UMLS as Knowledge Base - A Rule-Based Expert System Approach to Controlled Medical Vocabulary Management

James J. Cimino, M.D.; George Hripcsak, M.D.; Stephen B. Johnson, Ph.D.;
Carol Friedman, Ph.D.; Daniel J. Fink, M.D. and Paul D. Clayton, Ph.D.

Center for Medical Informatics - Columbia University
New York, New York 10032

The National Library of Medicine is developing a Unified Medical Language System (UMLS) which addresses the need for integration of several large, nationally accepted vocabularies. This is important to the clinical information system under development at the Columbia-Presbyterian Medical Center (CPMC). We are using UMLS components as the core of our effort to integrate existing local CPMC vocabularies which are not among the source vocabularies of the UMLS. We are also using the UMLS to build a knowledge base of vocabulary structure and content such that logical rules can be developed to assist in the management of our integrated vocabularies. At present, the UMLS Semantic Network is used to organize terms which describe laboratory procedures. We have developed a set of rules for identifying undesirable conditions in our vocabulary. We have applied these rules to 526 laboratory test terms and have found ten cases (2%) of definite redundancy and sixty-eight cases (13%) of potential redundancy. The rules have also been used to organize the terminology in new ways that facilitate its management. Using the UMLS model as a vocabulary knowledge base allows us to apply an expert system approach to vocabulary integration and management.

Introduction

A crucial component of medical information management is a controlled medical vocabulary which represents information in a way that enables computers to use it meaningfully (for example, to generate clinical alerts), rather than simply to manipulate it as text (e.g., storing and printing reports). There are currently many "standard" vocabularies from which to choose, but none is universally accepted by medical application designers and users. The National Library of Medicine is addressing this problem by bringing existing standards together so that translation can be achieved among the standards and the deficiencies of one can be rectified by another. The result of this effort is the Unified Medical Language System (UMLS) [1-4]. The UMLS includes a metathesaurus (Meta-1), with some 98,000 concepts drawn from national vocabularies (MeSH, ICD9-CM, SNOMED, DSM-III, CPT4, COSTAR, and LCSH) [5], and a Semantic Network containing 133 semantic types used to classify Meta-1 concepts [6].

The clinical information system now under development at the Columbia-Presbyterian Medical Center (CPMC) will include a number of applications, many of which already exist as stand-alone programs. The existing applications each make use of their own controlled vocabulary, which is either a national one (such as ICD9) or a local one created specifically for the application (such as that used by the laboratory information system). Integration of these applications in the CPMC information system will require, among other tasks, the integration of their controlled vocabularies. The UMLS provides a solution to the problem of integrating medical applications which use disparate controlled vocabularies, so long as those vocabularies are among the UMLS source vocabularies. The integration into a comprehensive medical center information system of applications which use their own local vocabularies requires some method for translating between these local terms and the national vocabularies used by other applications. It is unlikely that the designers of existing applications will adopt the UMLS for their systems. Our goal is to facilitate information exchange between applications by integrating local CPMC terminologies with the UMLS, rather than with each other. The result will be a composite terminology for intervocabulary translation. We desire this vocabulary to have several properties: domain completeness, synonymy, nonredundancy, nonambiguous and precise definitions, multiple classification, consistent views of terms and explicit relationships among terms [7].

We are applying the resources of the UMLS to create a model for vocabulary representation which supports our requirements for vocabulary integration. The UMLS Semantic Network provides the organization and structure of the local terms. The UMLS Meta-1 forms our core terminology, allowing us to integrate applications which use UMLS source vocabularies. We provide two additional features to complete the model: semantic definitions of terms and a rule base for vocabulary maintenance. This paper describes our knowledge-based approach and the initial results of rule-based maintenance of the composite UMLS-CPMC vocabulary.

Methods

Integration of Laboratory Terms with the UMLS

We have developed our vocabulary management system using the Knowledge Engineering Environment (KEE, Intellicorp, Mountain View California) and used it to represent the preliminary version of the UMLS Semantic Network, distributed in July of 1990.

We selected the terminology used by our laboratory information system as the first local vocabulary to be integrated with the UMLS because: 1) it is a good example of a local vocabulary which is well-established and successful; 2) it is used to represent important coded clinical information within our institution; 3) it is used by the clinical decision support system now under development [8]; and 4) there is overlap between our own laboratory

This work was supported in part by Integrated Academic Information Management Systems (IAIMS) Grant LM04419 from the National Library of Medicine and by the IBM Corporation.

