Using the UMLS to Bring the Library to the Bedside

JAMES J. CIMINO, MD, ROBERT V. SIDELI, MD

This paper presents an algorithm that can be used to convert ICD9 terms to related MeSH terms. Preliminary evaluation indicates that together, the algorithm and the UMLS provide a reasonable resource for facilitating such conversions. *Key words*: Unified Medical Language System; bibliographic searches. **(Med Decis Making 1991;11(suppl):S116–S120)**

An important (and arguably the ultimate) goal of medical informatics is to provide support for clinicians faced with patient care decisions. Decision support can take many forms; perhaps the most widely used is retrieval of medical literature citations from bibliographic databases. MEDLINE, developed and maintained by the National Library of Medicine (NLM), is the largest of these databases and has been advocated repeatedly as a useful tool for supporting clinical decision-making.[1–5] Literature searches not made at the moment of decision making are typically deferred or forsaken, unless the decision itself can be delayed. The incorporation of access to MEDLINE with other computer applications involved in the decision-making process (such as order entry and results review) has the effect of "bringing the library to the bedside." An obvious advantage of this approach is the ability to provide medical literature in a timely manner. This approach also offers the opportunity to facilitate the search process by providing relevant clinical data for direct incorporation in the search strategy. For example, if a physician is planning to order a test and needs more information about the test,[5] why not let the order entry program provide the name of the test directly to the search engine?

One difficulty with performing bibliographic searches using clinical data is that the terminology used in the clinical setting may not prove useful for the task of literature retrieval. The NLM expends a great deal of effort in indexing literature citations using the Medical Subject Headings[6] (MeSH). Medical librarians trained in the use of MeSH for MEDLINE searches convert the desired clinical terms to the appropriate MeSH terms to improve their retrievals. Such a resource would be difficult to provide at the bedside. However, in September 1990, the NLM released the experimental first version of its Unified Medical Language System,[7] which is intended to facilitate access to computer-based medical knowledge resources (such as MEDLINE).[8] One of the intended uses of the UMLS is to provide a means for automating the translation of medical terms to appropriate controlled vocabulary terms (like MeSH).[9]

At the Columbia–Presbyterian Medical Center, we are experimenting with the UMLS to help translate from clinical terms to MeSH terms for the purpose of automating literature searches during entry and review of a computer-based clinical information. The clinical information of interest consists of intensive care unit (ICU) diagnoses and discharge diagnoses, both of which are recorded using the *International Classification of Diseases with Clinical Modifications, Ninth Revision*, commonly referred to as ICD9.[10] Although the UMLS includes some ICD9 terms and all MeSH terms, it provides neither the software nor the algorithms for translating between them. This paper describes our experience with developing and testing algorithms for using the UMLS to translate ICD9 to MeSH.

Methods

UMLS DATA STRUCTURES

Information deemed relevant to the translation process was drawn from the UMLS Metathesaurus (Meta-1). This included the main concept name or preferred name {MC} (which is preferentially taken from MeSH), the unique identifier {UI}, the source field {SO}, the synonym field {SY} and the related term fields {RRT and URT}. (Identifiers in { } refer to field names found in the UMLS documentation.[7]) From these data, two new data structures were constructed with the use of MUMPS (DTM-PC, Data Tree, Waltham, MA) globals.

The first of global is a table of ICD9 codes found in the source fields, with pointers to the Meta-1 concepts

Received from the Center for Medical Informatics, Columbia University, New York, New York.

Address correspondence and reprint requests to Dr. Cimino: Center for Medical Informatics, Atchley Pavilion, Room 1310, Columbia-Presbyterian Medical Center, New York, NY 10032.

*The Meta-1 documentation states that 2,808 preferred terms are included, but we find 2,402 disease codes, 334 procedure codes, 40 external causes of injury ("E" codes), and 67 health status factors ("V" codes).

(unique identifiers) in which they were found. The pointers for each ICD9 code are differentiated by whether the corresponding Meta-1 concept is an ICD9 Preferred Term (PT), Abbreviation (AB) or Index Term (IT).

The second global is a table of Meta-1 concepts with pointers to other Meta-1 concepts. The pointers for each Meta-1 code are differentiated by whether they point to synonyms or related terms. Related term pointers are further differentiated by whether they refer to broader terms, narrower terms or "other." Also stored in this global is the source information for each term, particularly if it is a MeSH Main Heading (MH) or Entry Term (ET).

