Representation of Clinical Laboratory Terminology in the Unified Medical Language System

James J. Cimino, M.D.
Center for Medical Informatics
Columbia University
New York, New York 10032

The Unified Medical Language System (UMLS) was examined to determine its coverage of clinical laboratory terminology in use at the Columbia-Presbyterian Medical Center (CPMC). The Metathesaurus (Meta-1) contains exact matches for 30% of 1460 CPMC laboratory terms and near matches for an additional 42%, with better coverage of atomic-level concepts ("substance" terms) than complex ones (tests and panels). The Semantic Network includes types for representing laboratory procedures (2), measured substances (at least 56) and sampled substances (at least 14), but no type to represent specimens. Few of the UMLS semantic relationships are applicable to the CPMC vocabulary. These results have implications for the utility of the UMLS for linking clinical databases to electronic medical information sources.

Introduction

The Unified Medical Language System (UMLS) was developed by the National Library of Medicine (NLM) to facilitate use of information from computerized biomedical information sources. The UMLS was released in September of 1990 for experimental purposes with the understanding that experience obtained during evaluation and use would be provided to the NLM [1]. This feedback may influence future UMLS versions [2].

The use of electronically available patient information is envisioned as ideal input to systems which will make use of the UMLS to retrieve biomedical information relevant to solving patient problems [2]. Other workers have explored the use of the UMLS for representing clinical data [3], for classification of patient records [4], and for performing bibliographic retrievals from medical reports [5]. This paper describes results of a preliminary evaluation of UMLS coverage of clinical laboratory terminology in use at Columbia-Presbyterian Medical Center (CPMC).

Background

Controlled vocabularies used by clinical applications at CPMC are being integrated into a Medical Entities Dictionary (MED) which makes use of a semantic network representational scheme to facilitate browsing and maintenance functions [6]. As controlled terms from CPMC applications are added to the MED, two additional kinds of information are added with them: classification of the terms and semantic relationships between the terms being added and other controlled terms in the MED. For example, the term "Serum Glucose Measurement" is classified as "Laboratory Diagnostic Procedure". It is capable of having a number of semantic relationships including "Specimen", "Part of" and "Substance Measured". It is actually linked through these relationship to the terms "Serum Specimen", "Chem-7" and "Glucose", respectively. The terms "Serum Specimen" and "Chem-7" are included in the MED because they are CPMC laboratory terms. "Glucose" is not found in the current CPMC laboratory terminology and must be added to the MED. The controlled terminology examined in the present study includes 900 terms for laboratory tests, 155 terms for laboratory panels (composed of two or more tests), and 170 specimen terms. There are 306 terms for substances which are measured by tests and 99 additional terms for substances which are sampled to produce specimens.

The UMLS is an attractive source for providing classes, terms and semantic links used in the MED, alleviating some of the burden of MED maintenance. For example, there is no need to create a chemical vocabulary if needed chemical terms (such as "glucose", in the above example) already exist in the UMLS. Using UMLS components wherever possible offers the opportunity map local CPMC terms to those of the external information sources from which the UMLS has been constructed. For example, literature searches based on laboratory results could be facilitated by translating the laboratory terms into Medical Subject Heading (MeSH) terms for Medline searches. UMLS design has influenced the MED structure [7]. This study examines the potential contribution by the UMLS to MED content.

0195-4210/91/$5.00

Methods

This study sought to determine the suitability of the UMLS as a source for classes, terms and semantic links needed to build the CPMC MED. The approach taken was to enumerate the different components of the current version of the MED and them to the UMLS Semantic Network and Metathesaurus (Meta-1).

Data Structures

Semantic information about the 2248 terms in the MED was read using a PC-based MUMPS package (DTM-PC, DataTree Inc., Waltham, MA) to create a hierarchical data structure (a MUMPS global) which contains all of the hierarchical information and the semantic links between terms. The UMLS files were read from the CD-ROM to DOS ASCII files on a hard disk. From there, they were processed using MUMPS to create a MUMPS global containing the Meta-1 Unit Record fields, indexed by the Meta-1 Unique Identifier. A word index was created which provided access to each preferred name for each of the 78,892 Concept, Synonym and Related Term entries in Meta-1. Analysis involving Meta-1 terms made use of the Semantic Type and Synonym fields and the word index global to help focus the manual reviews. The analysis of the UMLS Semantic Network was performed manually by reviewing all fields in the unit-record ASCII file provided on the CD-ROM.

Term Classification

The 131 "semantic types" in the UMLS Semantic Network [8] were reviewed manually to identify appropriate types for laboratory terms (tests, panels, and specimens) and laboratory-related terms (substances measured by tests and substances sampled to produce specimens). Each CPMC term was matched to the most specific appropriate semantic type available, based on the UMLS definitions. For example, while "glucose" could be appropriately classified as a "chemical", a more specific type, "carbohydrate" exists.