0195-4210/90/0000/0175$01.00 © 1990 SCAMC, Inc.

terms and those in the UMLS Meta-1, offering the opportunity to integrate local terms with national ones.

The terminology taken from the laboratory information system included 524 "procedures", 526 "tests" (the individual parts of a procedure) and 168 "specimens". Integration of the laboratory terms consisted of adding each term to the KEE knowledge base as a descendant of the appropriate class in the UMLS Semantic Network.

Creation of Semantic Definitions

Once the terms of the CPMC laboratory vocabulary were added to the knowledge base, semantic definitions were created for each new term. A semantic definition consists of two sets of information: The first is a list of attributes, called link attributes, which have controlled terms as their values. Each attribute corresponds to a semantic relationship between the term being defined and other terms in the controlled vocabulary. For example, the semantic definitions for our laboratory tests represent the fact that they "measure" a particular substance in a "specimen" of a particular body part. The substance measured and the body part sampled are also represented as terms in the knowledge base. This representational scheme has been described previously in detail [7]. The links between the laboratory tests and other related terms form a semantic network [9] which, together with the UMLS Semantic Network, forms the CPMC Semantic Network. Specimen terms needed to construct laboratory test definitions are supplied by the laboratory information system. Our laboratory system does not include measured substance terms; these were taken from Meta-1.

The second set of information in the semantic definition is a list of attributes, called non-link attributes, which have values that do not correspond to controlled terms. For example, each laboratory term has a code and a name taken from the laboratory information system. Such information is represented as literal data (strings and numbers).

Rules-Based Vocabulary Maintenance

We have developed a set of vocabulary maintenance rules which express conditions which we find acceptable and unacceptable in a controlled vocabulary. Each rule makes use of some knowledge in the semantic definitions to detect situations which appear unacceptable. For example, two terms with identical semantic definitions are, or at least have the appearance of being, redundant. We can express this undesirable condition as a rule, with the form:

IF	Each semantic relationship of Term 1 EQUALS Each semantic relationship of Term 2
THEN	Terms 1 and 2 appear redundant.

Rules were created to address a number of specific conditions in our vocabulary: 1) the presence of pairs of terms with apparently identical meanings (redundancy rules); 2) the presence of single terms with multiple apparent meanings (ambiguity rules); 3) the presence of terms with incomplete meanings (vagueness rules); and 4) likely locations in the semantic network in which terms could be logically organized into semantic classes (classification rules). Rules were written in the KEE rule language, forming a rule base for vocabulary maintenance. This paper describes the results of applying three rules to the CPMC Semantic Network. The "Redundant Semantic Relations Rule" embodies the logic in the above example. The "Non-Unique Rule" looks for two terms with the same value an for attribute which should be unique. For