ALGORITHMS FOR TRANSLATION

Given the above data structures, an algorithm for finding an exact translation between an ICD9 term and MeSH is straightforward. Using the ICD9 code as an index into the first global, the Meta-1 concept which is the ICD9 PT is found. Using the identifier found in the first global as the index to the second global, the source of the Meta-1 term can be determined. If the source indicates that the term is a MeSH MH, the translation process is complete. Of the 78,862 Meta-1 concepts, 15,610 MeSH MH's and 2,843 ICD9 PT's,* only 686 such confluences occur. More typically, either the Meta-1 concept identified with the ICD9 code is not a MeSH MH or the ICD9 code does not appear in Meta-1 at all.

In the first case (the ICD9 code is not a MeSH MH), an algorithm was developed for traversing the pointers between entries in the second global. The algorithm includes precedence rules for choosing among multiple pointers, when they occur. The algorithm first examines synonyms of the ICD9 term to determine whether any are MeSH MH's. If not, the "broader related terms" are then examined, followed by the "other related terms," followed by the "narrower related terms."

In the second case where an exact match is not found (the ICD9 code is not in Meta-1) or when the precedence search described above fails to locate a related MeSH MH, a second algorithm is applied which takes advantage of the hierarchical coding scheme employed in ICD9. Specifically, the removal of the fourth or fifth digit of an ICD9 code yields an ICD9 code that corresponds to a term that is "broader than" the original code. For example, the code 123 is broader than the code 123.1 and both are broader than the code 123.11. By exploiting this feature of ICD9, the "hit rate" for finding ICD9 codes in Meta-1 is expanded. Of the 12,278 disease codes in ICD9, 3,153 are found in Meta-1 and another 2,369 can be "found" by using this truncation technique.

Figure 1 shows one example of how the algorithms are employed to locate a MeSH terms which corresponds to an ICD9 term. The ICD9 term "TOBACCO ABUSE-CONTINUOUS" (code 305.11) does not appear in Meta-1. However, the term "Tabagism syndrome," which is an index term for the ICD9 term "TOBACCO ABUSE" (code 305.1), does appear (Meta-1 UI 39221). It is identified as a synonym of the SNOMED (*Systematized Nomenclature of Medicine*[11]) Preferred Term "Smoker's respiratory syndrome" (SNOMED code D-7760; Meta-1 UI 37368). This new term is, in turn, identified as being related (by the "Other" relationship) to the MeSH MH "Tobacco" (MeSH unique identifier D014026; Meta-1 UI 40329). Thus, in this example, the algorithm would find "Tobacco" as the closest MeSH term to the ICD9 term "TOBACCO ABUSE-CONTINUOUS."

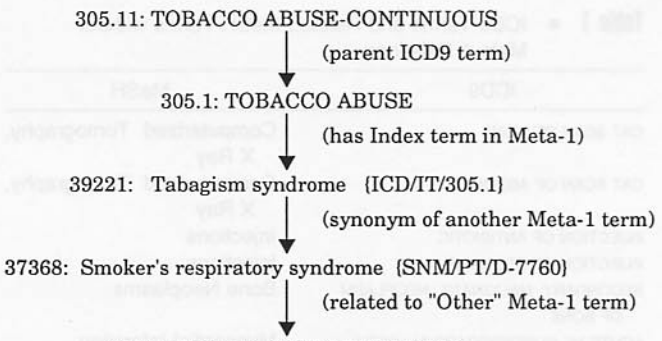

FIGURE 1. Example of locating a MeSH term related to an ICD9 term in Meta-1. The decimal numbers (305.11 and 305.1) are ICD9 codes, the other five-digit numbers (39221, 37368 and 40329) are Meta-1 codes, and the codes enclosed in {}s represent the Meta-1 source field (SO) information. See text for a complete description of the translation process.

DATA SETS

In order to measure the algorithm's ability to match MeSH terms to clinically relevant ICD9 terms, two test sets were constructed. The first was a list of diagnoses drawn from a database of 3,590 patient records in the intensive care unit system (the ICU Test Set). The second was a list of all ICD9 diseases and procedures drawn from hospitalizations for the ten most recently hospitalized patients (as of November 1, 1990) in one physician's practice (JJC) (the Inpatient Test Set). The translation algorithm was applied to each of the terms in the test sets. In each case, the resulting MeSH term (if any) and the method used (precedence of link and, if needed, truncation) were noted.

