AUTOMATED TRANSLATION BETWEEN MEDICAL TERMINOLOGIES USING SEMANTIC DEFINITIONS

JAMES J. CIMINO, M.D., AND G. OCTO BARNETT, M.D.

The abundance of standard medical vocabularies, such as the National Library of Medicine's medical subject headings (MeSH terms), the International Classification of Diseases, ninth revision (ICD-9), Current Procedural Terminology, fourth edition (CPT-4), and the Systematized Nomenclature of Medicine (SNOMED) [1-3], has produced a tower of Babel, with no one vocabulary proving adequate for all purposes [4-6]. It is clear that the production of a universal medical language is an elusive goal and that such a language would be unlikely to find rapid acceptance among the speakers of each of the current medical tongues. Yet the development of a common language would be desirable. Computerized clinical record systems, for example, could convert local terminologies to a standard vocabulary for automatic medical record coding (for billing) or for searches of the medical literature.

Much work has been done to provide a way to translate free text into controlled vocabularies; a common application is found in the user interfaces of several programs that assist in medical diagnosis. RECONSIDER, for example, uses synonyms to match the clinician's input with terms used in descriptions of diseases [7]. The QMR program uses a word-stem algorithm to identify terms for diseases and manifestations [8], and DXplain uses word stems, a spelling checker, and synonyms for the same purpose [9]. Attention has also been focused on ways to convert free text into MeSH terms. MeSH itself contains tens of thousands of "entry terms" that, when entered by a user searching the MEDLINE database, serve to

ABSTRACT
Automatic translation of medical terms from one controlled vocabulary into another is essential to the integration of diverse medical informatics systems. We have developed a strategy in which medical terms are represented in a standard format that provides semantic description of the terms. We demonstrate the representational power of our method by showing that a subset of medical terms (procedures) from diverse vocabularies can be described in this manner. We assess the potential usefulness of our approach for facilitating automatic translation by finding the closest match for MeSH cardiovascular procedures with ICD-9 procedures.

[KEYWORDS: automated translation, controlled vocabularies, Unified Medical Language System, semantic nets]
RATHER THAN CHOOSING AN ARBITRARY STRING OF CHARACTERS, WE APPLY A SET OF RULES TO CONSTRUCT A FORMAL REPRESENTATION OF THE TERM.

METHODS

Background
We have developed a standard way of representing medical terms. Rather than choosing an arbitrary string of characters, we apply a set of rules to construct a formal representation of the term. The representation chosen is that of the frame, in which each high-level concept is represented by a name and a collection of properties, or slots, that denote semantic features of the term. Each property may contain a value that conveys specific information about the term. Each frame represents a class of objects or concepts (in this case, medical terms), and classes are arranged hierarchically, with superclasses at the top. A class derives its properties from its superclasses but may also have properties of its own. The class may also be related to subordinate classes, called subclasses. The class acts as a superclass in relation to these subclasses, passing properties on to them.

For example, we might have a class called "diseases" that has the property "host organisms." Two subclasses of "diseases" might be "treatable diseases" (which has the property "therapeutic agents") and "infectious diseases" (which has the two properties "infectious agents" and "vector"). By adding these properties to the class "diseases," we create the subclasses. In addition to their unique properties, the two subclasses have properties they have inherited from their superclass, "diseases" (such as "host organisms"). Some restrictions are imposed on the values that may be assigned to a property. For example, the property "infectious agents" would be restricted to members of the class "organisms.

We can create more specific classes (subclasses) in two other ways. One way is to restrict the values of a property further. For example, "parasitic diseases" is a subclass of "infectious diseases" in which the property "infectious agents" is further restricted to "parasites" (a subclass of "organisms"). Another way is to represent the intersections of one or more classes. The class "treatable parasitic diseases," for example, inherits all its properties from "parasitic diseases" and "treatable diseases."