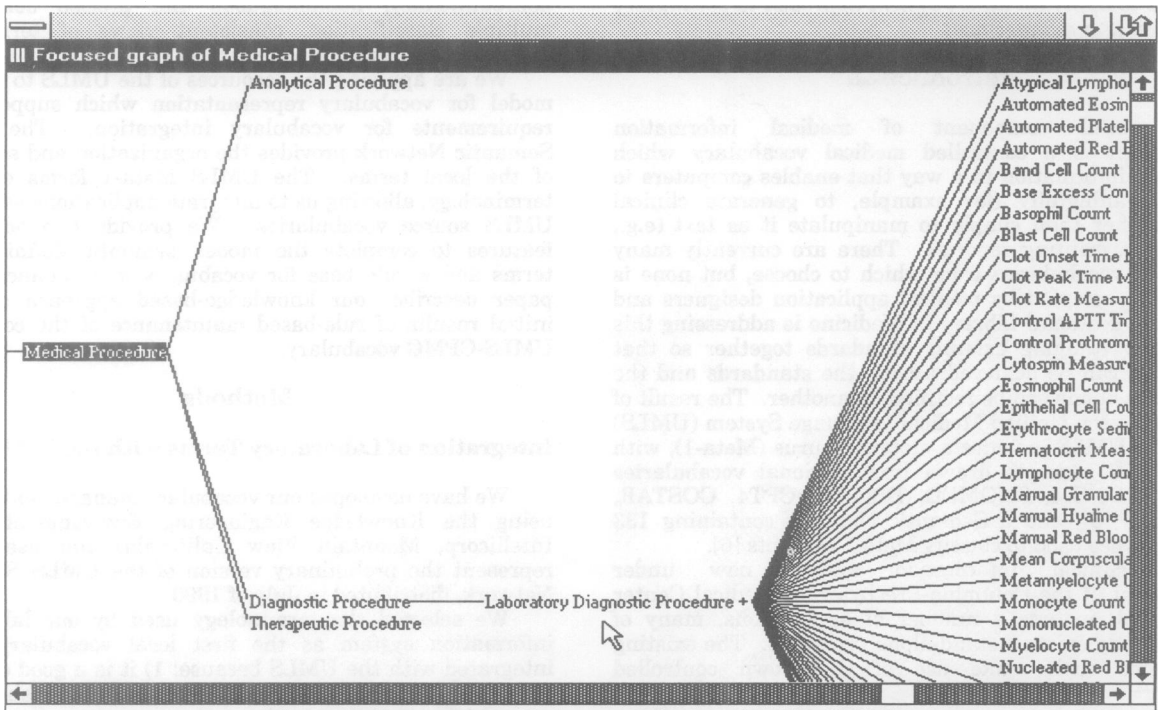

Figure 1 - Integration of Laboratory Terms with The UMLS
Graph showing partial view of the semantic network under "Medical Procedure" ("+" indicates omitted nodes). Note the large number of descendants under the term "Laboratory Diagnostic Procedure" (arrow).

example, no terms should have the same laboratory code. The Non-Unique Rule identifies pairs of terms with the same code.

The "New Class Rule" is used to identify terms which might constitute a new semantic class. This rule identifies some semantic attribute that is shared by a set of terms. It then takes the value of that attribute for one of the terms and identifies some semantic class of which that value is a member. Armed with this class, it determines which of the original set of terms have a value (for the selected attribute) in the same class and proposes a new semantic class. An example of application of the New Class Rule is described below, taken from an actual application of the classification rule to the semantic network.

Results

Integration of Laboratory Terms with the UMLS

The preliminary version of the UMLS Semantic Network contains 133 semantic classes. Our vocabulary knowledge base was established by representing these classes in KEE. We added a new semantic class, "Laboratory Diagnostic Procedure", as a subclass of the UMLS class "Diagnostic Procedure". This was done to allow the introduction to the network of semantic relationships that might not be appropriate for all diagnostic procedures (e.g., "Specimen" and "Measures"). The 1050 procedure and test terms were then added to the knowledge base as descendants of "Laboratory Diagnostic Procedure" (see Figure 1). The 168 specimen terms were added to the knowledge base as descendants of other, appropriate UMLS semantic classes (most are anatomic structures such as "plasma" and "biopsy tissue", but some are objects, such as "IV catheter" and "instrument").

LIPID.PROFILE
Superclasses: CLINICAL.CHEMISTRY.PROCEDURE

HAS-PARTS from LIPID.PROFILE
Has value: SERUM.TRIGLYCERIDE.CONCENTRATION.MEASUREMENT,
QUALITATIVE.SERUM.LIPEMIA.MEASUREMENT,
QUALITIATIVE.SERUM.CHYLOUS.MEASUREMENT,
SERUM.CHOLESTEROL.CONCENTRATION.MEASUREMENT

SPECIMEN from LIPID.PROFILE
Has value: SERUM

CPMC-LAB-PROC-CODE from LIPID.PROFILE
Has value: "CC000012"

CPMC-LAB-PROC-NAME from LIPID.PROFILE
Has value: "LIPID PROFILE"

SERVICE-CODE from LIPID.PROFILE
Has value: "06210654"

Figure 2 - Semantic Definition of "Lipid Profile Test"

Creation of Semantic Definitions

Semantic definitions for the laboratory procedures were created automatically, using information from the laboratory information system. The link attributes were "Has-Parts" and "Specimen"; these were assigned values corresponding to laboratory test and specimen terms respectively. The non-link attributes were "CPMC-Lab-Proc-Code", "CPMC-Lab-Proc-Name" and "Service-Code"; these received literal (character string) data directly from the laboratory system data dictionary. Figure 2 shows the semantic definition for the laboratory procedure "Lipid Profile". This definition links the procedure is linked to four terms corresponding to its component tests and a fifth term corresponding to its specimen.

Semantic definitions were created for the laboratory tests through a combination of automatic and manual means. The "Specimen" link attribute and the two non-link attributes ("CPMC-Lab-Test-Code" and "CPMC-Lab-Test-Name"), were provided by the laboratory information system. An additional link attribute, "Measures" was needed to represent the meaning of each test. Values for the "Measures" attribute for each test were obtained, where possible, from Meta-1. The Meta-1 terms, drawn from UMLS semantic classes such as "Chemical", "Cell", and "Bacterium", were added to the knowledge base. For example, the semantic definition for "Serum Cholesterol Concentration Measurement" includes the attributes "Specimen" and "Measures" with the values "Serum" and "Cholesterol", respectively.

