From ICD9-CM to MeSH Using the UMLS: A How-to Guide

James J. Cimino, M.D., Stephen B. Johnson, Ph.D.,
Ping Peng, Ph.D., Anthony Aguirre, M.S., M.L.S.
Center for Medical Informatics
Columbia Presbyterian Medical Center
New York, New York 10032

One purpose of the Unified Medical Language System (UMLS) is to facilitate conversion of terms from one controlled medical vocabulary to another. We examined our ability to convert International Classification of Diseases, 9th Edition, Clinical Modifications (ICD9-CM) to Medical Subject Headings (MeSH) using the UMLS. We describe a method which mapped 30.4% of ICD9-CM to UMLS. Of these, 95.0% were linked to MeSH, of which translation was straightforward in 90.4%. We discuss the use of these translations for retrieval from MeSH-indexed databases, such as Medline.

INTRODUCTION

Satisfying clinician information needs which arise in patient encounters often requires access to up-to-date literature [1]. On-line literature is a potential[2], but underutilized [3] solution. Barriers to access, lack of familiarity with use, and time requirements are cited as reasons for this underutilization. Programs have been developed to address these problems by using computer-based patient information to formulate search strategies and carry out retrievals [4,5,6]. These programs allow the user to identify free-text information in an electronic patient record, translate it to the Medical Subject Headings (MeSH) [7], and conduct an automated search of bibliographic databases. We have described a similar program (the Medline Button [8]) which allows a user to select information coded in the International Classification of Diseases, 9th Edition, Clinical Modifications (ICD9-CM) [9]. This paper examines the use of the National Library of Medicine's Unified Medical Language System (UMLS) [10] for the translation between two of its source vocabularies (ICD9-CM to MeSH), exposes some pitfalls of automated translation and explores potential remedies.

BACKGROUND

The UMLS is being developed to serve many information retrieval needs, including identifying appropriate information sources, constructing source-specific retrieval strategies, and converting user terms to source-specific terminology [11]. For this last feature, the UMLS includes a Metathesaurus [12], constructed from several terminologies (including all of MeSH and a substantial portion of ICD9-CM) with disparate terms integrated as synonyms and related terms. Translation of an ICD9-CM term to MeSH, using the Metathesaurus, depends on several factors. First, the ICD9-CM term must be in the UMLS. Second, it must be linked to one or more appropriate MeSH terms. Third, conversion to MeSH must be conducted properly. This is straightforward when the ICD9-CM term is identical to or synonymous with a single MeSH term; it is more complex when the ICD9-CM term is related to multiple MeSH terms.

METHODS

Step 1: ICD9-CM Terms in the UMLS

The 1992 ICD9-CM Codes and Titles were obtained from HCIA, Inc. (Ann Arbor Michigan). The MRSO (sources) file of the 1992 UMLS Metathesaurus was examined for records containing "ICD" in the Source Abbreviation (SAB) field and "PT" ("preferred term") in the Term Type (TTY) field. The codes in the Source Code Unique Identifier (SCD) field were identified as *represented codes*; remaining ICD9-CM codes, not encountered in MRSO, are *unrepresented codes*. Metathesaurus concepts with represented codes are *ICD9 concepts*. Previous work has resulted in a reorganization of the ICD9-CM hierarchy to reflect a more natural classification, unrestricted by the ICD9-CM decimal coding scheme [13]. For example, the ICD9-CM term "Early Syphilis, Latent, Unspecified" (092.9) appears under the "Early Syphilis, Latent" (092). In our hierarchy, it appears under "Latent Syphilis, Unspecified" (097.1) - a relationship which is not evident from the ICD9-CM coding. We expanded coverage by the UMLS for unrepresented ICD9-CM by looking for parent terms in the hierarchy which correspond to represented codes. Thus, if 092.9 is not a represented code, but 092 or 097.1 is, then 092.9 would be identified as a *surrogate code*.