LITERATURE SEARCHES

In order to form a rough estimate of the usefulness of translating ICD9 terms to MeSH, literature searches were performed using the ICD9 terms in the Inpatient Test Set and, where found by the translation algorithm, the corresponding MeSH term. Searches were performed using BRS/Onsite, running under MVS on an IBM 3090, Model 200 and containing five years of MED-

Table 1 • ICD9 Terms and Related MeSH Terms without Meta-1 Relations

ICD9	MeSH
CAT SCAN OF HEAD	Computerized Tomography, X Ray
CAT SCAN OF ABDOMEN	Computerized Tomography, X Ray
INJECTION OF ANTIBIOTIC	Injections
INJECTION OF STEROIDS	Injections
SECONDARY MALIGNANT NEOPLASM OF BONE	Bone Neoplasms
ACUTE MI, SUBENDOCARDIAL, INITIAL	Myocardial Infarction
OLD MYOCARDIAL INFARCT	Myocardial Infarction

LINE files. The ICD9 terms were regularized in an attempt to improve the search results: abbreviations were expanded, expressions such as "not elsewhere classified" and "not otherwise specified" were dropped, and phrases such as "TOBACCO ABUSE-CONTINUOUS "were rearranged to make them more English-like (i.e., CONTINUOUS TOBACCO ABUSE). The regularized ICD9 terms and the MeSH terms were entered into BRS manually, each as an individual search, and the number of citations was noted.

Results

TRANSLATING THE ICU TEST SET

The ICU Test Set contained 699 unique ICD9 terms, with each term appearing an average of 5.14 times (range, one to 140). Of these 699 ICD9 terms, MeSH terms were found in Meta-1 for 271 (39%): 78 matched MeSH terms exactly, 38 were synonyms of MeSH terms, and 155 were related to MeSH terms. For the 428 ICD9 terms that could not be found in Meta-1, 97 (14% of the original 699) were found using the truncation technique: 53 exact matches, 12 synonyms, and 32 related terms. Thus, there was an overall "hit rate" of 53%. The 368 terms which matched corresponded to 82% of the patient diagnoses in the test set. The hit rate corresponded to the frequency with which terms appear in the test set. For example, for terms appearing more than seven times, the hit rate was 76%; for terms appearing more than 30 times, the hit rate was 100%.

TRANSLATING THE INPATIENT TEST SET

The Inpatient Test Set contained 55 unique ICD9 terms, 40 of which were disease terms and 15 of which were procedure terms. Of these 55 ICD9 terms, MeSH terms were found in Meta-1 for 40 (73%): 16 matched MeSH terms exactly, seven were synonyms of MeSH terms, and 17 were related to MeSH terms. For the 15 ICD9 terms that could not be found in Meta-1, five (9% of the original 55) were found using the truncation technique: four exact matches and one related term. Thus, there was an overall "hit rate" of 82%. The hit rate was somewhat higher for disease terms (90%) than for procedure terms (60%).

For the four diseases and six procedures that were not matched to MeSH terms by any automated algorithm, a manual search was performed of the Meta-1 terms to attempt to identify potentially related MeSH terms. In seven cases, related MeSH terms were identified (see table 1).

Table 2 • Comparison of MEDLINE Searches Using ICD9 and Related MeSH Terms

Search Terms	Successful Searches	Average Number of Citations	Unsuccessful Searches
ICD9 Terms with MeSH Matches	33	1,321	12
ICD9 Terms without MeSH Matches	4	27	6
MeSH Terms	45	4,319	0

LITERATURE SEARCHES

A total of 100 literature searches were performed (for 55 ICD9 terms and 45 MeSH terms). Table 2 shows, for translated and untranslated ICD9 terms and for MeSH terms, the number of successful searches (any citations) and the average number of citations in the successful searches.

Discussion

The work discussed in this paper is relevant to those who seek to use the UMLS information sources for translating between the controlled vocabularies found in Meta-1. Although the current work deals exclusively with ICD9-to-MeSH translation, the general methods can be extended, with two caveats. First, the truncation method used here is not generally applicable for translating *from* other controlled vocabularies; however, the general notion of substituting a general term when the more specific term is unsatisfactory can be applied when using any hierarchical vocabulary, so long as the hierarchy embodies the class–subclass relationships. Second, translation *to* vocabularies other than MeSH is unlikely to be as successful in the present version of the UMLS, since most other UMLS source vocabularies (such as ICD9) are not fully represented in Meta-1.

While the algorithm presented here should be transferable to other settings and applications, much work remains to be done with respect to its refinement and evaluation. No attempt was made here to rate the "quality" of the match. Certainly, one could take issue with the example shown in figure 1: "Tobacco" would

not generally be considered a proper translation of "TOBACCO ABUSE-CONTINUOUS." However, it makes little sense to attempt to rate such "translations" out of context. If the purpose is to retrieve appropriate MEDLINE citations using MeSH terms, and if no other MeSH term is applicable, then such "poor" translations might be deemed acceptable.