Rules-Based Vocabulary Maintenance

The application of the rule base to the knowledge base produced many cases in which rule premises were satisfied, resulting in conclusions being made regarding undesirable conditions in the vocabulary. One stipulation of our vocabulary is that no two terms may have the same value for the "CPMC-Lab-Test-Code" attribute. The Non-Unique Rule was applied with respect to this attribute and 10 of the 526 test terms (2%) were found to have duplicate codes.

The stipulation that no two terms may have the same link attributes is expressed in the Redundant Semantic Relations rule. When this rule was applied to the test terms, 266 cases were found in which a test measured the same substance in the same specimen as some other test. This condition occurred for a variety of reasons. In 198 (74%) of the cases, the "redundant" test terms represented tests which are distinguished in the laboratory system by the location in which they are performed. Since this information was not included in the semantic definitions, the rule was unable to detect any differences between the terms. In the remaining 68 cases, the reason for the apparent redundancy is unclear. In some cases, the specimen given by the laboratory system may, in fact, be incorrect; inserting the correct specimens into the semantic definitions will abolish the apparent redundancy. In other cases, the tests may actually differ based on the analytic method used. Like the test location, this information is not included in the semantic definitions; however, it does not appear in the laboratory system either.

The New Class Rule was used to explore ways in which the terms in the semantic class "Laboratory Diagnostic Procedure" could be organized. Referring back to Figure 1, it can be seen that this class initially included many terms (i.e., 1050 procedures and tests). The rule was applied to this class and the following sequence of events occurred. First, "Serum Glucose Concentration Measurement" was selected as an arbitrary member of the class. Next, "Measures" was identified as a link attribute of that term and was found to have the value "Glucose". Then, "Chemical" was recognized as a semantic class for "Glucose". At this point, the members of the class "Laboratory Diagnostic Procedure" were divided into those which had a "Chemical" value for their "Measures" attribute and those which did not. Finally, the rule requested an acknowledgement (from the vocabulary expert who was applying the rule) that this was a valid criteria for subclassifying "Laboratory Diagnostic Procedure". Upon confirmation by the expert, a new class was created in the knowledge base to correspond to the set of terms with the proper value. This rule was applied repeatedly, until no new classes could be proposed. For each proposed class, the expert decided whether or not it should be included in the