Step 2: Links from ICD9-CM to MeSH

The UMLS maps terms from source vocabularies to unifying concepts. We identified ICD9-CM terms in the MRSO file which mapped to the same concepts as MeSH terms (Main Headings and Entry Terms); These were considered to be synonymous. We examined the UMLS file MRATX for *associated expressions* for all ICD9-CM concepts not found to be synonymous with a MeSH term. (Associated expressions are complex expressions of MeSH Main Headings and Subheadings or Library of Congress Headings which are synonyms of the associated. For example, the Metathesaurus concept "Neck Pain" (ICD9-CM code 723.1) has the associated expression "<Neck> AND <Pain>".) Finally, the UMLS file MRREL was examined for interconcept relations ("broader", "narrower", "like" and "other") between ICD9-CM concepts and MeSH terms.

Step 3: Converting ICD9-CM to MeSH

Links between ICD9-CM terms and MeSH terms are necessary but not sufficient to accomplish translation from the former to the latter. Given the potential variety of links from an ICD9-CM term to one or more MeSH terms, a precedence was set for choosing among multiple links where they existed: (1) when the ICD9-CM term was synonymous with a MeSH Main Heading, the Main Heading was used; (2) when the ICD9-CM term was synonymous with a MeSH Entry Term, the Main Heading for that Entry term was used; (3) when an Associated Expression was available, it was chosen; and, if none of the above were present, then (4) an expression was created from the MeSH terms which were available through the Metathesaurus relations. Since there is no set algorithm for arranging related terms into MeSH expressions, we studied the existing patterns of relationships and developed one. We also translated ICD9-CM terms to MeSH terms through lexical means, using a simple term look-up facility which accepts a character string from the user and retrieves MeSH concepts which contain the string in their name, synonyms or lexical variants. This yielded a list of MeSH terms which were arranged manually into a MeSH expression.

RESULTS

Step 1: ICD9-CM Terms in the UMLS

The 1992 version of ICD9-CM contains 23,942 coded terms. Of these, 2,968 (12.4%) were identified as being represented in the Metathesaurus. An additional 4,305 codes (18.0%) were mappable as surrogate codes. The 2,968 represented codes

correspond to 2,599 ICD9-CM concepts in the Metathesaurus. This occurs because many ICD9-CM terms are treated as synonyms in the UMLS. For example, the ICD9-CM terms "Benign Mammary Dysplasias" (610.), "Benign Mammary Dysplasia, Unspecified" (610.9), "Diffuse Cystic Mastopathy" (610.1), and "Fibroadenosis of Breast" (610.2) all map to the UMLS term "Fibrocystic Disease of Breast". This set of 2,599 ICD9-CM concepts was the subject of subsequent analysis.

Step 2: Links from ICD9-CM to MeSH

The UMLS includes all 16,641 of the 1992 MeSH Main Headings and 8,254 Entry Terms. Table 1 summarizes the links from ICD9-CM to MeSH which we were able to extract from the Metathesaurus. Of the 2,599 ICD9-CM codes, 1112 were found to be synonymous with MeSH terms. For example, "Corticoadrenal Insufficiency" (ICD9-CM code 255.4) was a synonym of MeSH Main Heading "Addison's Disease", while "Schilder's Disease" (341.1) was a synonym of MeSH Entry Term "Encephalitis Periaxialis Diffusa" which, in turn, maps to the Main Heading "Encephalitis Periaxialis". Sixty-one of these terms had associated expressions; two were MeSH expressions: "Chalazion" (373.2) was assigned "<Eyelid Diseases> AND <Cysts> AND <Meibomian Glands>" (all MeSH Headings) and "Diabetic Retinopathy" was assigned "<Diabetes Mellitus>/<complications> AND <Eye Diseases>" (two Headings and one Subheading).

Of the 1,487 ICD9-CM concepts which were not synonymous with any MeSH term, 19 had associated expressions. In 18 of these, the expression consisted of a pair of MeSH Main Headings (for example "Cocaine dependence" was assigned "<Cocaine> AND <Substance Dependence>"). In the remaining case, the expression consisted of a Main Heading-Subheading pair ("Hemorrhoidectomy" was assigned "<Hemorrhoids>/<surgery>").

