, say, blood glucoses need not be updated when a new blood glucose test is added to the laboratory repertoire. Instead, the summary reporter can refer to the MED class "blood glucose tests" and retrieve all clinical information coded with MED Codes of terms in that class. The knowledge-based MED maintenance facilitates the inclusion of the new test into the proper MED class.

Fourth, retrievals performed on the patient database for quality assurance or clinical research can benefit by issuing queries based on MED coding, rather than that provided by an ancillary system. Returning to the example at the beginning of the paper, the disease "staphylococcal pneumonia" would, ideally, be added to the MED with the information that it was caused by Staphylococcus and that it occurs in the lung. If this information were sufficient, the disease would be classified in the MED both as a pneumonia and as a staphylococcal disease. Thus, a pulmonologist studying lung diseases could identify patients coded

as having staphylococcal pneumonia as easily as could the epidemiologist studying pathogens.

Fifth, clinical decision support systems[16] can take advantage of the MED in several ways. One clear benefit is that maintenance of the system logic, just like other clinical applications, is independent of minor changes in the ancillary systems providing the clinical data on which the decision support system operates. More important, however, the MED provides medical knowledge that the ancillary systems do not include. For example, if the pharmacist enters the fact that a patient has an aspirin allergy and then attempts to order the medication "Ascriptin 325mg Tab" (see Figure 5), the pharmacy system will not raise any objection, since it believes that medication to be nonallergenic (allergy code "00"). However, the MED provides information that allows the decision support system to recognize that this medication is, in fact, in an allergy class relevant to the patient.

Sixth, as the ancillary terminologies evolve, the knowledge now in the MED will facilitate its own maintenance. For example, it is likely that new brands of existing medications will be added to the pharmacy system much more frequently than will entirely new medications. The knowledge in the MED allows the former to be automated and identifies the latter for manual processing and review. As new ancillary system terminologies are added to the MED, we anticipate that there will be significant overlap between the new terms and those already in the MED. This overlap will require merging of current and future terminologies into a unifying classification and perhaps require merging of redundant concepts. We believe that the knowledge in the MED can be used to facilitate this process.

The ancillary clinical systems at CPMC are not unique to our environment. Many other institutions have installations of the same commercial products or have locally built systems with similar characteristics. Likewise, the central pooling of clinical data into a central database for multiple purposes is a common scheme. What we believe is unique to the CPMC environment is the use of knowledge to assist in the coordination of the terminology tasks that are needed to accomplish the centralized scheme. It is therefore conceivable that other institutions could adopt the MED approach with the same potential for benefit. While the work required to build a MED from scratch is significant, a standard for the knowledge representation could be developed that would allow sharing of terminologic work across institutions[40] in the same way that the Arden Syntax facilitates the sharing of medical logic.[16]

Conclusion

The CPMC Medical Entities Dictionary is a complex data structure that has required extensive effort to maintain. Fortunately, the work being invested in knowledge representation, for use by other systems, is proving to be of benefit to the terminology maintenance system itself, through the development of sophisticated terminology editing tools. As a result, we have prevented ambiguity, misclassification, and redundancy that would otherwise have occurred.

The authors thank Dr. Robert Sideli, Dr. Soumitra Sengupta, Dr. Carol Friedman, and Dr. Daniel Fink for helpful discussions throughout the project. Thanks are also due to Ms. Leslie Juccam for assistance with editing the manuscript.

References ■

1. International Standards Organization. International Standard ISO 1087: Terminology—Vocabulary. Geneva, Switzerland: International Standards Organization, 1990.
2. American Society for Testing and Materials (ASTM). Preparation and use of terminology in ASTM standards. In: Form and Style for ASTM Standards. Eighth edition. Philadelphia: ASTM, 1989;27–34.
3. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD. Designing an introspective, controlled medical vocabulary. In: Kingsland LW, ed. Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care. Washington, DC: IEEE Computer Society Press, 1989:513–8.
4. United States National Center for Health Statistics. International Classification of Diseases, Ninth Revision, with Clinical Modifications. Washington, DC: U.S. National Center for Health Statistics, 1980.
5. Côté RA, ed. Systematized Nomenclature of Medicine. Second edition. Skokie, IL: *College of American Pathologists*, 1982.
6. National Library of Medicine. Medical Subject Headings (NTIS NLM-MED-92-01). Bethesda, MD: NLM, 1992.
7. Gordon BL, ed. Current Medical Information and Terminology. Fourth edition. Chicago: American Medical Association, 1971.
8. *Lindberg DAB, Humphreys BL.* Toward a unified medical language system. Medical Informatics Europe '89; Proceedings of the Seventh International Congress. Berlin, Germany: Springer-Verlag, 1987;1:23–31.
9. Lindberg DAB, Humphreys BL. Computer systems that understand medical meaning. In: Scherrer JR, Côté RA, Mandil SH, eds. Computerized Natural Medical Language Processing for Knowledge Representation. Amsterdam, The Netherlands: Elsevier Science Publishers, 1989;5–17.
10. Woods W: What's in a link: foundations for semantic networks. In: Bobrow DG, Collins A, eds. Representation and Understanding: Studies in Cognitive Science. New York: Academic Press, 1975:35–82.
11. Komorowski HJ, Greenes RA, Pattison-Gordon E. The use of fisheye views for displaying semantic relations in a medical taxonomy. In: Stead WW, ed. Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care. Los Angeles: IEEE Computer Society Press, 1987:113–6.