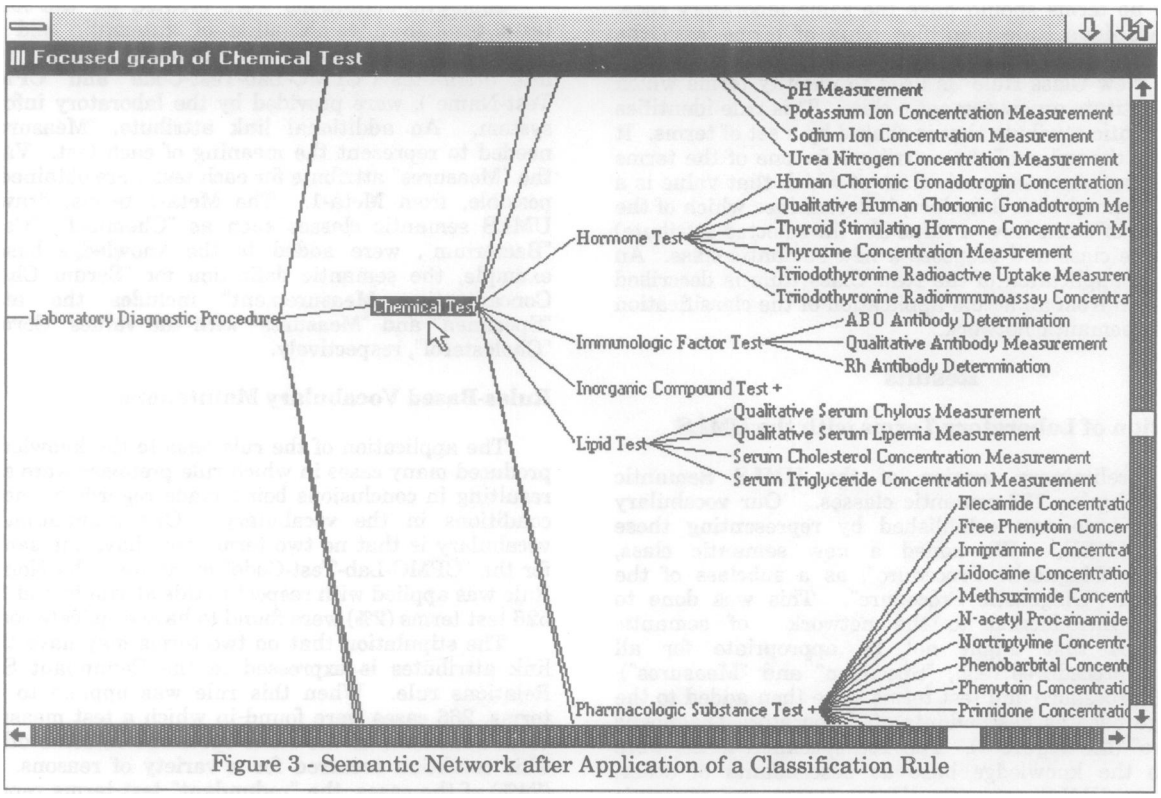

Figure 3 - Semantic Network after Application of a Classification Rule

knowledge base. The system identified and separated complex laboratory procedures (such as the complete blood count) from the individual component tests (such as hematocrit). It also subclassified various chemical tests according to the classes of chemicals measured. Figure 3 shows the resulting organization of the "Laboratory Diagnostic Procedure" class after the repeated application of the New Class Rule.

Discussion

The design of the CPMC clinical information system requires the development of a controlled medical vocabulary that is able to encompass the terminologies of all existing component systems. We have expressed our vocabulary requirements as formal rules; the representational scheme of our vocabulary provides the semantic information needed by the rules. The results obtained by applying the rule base to a network of semantic definitions builds upon previous work which has shown that the knowledge-based approach is a viable one for management of both local vocabularies [10] and the UMLS [11-14]. Through the use of our semantic model (that is, based on term meaning), we have been able to augment the traditional vocabulary maintenance methods which are based on lexical (e.g., word matching) and hierarchical (e.g., tree-walking) techniques.

The work presented here demonstrates the success of our approach. The ten tests with nonunique codes are clearly in violation of the vocabulary requirements of the laboratory information system. The sixty-eight additional cases of apparent redundancy may also represent such violations; however, the present semantic model is inadequate to make this judgement. A number of possible actions can be taken to resolve these apparent redundancies. One response could be to make the two terms synonymous; however, this would mask distinctions that could be clinically relevant. For example, the reliability of a test result will vary with the method used to acquire it. A second response would be to add new attributes such as "location" and "method", which allow for further distinctions to be represented. A current challenge is to decide when such additions are appropriate (e.g., clinically relevant). A third response is to create a new class and make the "redundant" terms its members. This permits the retention of the original terms and allows applications such as the clinical decision support system to refer to them as a generic way.

It is clear from these results that the modeling effort will be an iterative one, requiring additional effort to formalize features of a controlled vocabulary that might otherwise be left unstated. For example, the creation of the semantic definitions requires the establishment of semantic relationships between terms and the addition of new terms to the vocabulary. In some cases, this information is available (as in the case of our laboratory system), but in other cases domain experts will be required to provide the missing information. Some assistance may be available in the form of semantic relationships which can be generated automatically from the medical literature [15].

The use of the UMLS Semantic Network as a basis for our vocabulary is important for three reasons. First, the UMLS Semantic Network plays an important role in the knowledge base. For example, the reorganization of "Laboratory Diagnostic Procedure" was possible only through the use of the UMLS Semantic Network classification of chemicals. Second, the integration of terms from the UMLS Meta-1 with the local CPMC vocabularies will be more easily accomplished, since each Meta-1 term will already be a member of one or more UMLS semantic types [4]. Thus, assuming that our semantic definitions are

generated in a consistent manner, UMLS terms and CPMC terms which are potentially synonymous or otherwise related will co-occur in the same semantic classes, facilitating their comparison. Third, the use of Meta-1 concepts in our vocabulary will save much of the work required for local vocabulary integration since some of these "local" vocabularies are actually drawn from UMLS source vocabularies (although laboratory terms are not). In creating our semantic definitions, we have been careful to select terms which appear in Meta-1. Fourth, by representing our terminology with the UMLS, we can promote the sharing of medical information between our system and other systems which adopt the UMLS.