Of the 1,468 ICD9-CM concepts which were not synonymous with MeSH terms and which were without associated expressions, 1337 were related to MeSH concepts in the Metathesaurus; the remaining 131 concepts were not. A variety of patterns were found for the use of the different types links between a single ICD9-CM concept and one or more MeSH concepts. The "Like" link did not appear; however, all other permutations of links ("Broader, "Narrower", and "Other") were present. Table 1 shows the counts for each of the patterns. The predominant pattern, found in 615 of the concepts,

Table 1 - Links between ICD9-CM and MeSH in the Metathesaurus

2,599 ICD9-CM concepts in UMLS

 1112 - synonymous with MeSH terms (1033 Main Headings and 79 Entry Terms)

 1487 - not synonymous with MeSH Terms

   19 - had an associated MeSH expression

  1337 - had links to MeSH terms

     615 - "other" links only (509 had a single link)

     301 - "broader" links only (269 had a single link)

     190 - "narrower" links only (160 had a single link)

     113 - "broader" and "other" links (91 had one of each)

      73 - "narrower" and "other" links (42 had one each)

      25 - "broader" and "narrower" links (19 had one each)

      20 - "broader", "narrower", and "other" links (9 had one each)

       0 - "like" links

   131 - had no links to MeSH terms

consisted of one to six "Other" links. In 509 of these, only one "Other" link was present. Overall, 938 of these ICD9-CM concepts were each linked to a single MeSH concept. Table 2 shows examples of patterns.

Step 3: Converting ICD9-CM to MeSH

The majority of the 2,599 ICD9-CM concepts were converted to MeSH: 39.8% were synonymous with a MeSH Main Heading, 3% were synonymous with a MeSH Entry term, 0.7% had an MeSH Associated Expression, and 51.4% were linked to one or more MeSH terms. No MeSH term was found in 5.0% of the cases. Automatic creation of expressions for those cases with multiple MeSH terms was problematic. Table 2 shows some examples of the patterns that were encountered. In some cases, the appropriate expression was an "OR'ed" list of terms; in others, it was an "AND'ed" list; in still others it was a combination of the two. In the end, we selected two methods. In each method, the "broader", "narrower" and "other" terms were assembled into a respective OR'ed lists. In one

method, these three lists were then AND'ed together. In the second method, the "other" list was ignored if there were broader and/or narrower terms available. Thus, the ICD9-CM term "Dyschromia" (from Table 2) was assigned the expression "(Pigmentation) AND (Lentigo OR Melanosis OR Vitiligo)", while the ICD9-CM "Viral Warts" was assigned two expressions: "Warts AND Condylomata Acuminata AND Epidermodysplasia Verruciformis" and "Warts AND Condylomata Acuminata". The alternative methods provide alternative translations: if information retrieval with one fails, the second is less specific. Table 3 shows examples of the expressions.

Manual translation was carried out on 1447 terms (744 MeSH-synonymous UMLS concepts, 253 non-MeSH-synonymous UMLS concepts, and 450 not in the UMLS). Analysis of the results showed arrangements of linked terms consisting of one or more primary terms with zero or more secondary terms. Primary terms were usually broader terms with the same semantic type (i.e., disease or

Table 2 - Patterns of Metathesaurus Relationships

ICD9-CM	MeSH
078.1 : Viral Warts	Warts[b]; Condylomata Acuminata[n]; Epidermodysplasia Verruciformis[o]
382.9 Otitis Media	Otitis[o]; Otitis Media[o]
423.9 Disease of Pericardium	Cardiac Tamponade[n]; Pericardial Effusion[o]; Pericarditis[o]
431. Intracerebral Hemorrhage	Cerebrovascular Disorders[b]; Cerebral Hemorrhage[b]
583. Nephritis and Nephropathy	Nephritis[n]
586. Renal Failure, Unspecified	Kidney Failure, Acute[b]; Kidney Failure, Chronic[b]; Uremia[o]
709.0 Dyschromia	Pigmentation[b]; Lentigo[n]; Melanosis[n]; Vitiligo[n]
808. Fracture OF Pelvis	[None]

[b] - Broader link; [n] - Narrower link; [o] - Other link

syndrome, or medical procedure) as the ICD9-CM term. Secondary terms acted as modifiers to the broader term or terms to restrict their meaning. Expressions were created of the form (primary OR primary...) AND (secondary OR secondary...). Examples of these can be found in Table 3.