Laboratory Terms

Meta-1 Main Concepts, Synonyms or Related Concepts were reviewed to identify those which were either synonyms of CPMC terms ("exact matches") or more general concepts which could be considered to represent classes containing the CPMC terms ("near matches"). The first step in the review was to retrieve Main Concepts with an appropriate semantic type (and Synonyms of Main Concepts with an appropriate semantic type) which have at least one word in common with each CPMC term. Frequently-appearing words such as "measurement" and "test" were excluded from the process. The resulting list was scanned to identify a matching term. If none was found, the retrieval was expanded with additional words. For example, in matching the CPMC term "O2 Partial Pressure Measurement", the reviewer choose to add "PO2" to the search, producing a list of all Meta-1 terms which had been assigned one or more of the desired semantic types (see below) and contained either the "O2" "Partial", "Pressure" or "PO2" ("Measurement" was ignored). "Arterial pO2" was found (a near match).

Semantic Relationships

For each semantic type found to be relevant to laboratory terms, the semantic relationships in the UMLS Semantic Network were identified and their definitions were compared to those used for incorporating terms into the CPMC MED semantic network.

Supplemental Terms

For each of the supplemental terms ("Measured Substances" and "Sampled Substances"), Meta-1 was searched using the method described above for laboratory terms. No restriction was placed on the semantic types reviewed, to facilitate discovery of additional useful types for representing these terms.

Results

Term Classification

Two semantic types in the UMLS Semantic Network may be appropriate for classifying test and panel terms. These are "Laboratory Procedure" and "Diagnostic Procedure". The former is defined as "A procedure, method or technique used to determine the composition, quantity or concentration of a specimen, and which is carried out in a laboratory", while the latter is defined as "A procedure, method or technique used to determine the nature or identity of a disease or disorder [and] *excludes procedures which are primarily carried out on specimens in a laboratory*" (italics added). Review of the 1055 Meta-1 terms with these types showed that semantic types do not always appear to be assigned consistently.

Most Meta-1 "Laboratory Procedures" are laboratory analytic techniques which are devoid of specific diagnostic implications (e.g., "Gel Filtration), while most "Diagnostic Procedures"

refer to non-laboratory tests with specific diagnostic implications (e.g., "Heart Auscultation"). In many cases, however, the reason for choice of semantic type assignment is obscure. For example, it is unclear why "Adrenal Cortical Function Tests", "Blood Tests", "Densitometry, X-Ray" and "Erythrocyte Size Determination" are considered diagnostic procedures while "Acoustic Stimulation" and "Densitometry" are considered "laboratory procedures". A few "laboratory" terms appear to refer to statistical procedures (e.g "Discriminant Analysis") and 14 Meta-1 terms are listed as both "laboratory" and "diagnostic", despite the mutual exclusivity (by definition) of these types. I an random sample of 5% of other Meta-1 terms, none were found which might have been assigned to one of these two types but was not.

The 306 substances measured by procedures and tests can be represented by the types "Anatomical Structure", "Chemical", "Phenomenon or Process" or "Organism", or one of 56 more specific types (e.g., "Cell", "Immunologic Factor", "Pharmacologic Substance" and "Virus"). CPMC specimen terms are not properly classifiable by UMLS semantic types. It might seem that types such as "Body Substance" would serve this purpose; however, CPMC specimen terms also refer to collection method (sterile collection, anaerobic conditions, biopsy, etc.). However, the 99 sampled substances can be classified under the UMLS types "Anatomical Structure" or "Physical Object", or one of 14 more specific types (e.g., "Tissue", "Body Substance" and "Medical Device").

Laboratory Terms

When the 900 laboratory test terms and the 155 panel terms were compared to Meta-1, 666 (63%) (570 tests and 96 panels) were judged to be in Meta-1 as either exact matches (119 of the tests and 33 of the panels) or as more specific forms of Meta-1 terms (448 of the tests and 63 of the panels). No match was found for the remaining 389 CPMC panel terms (such as "Insulin Tolerance Test", "Clinical Chemistry Lipid Profile" and "Sweat Test") and test terms (such as "CSF Protein Measurement", "Plasma Zinc Measurement" and "Sweat Chloride Ion Measurement").

Semantic Relationships

The type "Diagnostic Procedure" has three links in the UMLS Semantic Network: "assesses_effect_of" (linked to "Chemical"), "diagnoses" (linked to "Pathologic Function", "Injury or Poisoning", "Congenital Abnormality", and "Acquired Abnormality"), and "measures" (linked "Organism Attribute" and "Physiologic Function"). The type "Laboratory Procedure" has these three links and plus the link "has_result" (linked to "Laboratory or Test Result").