The precedence for selecting among possible MeSH terms was chosen in a purely empiric manner. In many cases, the precedence was of little consequence, since there was only one possible path to a MeSH term. In other cases, the precedence appears sensible, e.g., a direct match is desirable and a synonym match is the next best thing. However, the simplicity begins to deteriorate when additional choices are considered. For example, is "narrower" better than "other"? Would it be better to truncate (a *de facto* "broader" link) than to use "narrower"? When is "other" acceptable? Should the result of using "narrower" be the list of all MeSH terms found this way? And finally, should an "append" method be applied to the ICD9 codes to synthesize *de facto* "narrower" links? Answers to these questions await an evaluation which compares the methods and the quality of translations.

Developing such an evaluation would best be done in the context of literature retrieval, since the broad purpose is to retrieve citations, not MeSH terms. Such an evaluation would require knowing the reason for the search in order to compare different strategies. The searches presented here were done in the absence of an associated clinical question. The results, therefore, are of limited value. It should be no great surprise that MeSH terms retrieve more citations than do ICD9 terms. The results suggest that the sensitivity of the MeSH searches is greater than the ICD9 searches, especially when the ICD9 searches failed to find any citations. However, without having a purpose behind each search, this remains no more than a suggestion. In fact, the possibility exists that the specificity of the ICD9 searches may be much greater than that of the MeSH searches. If this were the case, then a collection of 27 high-quality citations would be preferable to wading through a collection of 4,319 citations less specific ones. The most that can be said of the searches performed in this study is that they suggest that performing such translations is beneficial but specific evaluation is needed.

Further evaluation will await more convenient methods for performing the searches. At the time that this study was carried out, only the translation of terms was automated. The accumulation of ICD9 codes from patient records and the performance of the searches were carried out by hand. Since that time, a new application that displays the ICD9 diagnoses and procedures associated with all of a patient's hospital admissions has been made available to clinicians at the Columbia–Presbyterian Medical Center. The establishment of a means for inserting these terms into a search strategy and transferring it to the on-site BRS search engine (with access to five years of MEDLINE) has yet to be realized.

Despite the lack of a robust evaluation of algorithm performance, as an evaluation of the UMLS, this work has produced some interesting results. First, despite the statement by the NLM that Meta-1 contains 2,808 of ICD9 (less than 16% of the 18,317 codes in all of ICD9),[7] we found that the ICD9 terms used in two different settings in our hospital seemed to be much better represented. The representation seems even more favorable when the "false negatives" of the translation (shown in table 2) are considered. For example, in the small inpatient sample, only one of 40 diseases could *not* be found when automated and manual methods were combined. This suggests that the UMLS may cover a larger portion of the domain of clinical medicine than might be expected from the initial description of the UMLS content. It also suggests that the inclusion of more ICD9 terms would be consistent with the current domain of the UMLS, since the terms already in Meta-1 appear to be related to the as-yet-unadded ICD9 terms (although not, of course, by explicit links).

Conclusion

This paper presents an algorithm that can be used to convert ICD9 terms to related MeSH terms. Preliminary evaluation indicates that together, the algorithm and the UMLS provide a reasonable resource for facilitating such conversions. The conversion of ICD9 terms to MeSH through the UMLS leads the way toward recruiting a wealth of electronic clinical information for automating access to literature which can be of use in medical decisions, at the time those decisions are being made.

This work was supported in part by a National Library of Medicine Integrated Academic Information Management Systems (IAIMS) grant. Dr. Cimino receives support from the International Business Machines Corporation. Dr. Sideli is supported by an NLM Training Grant.

References

1. Haynes RB, Walker CJ. Computer-aided quality assurance. A critical appraisal. Arch Intern Med. 1987;147:1297–1301.
2. Huth EJ. The underused medical literature. Ann Intern Med. 1989;110:99–100.
3. Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med. 1990;112:78–84.
4. Safran C, Herrmann F, Rind D, Kowaloff HB, Bleich HL, Slack WV. Computer-based support for clinical decision making. MD Comput. 1990;7:319–22.
5. Guyatt GH. Evidence-based medicine. Ann Intern Med. 1991;114(suppl.2):A-16.
6. National Library of Medicine, Library Operations. Medical subject headings. Bethesda, MD, 1989.

7. National Library of Medicine. UMLS knowledge sources—experimental edition documentation. Bethesda, MD, September 1990.
8. Lindberg DAB, Humphreys BL. The UMLS knowledge sources: tools for building better user inferfaces. In: Miller RA, ed. Proc Fourteenth Annual Symposium on Computer Applications in Medical Care, Washington, DC, November 1990;121–5.
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 engineering. International Medical Informatics Association. Amsterdam: Elsevier (North-Holland). 1989;5–17.
10. United States National Center for Health Statistics. International classification of diseases, ninth revision, with clinical modifications. Washington, DC, 1980.
11. Côté RA, ed. Systematized nomenclature of medicine. 2nd ed. College of American Pathologists, Skokie, IL, 1982.