DISCUSSION

We were able to map 30.4% of our version of ICD9-CM to the 1992 UMLS Metathesaurus, where we found 95.0% of the terms linked to MeSH in some way. Conversion was straightforward when the ICD9-CM term was synonymous with a MeSH term, when a MeSH associated expression was available or when there was link to a single MeSH term; this occurred 90.4% of the time. Thus, the content of the

in three different ways (synonyms, associated expressions and related terms). The relative appropriateness of these ways remains to be tested, but it is clear that a number of questions remain unanswered. For example: if an ICD9-CM term is synonymous with a MeSH term and has an associated expression, which should be used for retrieval? In a previous example, we noted that the ICD9-CM term "Chalazion" is synonymous with a MeSH term and has an associated MeSH expression. Searching a five-year subset of the Medline database for the former term yields 28 citations, while searching for the latter yields 13, with only 3 citations in common.

The interconcept relations in the Metathesaurus represent a form of medical knowledge that may be put to use for a variety of purposes. Such knowledge

Table 3 - Sample Translations of ICD9-CM to MeSH		
ICD9-CM	From UMLS MeSH Links	From Manual Links
078.1 Viral Warts	(Warts AND Condylomata Acuminata AND Epidermodysplasia Verruciformis*)	(Warts)
382.9 Otitis Media	(Otitis OR Otitis Media)	(Otitis Media)
423.9 Disease of Pericardium	(Cardiac Tamponade AND (Pericardial Effusion OR Pericarditis))	(Mediastinal Diseases) AND (Pericardium)
431. Intracerebral Hemorrhage	(Cerebrovascular Disorders OR Cerebral Hemorrhage)	(Cerebral Hemorrhage)
583. Nephritis and Nephropathy	Nephritis	(Nephritis OR Nephrosis)
586. Renal Failure, Unspecified	((Kidney Failure, Acute OR Kidney Failure, Chronic) AND Uremia*)	((Kidney Failure, Acute OR Kidney Failure, Chronic) AND Uremia)
808. Fracture of Pelvis	N/A	(Fractures AND Pelvic Bones)

* - Term not included if ignoring "other" terms

Metathesaurus is insufficient for 100% reliable translation. However, as the UMLS continues to evolve, the current deficiencies will be addressed. The National Library of Medicine plans to include all of the ICD9-CM terms in the 1993 version of the UMLS. Eventually, mappings to MeSH will be provided for all ICD9-CM terms that can be represented reasonably in MeSH. Approximately 2,000 such mappings (as MeSH associated expressions) will appear in the 1993 UMLS. These 2,000 codes cover approximately 80% of the ICD9-CM codes used in 5 surveyed hospitals (Humphreys, BL: personal communication).

The primary reason for converting ICD9-CM to MeSH is to carry out retrievals of information indexed by MeSH. We accomplish this conversion

has been used for automated translation [14,15]; however, the type of translation needed to search indexed databases such as Medline may be too term-specific to yield to a generalizable method. The examples of automated versus manual translation shown in Table 3 illustrate this point. The Metathesaurus contains many relationships which are appropriate as broader, narrower and other relations; however, there is no indication as to which are appropriate for translation.

The solution appears to be to use associated expressions. Consider the ICD9-CM term "Jaundice, Not of Newborn" (code 782.4). This term is linked to three "other" terms: "Jaundice", "Jaundice, Cholestatic", and "Hyperbilirubinemia". The correct expression for searching Medline would be "Jaundice

OR (EXPLODE Cholestasis) OR (EXPLODE Hyperbilirubinemia) OR (EXPLODE Jaundice, Chronic Idiopathic))" (simply exploding "Jaundice" would also include "Jaundice Neonatal"). There does not appear to be any way at present to generate such an expression from UMLS interconcept relations.