12. Masarie FE Jr, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24:379–400.
13. Hendrickson G, Anderson RK, Clayton PD, et al. The integrated academic information management system at Columbia–Presbyterian Medical Center. MD Comput. 1992;9:35–42.
14. Clayton PD, Sideli RV, Sengupta S. Open architecture and integrated information at Columbia–Presbyterian Medical Center. MD Comput. 1992;9:297–303.
15. Johnson SB, Friedman C, Cimino JJ, Clark T, Hripcsak G, Clayton PD. Conceptual data model for a central patient database. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1992:381–5.
16. Hripcsak G, Cimino JJ, Johnson SB, Clayton PD. The Columbia–Presbyterian Medical Center decision-support system as a model for implementing the Arden Syntax. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1992:248–52.
17. Cimino JJ, Aguirre A, Johnson SB, Peng P. Generic queries for meeting clinical information needs. Bull Med Library Assoc. 1993;81:195–206.
18. Cimino JJ, Hripcsak G, Johnson SB, Friedman C, Clayton PD. Prototyping a vocabulary management system in an object-oriented environment. In: Timmers T, Blum BI, eds. Proceedings of the International Medical Informatics Association Working Conference on Software Engineering in Medical Informatics. Amsterdam, The Netherlands: North-Holland Press, 1991:429–39.
19. Nowlan WA, Rector AL, Kay S, Horan B, Wilson A. A patient care workstation based on user centered design and a formal theory of medical terminology: PEN&PAD and the SMK formalism. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1992:855–7.
20. Humphrey SM, Miller NE. Knowledge-based indexing of the medical literature: the Indexing Aid Project. J Am Soc Info Sci. 1987;38:184–96.
21. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. MD Comput. 1990;7:104–9.
22. Huff SM, Warner HR. A comparison of Meta-1 and HELP terms: implications for clinical data. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care. Los Alamitos, CA: IEEE Computer Society Press, 1990:166–9.
23. Cimino JJ. Saying what you mean and meaning what you say: coupling biomedical terminology and knowledge. Acad Med. 1993;68:257–60.
24. Barr CE, Komorowski HJ, Pattison-Gordon E, Greenes RA. Conceptual modeling for the Unified Medical Language System. In: Greenes RA, ed. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. Silver Spring, MD: IEEE Computer Society Press, 1988:148–51.
25. Packer MS, Cimino JJ, Barnett GO, Kim R, Hoffer EP, Zatz S. Updating the DXplain database. In: Greenes BA, ed. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. Silver Spring, MD: IEEE Computer Society Press, 1988:96–100.
26. Gabrieli ER. Interface problems between medicine and computers. In: Cohen GS, ed. Proceedings of the Eighth Annual Symposium on Computer Applications in Medical Care. Washington, DC: IEEE Computer Society Press, 1984:93–5.
27. Held JP, Chute CG, Gatewood LC, Connelly D. A schematic analysis of medical classification systems (abstr). In: Mitchel JA, ed. Proceedings of the First Annual Education and Research Conference of the American Medical Informatics Association. Bethesda, MD: American Medical Informatics Association, 1990:62.
28. Wingert F. Reduction of redundancy in a categorized nomenclature. In: Côté RA, Protti DJ, Scherrer JR. eds. Role of Informatics in Health Data Coding and Classification Systems. Amsterdam, The Netherlands: Elsevier, 1985:191–202.
29. National Library of Medicine. UMLS Knowledge Sources–Second Experimental Edition. Bethesda, MD: NLM, 1991.
30. Cimino JJ, Barnett GO. The physician's workstation: recording a physical examination using a controlled vocabulary. In: Stead WW, ed. Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care. Los Angeles: IEEE Computer Society Press, 1987:287–91.
31. Cimino JJ, Hripcsak G, Johnson SB, Friedman C, Fink DJ, Clayton PD. UMLS as knowledge base—a rule-based expert system approach to controlled medical vocabulary management. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care. Los Alamitos, CA: IEEE Computer Society Press, 1990:175–79.
32. Cimino JJ, Barrows RC, Allen B. Adapting ICD9-CM for clinical decision support (abstr). In: Musen MA, ed. Proceedings of the 1992 Spring Congress of the American Medical Informatics Association. Bethesda, MD: American Medical Informatics Association, 1992:34.
33. American Society of Hospital Pharmacists. American Hospital Formulatory Service Drug Information. Bethesda, MD: ASHP, 1988.
34. Rector AL, Nowlan WA, Kay S. Conceptual knowledge: the core of medical information systems. In: Lun KC, Degoulet P, Piemme TE, Rienhoff O, eds. MEDINFO 92. Amsterdam, The Netherlands: North-Holland, 1992:1420–6.
35. Rossi-Mori A. CEN/TC251/PT003 model for representation of terminologies and coding systems in medicine. In: Proceedings of the Seminar "Opportunities for European and U.S. Cooperation in Standardization in Health Care Informatics." Geneva, Switzerland, 1992.
36. Tuttle MS, Blois MS, Erlbaum MS, Nelson SJ, Sherertz DD. Toward a bio-medical thesaurus: building the foundation of the UMLS. In: Greenes RA, ed. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care. Silver Spring, MD: IEEE Computer Society Press, 1988;191–5.
37. Humphreys BL, Lindberg DAB. Building the unified medical language system. In: Kingsland LW, ed. Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care. Washington, DC: IEEE Computer Society Press, 1989:475–80.
38. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The MEDLINE button. In: Frisse ME, ed. Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1992:81–5.
39. Johnson SB, Cimino JJ, Friedman C, Hripcsak G, Clayton PD. Using metadata to integrate medical knowledge in a clinical information system. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care. Los Alamitos, CA: IEEE Computer Society Press, 1990:340–4.
40. Evans DA, Chute CG, Cimino JJ, et al. CANON: towards a medical-concept representation language for electronic medical records (abstr). In Kahn MG, ed. Proceedings of the 1993 Spring Congress of the American Medical Informatics Association. Bethesda, MD: American Medical Informatics Association, 1993:26.