One of the requirements of users of the UMLS is that they agree to evaluate the UMLS resources and suggest necessary changes and additions. By integrating local vocabularies with the UMLS and by creating semantic definitions, we expect to provide extensive feedback on the values and limitations of the UMLS offerings. For example, we note that the fact that "Pharmacologic Diagnostic Procedure" could not be subdivided beyond its 32 children (Figure 3) reflects the fact that the UMLS Semantic Network does not subdivide chemicals into semantic classes by their pharmacologic function (antiarrhythmic agents, anticonvulsants, antibiotics, etc.). This suggests new types for the semantic network and new concepts for the next version of the UMLS Metathesaurus.

The laboratory system terminology is relatively small and stable, and is managed by a single expert; even so, it has significant inconsistencies. We can anticipate that larger, less stable vocabularies, coming from multiple sources and experts (such as those needed to express clinical findings) will be even more susceptible to such problems. Furthermore, with large vocabularies, the manual detection of irregularities will be much more difficult. Automated methods, such as we present here, will be required to address the task of vocabulary maintenance.

Conclusion

The results obtained support the notion that vocabulary integration and management will be facilitated using a knowledge base consisting of a semantic network of terms and a set of maintenance rules. By using the UMLS Semantic Network and Meta-1 as components of the CPMC Semantic Network, we simplify our task, obtain the additional value of the UMLS (in the form of links with UMLS source vocabularies) and make use of a vocabulary representation which is in accord with national standards.

References

1. Lindberg DAB, Humphreys BL: Toward a Unified Medical Language System. Medical Informatics Europe '89; *Proceedings of the Seventh International Congress*, Rome, September 11-15, 1987;1:23-31.

2. 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: Elsevier, 1989;5-17.

3. 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 SCAMC*; Washington, DC; Nov, 1988:191-5.

4. Humphreys BL, Lindberg DAB: Building the Unified Medical Language System. In Kingsland LW, ed.: *Proceedings of the Thirteenth SCAMC*; Washington, D.C.; November, 1989: 475-480.

5. Tuttle MS, Sherertz D, Erlbaum M, Olson N, Nelson SJ: Implementing Meta-1: the First Version of the UMLS Metathesaurus. In Kingsland LW, ed.: *Proceedings of the Thirteenth SCAMC*; Washington, D.C.; November, 1989:483-487

6. McCray AT: The UMLS Semantic Network. In Kingsland LW, ed.: *Proceedings of the Thirteenth SCAMC*; Washington, D.C.; November, 1989: 503-507.

7. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD: Designing an Introspective, Controlled Medical Vocabulary, in Kingsland LW, ed.: *Proceedings of the Thirteenth SCAMC*; Washington, D.C.; November, 1989: 513-518.

8. Hripcsak G, Clayton PD, Cimino JJ, Johnson SB, Friedman C: Medical Decision Support at Columbia-Presbyterian Medical Center. In Timmers T, ed.: *Proceedings of the International Medical Informatics Association Working Conference on Software Engineering in Medical Informatics*, Amsterdam, October, 1990 (in press).

9. Brachman RJ: On the Epistemologic Status of Semantic Networks, in Findler NV, ed.: *Associative Networks: Representation and Use of Knowledge by Computers*; Academic Press, New York, 1979:3-50.

10. Packer MS, Cimino JJ, Barnett GO, Kim R, Hoffer EP, Zatz S: Updating the DXplain Database, in Greenes RA, ed.: *Proceedings of the Twelfth SCAMC*; Washington, D.C.; November, 1988:96-100.

11. 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 SCAMC*; Washington, DC; Nov, 1988:148-51.

12. Komorowski HJ, Greenes RA, Barr C, Pattison-Gordon E: Browsing and Authoring Tools for a Unified Medical Language System. In: Fluhr C, Walker D, eds.: *Proceedings of RIAO 88 Conference on User-Oriented Context-Based Text and Image Handling*. Cambridge: MIT, 1988:624-641.

13. Masarie FE, Cimino JJ, Giuse NB, Miller RA: Mapping Between QMR Manifestations and DXplain Terms, Report to the National Library of Medicine for the Unified Medical Language System; Bethesda; April 7, 1988.

14. Cimino JJ, Barnett GO: Automated Translation Between Medical Terminologies Using Semantic Definitions, *MD Computing*, March-April, 1990:104-109.

15. Cimino JJ, Mallon LJ, Barnett GO: Automated Extraction of Medical Knowledge from Medline Citations, in Greenes RA, ed.: *Proceedings of the Twelfth SCAMC*; Washington, D.C.; November, 1988: 180-184.