The translation of ICD9-CM terms to MeSH is useful for our Medline Button; however, it is only one example of the kinds of operations the UMLS is intended to support. The steps we detail here should be applicable to other translations (e.g., SNOMED to MeSH or ICD9 to Library of Congress Headings). The ability to add necessary ATX expression added to the UMLS and to develop algorithms to use other UMLS information for automated translation remain questions for future research.

CONCLUSION

We were able to use the current UMLS to facilitate translation of a substantial part of ICD9-CM to MeSH through use of information in the MRSO, MRATX and MRREL files. However, the value of such translations is limited when the intended purpose is citation retrieval. More reliable translation appears to depend on expansion of associated expressions, created manually by the National Library of Medicine, rather than expansion of Metathesaurus relationships.

References
1. Covell DG, Uman GC, Manning PR. Information needs in the office practice: are they being met? *Ann Int Med* 1985;103(4):596-9.
2. Guyatt GH. Evidence-based medicine. *ACP Journal Club, Ann Int Med* 1991;114(sup 2):A16.
3. Huth EJ. The underused medical literature. *Ann Int Med* 1989;110(2):99-100.
4. Powsner SM, Riely CA, Barwick KW, Morrow JS, Miller PL. Automated bibliographic retrieval based on current topics in hepatology: Hepatopix. *Comp Biomed Res* 1989;22(6):552-64.
5. Cimino C, Barnett GO. Standardizing access to computer-based medical resources. In: Miller RA, editor. *Proceedings of the Fourteenth Annual SCAMC*; 1990 Nov 4-7, Washington. Los Alamitos (CA): IEEE Computer Society Press, 1990:33-7.
6. Miller RA, Gieszczykiewicz FM, Vries JK, Cooper GF. CHARTLINE: providing bibliographic references to patient charts using the UMLS Metathesaurus knowledge sources. In: Frisse ME, editor. *Proceedings of the Sixteenth Annual SCAMC*; 1992 Nov 8-11, Baltimore. New York: McGraw-Hill, 1992:86-90.
7. National Library of Medicine. *Medical Subject Headings*. Bethesda (MD): The Library, 1992 (NTIS NLM-MED-92-01).
8. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The Medline Button. In: Frisse ME, editor. *Proceedings of the Sixteenth Annual SCAMC*; 1992 Nov 8-11, Baltimore. New York: McGraw-Hill, 1992:81-5.
9. United States National Center for Health Statistics. *International Classification of Diseases, 9th Revision, Clinical Modifications*. Washington: The Center, 1980.
10. National Library of Medicine. *UMLS Knowledge Sources - 3rd Experimental Edition*. Bethesda, MD: The Library, 1992.
11. Lindberg DAB and Humphreys BL. The UMLS knowledge sources: tools for building better user interfaces. In: Miller RA, editor. *Proceedings of the Fourteenth Annual SCAMC*; 1990 Nov 4-7, Washington. Los Alamitos (CA): IEEE Computer Society Press, 1990:121-5.
12. Tuttle MS, Sherertz D, Erlbaum M, Olson N and Nelson SJ. Implementing Meta-1: the first version of the UMLS Metathesaurus. In: Kingsland LW, editor. *Proceedings of the Thirteenth Annual SCAMC*; 1990 Nov 5-8, Washington. Washington: IEEE Computer Society Press, 1990:483-7.
13. Cimino JJ, Barrows RC, Allen B: Adapting ICD9-CM for clinical decision support (abstract). In: Musen MA, editor. *Proceedings of the Second Annual American Medical Informatics Association Educational and Research Conference*; 1992 May 7-9, Portland (OR). Bethesda (MD): The Association, 1992:34.
14. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. *MD Computing* 1990 Mar-Apr;7(2):104-9.
15. Masarie Jr FE, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. *Comp Biomed Res* 1991;24(4):379-400.