The "has_result" link corresponds with a relationship in the MED linking test terms with result terms. The "diagnoses" link is not used because the MED deals with the definitions of terms rather than the sum of medical knowledge about them [6]. The "measures" link "ascertains or marks the dimensions, quantity, degree, or capacity of organism attributes and physiologic functions". The MED uses "substance measured" which has a similar definition but is associated with terms of type "Anatomical Structure", "Chemical", "Phenomenon or Process", or "Organism". The "assesses_effect_of" link "analyzes the influence or consequences of the function or action of a chemical". Thus, a UMLS procedure which "assesses_effect_of" digoxin might "measure" cardiac output (a physiologic function), while the CPMC procedure "Serum Digoxin Test", measures digoxin itself, rather than its effect.

Two additional semantic relationships are used to describe laboratory tests and panels in the MED which are not explicitly supplied by the UMLS Semantic Network. One of these, the "has part" relationship, is used to link panels to their component tests. Although the UMLS Semantic Network has such a relationship, it is not associated with the types "Laboratory Procedure" or "Diagnostic Procedure". The MED also includes a "specimen" relationship to allow association between test and panel terms and specimen terms.

Supplemental Terms

Of the 306 measured-substances terms, 216 (70.6%) exact matches were found in Meta-1. In 76 (24.8%) other cases, a more general Meta-1 term was found and in only 14 (4.6%) of cases could no appropriate Meta-1 term be found (e.g., "Helicobacter Pylori", "Leucocyte Cast" and Myelocyte"). The 292 matching Meta-1 terms fell into 19 types that were among the expected 56 types in the UMLS Semantic Network. Of the 99 sampled-substances terms, 74 exact matches were found in Meta-1. In 20 other cases, a more general Meta-1 term was found and in only 5 cases ("Fomite", "Labium", "Lochia", "Serum" and "Surface") could no appropriate Meta-1 term be found. The 94 matching Meta-1 terms fell into 11 types that were among the expected 14 types.

	CPMC Test Terms	Meta-1 Terms
*	Stat Blood Hematocrit Measurement	Hematocrit
*	Urine Protein Measurement	Protein, urine
	Presbyterian Red Cell Volume Measurement	Erythrocyte Indices
	Ureteral Urine Culture	Culture
	CPMC Panel Terms	**Meta-1 Terms**
*	Chem-7	Chem 7
*	Thyroid Screen	Thyroid Function Tests
	Cell Count, Cerebrospinal Fluid	Cell Count
	Influenza Virus Complement Fixation Test	Complement Fixation Tests
	CPMC Measured Substance Terms	**Meta-1 Terms**
*	Anaerobic Bacterium	Bacteria, Anaerobic
*	Anti DNA	Anti-DNA Antibodies
*	Doxepin	Doxepin <1>
	Mycoplasma Hominis	Mycoplasma
	Sickle Cell	Erythrocytes, Abnormal
	CPMC Sampled Substance Terms	**Meta-1 Terms**
*	Blood	Blood <1>
	Bowel Contents	Gastrointestinal Contents
	Facial Lesion	SKIN LESION
	Intravenous Catheter	Catheters, Indwelling
*	Cervix	Uterine cervix

Table 1: Samples of CPMC terms and matching Meta-1 terms (* indicates synonym matches)

Discussion

In this study, 666 of 1055 laboratory tests and panels, 292 of 306 measured substances, and 94 of 99 sampled substances which are represented in the CPMC laboratory vocabulary were found in Meta-1. Exact matches were found for 442, for a success rate of 30% and an overall success rate of 72%. It is important to note that this was a preliminary study with a single observer assisted by a simple lexical matching program. Multiple observers armed with more sophisticated matching techniques might have produced different results. Inadequacies in the present methods would tend to favor under-matching, rather than over-matching, suggesting that, to the extent which the CPMC vocabulary is representative of clinical laboratory terminology at other sites, 72% is a lower bound for UMLS *coverage* of laboratory terms.

With the exception of "Specimen", the UMLS Semantic Network appears to provide coverage of semantic types found in laboratory terminology. The UMLS semantic relationships (designed to represent relationships found in various knowledge sources) proved to be less useful as a source for relationships which exist in the CPMC vocabulary (which are needed to provide information about term meaning). If relationships such as "part of", "specimen", "substance measured" and "substance sampled" are found in UMLS-accessible information sources, their inclusion in the UMLS would be appropriate.

The 72% success rate could be rated as surprisingly good, considering that the UMLS was not derived from any vocabularies designed specifically for laboratories. The 30% rate for exact matches reflects generic nature of laboratory terms used in the information sources to which the UMLS provides access. If one's objective is to obtain information from those knowledge sources relevant to a particular CPMC laboratory test, the inexact match may prove to be sufficient for obtaining the desired information.

The lexical techniques used here for translation were tedious. More intelligent methods exist which make use of semantic pattern matching by taking advantage of semantic information in the MED [9]. For example, if a test is not found in Meta-1, its equivalent might be obtained by translating the terms for the substances sampled and measured by the test.

Rather than attempting to translate the test term "CSF Protein Measurement", the semantic approach can translate its measured substance and the sampled substance of its specimen (protein and cerebrospinal fluid, respectively - both are UMLS terms). Improving translation through such semantic decomposition has yet to be studied.

Semantic matching techniques could be further applied if the UMLS included more specific semantic information than is presently available. The framework for including such information is already in place, and it could offer advantages for other tasks, such as knowledge-based vocabulary maintenance [6,10]. The addition of such knowledge would be a large task in itself; however, there is evidence that some of the necessary information could be automatically obtained from information sources such as MEDLINE [11].

The NLM has expressed a commitment to expanding the UMLS; the direction of that expansion remains to be determined. Laboratory coverage can be increased through studies such as this one or by adding to the UMLS a vocabulary for representing laboratory terms (e.g., International Classification of Clinical Services [12]). However, translating local laboratory terms to Meta-1 terms is not an end in itself, but a means to facilitate access to relevant information sources. For example, a clinician viewing a laboratory result could initiate a literature search in which the laboratory term was translated to MeSH to retrieve citations about the result (such as about its meaning or reliability). Similarly, patient results could be translated for use by a medical diagnosis program to generate a differential diagnosis, without having to repeat any data entry. The feasibility and utility of such automated information retrievals remain to be tested. However, translation such as that available through the UMLS-based approach used in this study is a first step toward those retrievals.

Acknowledgements

This work was supported by IBM and an NLM Integrated Academic Information Management Systems (IAIMS) grant. The UMLS was provided by the NLM under experimental agreement. The author thanks his colleagues at the Center for Medical Informatics, the SCAMC referees and his wife for their constructive criticism of this paper.

References

1. Humphreys BL: *UMLS Knowledge Sources - Experimental Edition Documentation*. National Library of Medicine, Bethesda, Maryland; September, 1990.
2. Lindberg DAB, Humphreys BL: The UMLS knowledge sources: tools for building better user interfaces. In Miller RA, ed.: *Proceedings of the 14th SCAMC*; Washington, D.C.; 1990:121-5.
3. Huff SM, Warner HR: A comparison of Meta-1 and HELP terms: implications for clinical data. In Miller RA, ed.: *Proceedings of the 14th SCAMC*; Washington, D.C.; 1990:166-69.
4. Chute CG, Yang Y, Tuttle MS, Sherertz DD, Olson NE, Erlbaum MS: A preliminary evaluation of the UMLS Metathesaurus for patient record classification. In Miller RA, ed.: *Proceedings of the 14th SCAMC*; Washington, D.C.; 1990:161-65.
5. Powsner S, Miller P: Lexical versus conceptual links to a bibliographic database from a medical report (abstract). In Beck RJ, ed.: *Proceedings of Second AMIA Research and Education Conference*, 1991:64.
6. Cimino JJ, Hripcsak G, Johnson SB, Clayton PD: Designing an introspective, controlled medical vocabulary. In Kingsland LW, ed.: *Proceedings of the 13th SCAMC*; Washington, D.C.; 1989: 513-18.
7. 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 14th SCAMC*; Washington, D.C.; 1990:175-80.
8. McCray AT, Hole WT: The scope and structure of the first version of the UMLS semantic network. In Miller RA, ed.: *Proceedings of the 14th SCAMC*; Washington, D.C.; 1990:126-30.
9. Cimino JJ, Barnett GO: Automated translation between medical terminologies using semantic definitions, *MD Computing*, 1990; 7(2):104-9.
10. Tuttle MS, Sherertz DD, Olson N, Sperzel W, Erlbaum M, Fuller L: Adding user-specified terms to the Unified Medical Language System Metathesaurus (abstract). In Beck RJ, ed.: *Proceedings of Second AMIA Research and Education Conference*, 1991:45.
11. Cimino JJ, Mallon LJ, Barnett GO: Automated extraction of medical knowledge from Medline citations. In Greenes RA, ed.: *Proceedings of the 12th SCAMC*; Washington, D.C.; 1988: 180-4.
12. Mendenhall S: The ICCS code: a new development for an old problem. In Stead WW, ed.: *Proceedings of the 11th SCAMC*; Washington, D.C.; 1987; 703-9.