Creating a set of rules for the representation of a term involves establishing it as an instance of a class [20]. The instance takes on the properties (slots) of the class; the values for the class may be its default values or more specific values. For example, the term "malaria" could be used as an instance of "treatable parasitic diseases"; the various properties of "malaria" would be assigned values based on the restrictions imposed by the class from which the properties were inherited (the "host organism" is "human being" and the "vector" is "anopheles mosquito").

It should be noted that the
values used in the properties of descriptions of medical terms may themselves be medical terms and may, in turn, be defined as instances of classes. By using instances and classes as values for properties, we are in effect creating non-hierarchical links between concepts of different types. These links are bidirectional, so that "malaria" points to "chloroquine" through the property "therapeutic use" and "chloroquine" points back to "malaria" through the property "therapeutic use." This arrangement, known as a semantic network, is in our opinion a powerful and flexible representation for use in inferencing systems (20–22).

Developing the Semantic Network
To study the terminology for medical and surgical procedures, we began by using terms in ICD-9 (because of its broad scope and acceptance) to identify classes and properties that could be used in a semantic network. The properties were found to occur in distinct patterns, which suggested classes by which the terms could be categorized. The entire vocabulary was reviewed to detect all classes that would be required for the description of all ICD-9 procedures.

Validation of the Semantic Definitions Format
We tested the representational power of our class description by manually creating instances of formal definitions for procedure terms from each of the major terminologies: ICD-9, MeSH, CPT-4, and SNOMED. We expected the ICD-9 terms to be well-represented by the classification scheme we had developed, because it served as an initial guide to the delineation of the classes. We were therefore fairly exhaustive in validating the classes against ICD-9, including all medical and surgical procedures involving the cardiovascular system. For the remaining three vocabularies, we restricted our validation to cardiovascular procedures. The formal descriptions of procedures from the four vocabularies were organized into a MUMPS global structure and loaded into a program (23) written in DataTree Mumps and run on a Hewlett-Packard Vectra (an AT-compatible personal computer).

Testing Automatic Translation
We examined the ability of the semantic representation to facilitate automatic translation by comparing each of the MeSH terms for cardiovascular procedures with each of the ICD-9 terms. The comparison algorithm (see below) produces an integer that is a measure of similarity between any two procedures being compared. The result is the best match (or matches, in the case of a tie) for each of the MeSH terms.

The Comparison Algorithm
The difference between the semantic definitions of two procedures can be measured by the differences in class memberships and the differences in values for properties when the descriptions have the same properties. The classes provide a simple way to compare terms: the number of classes that two terms have in common

SIXTEEN MeSH TERMS AND THE CLOSEST MATCH FOUND IN ICD-9*

<table>
<thead>
<tr>
<th>MeSH</th>
<th>ICD-9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aortography</td>
<td>Contrast aortogram (88.42)</td>
</tr>
<tr>
<td>Cerebral angiography</td>
<td>Contrast cerebral arteriogram (88.41)</td>
</tr>
<tr>
<td>Portography</td>
<td>Portal contrast phlebogram (88.64)</td>
</tr>
<tr>
<td>Fluorescein angiography</td>
<td>Eye fluorescein angiography (95.12)*</td>
</tr>
<tr>
<td>Heart function tests</td>
<td>Other cardiac function test (89.5)</td>
</tr>
<tr>
<td>Echocardiography</td>
<td>Diagnostic ultrasound—heart (88.72)</td>
</tr>
<tr>
<td>Electrocardiography</td>
<td>Electrocardiogram (85.52)</td>
</tr>
<tr>
<td>Vectorcardiography</td>
<td>Vectorcardiogram (89.53)</td>
</tr>
<tr>
<td>Polarcardiology</td>
<td>Vectorcardiogram (89.53)</td>
</tr>
<tr>
<td>Exercise test</td>
<td>Cardiac stress test not elsewhere classified (89.44)</td>
</tr>
<tr>
<td>Phonocardiography</td>
<td>Phonocardiogram (89.55)</td>
</tr>
<tr>
<td>Plethysmography</td>
<td>Plethysmogram (89.58)</td>
</tr>
<tr>
<td>Myocardial revascularization</td>
<td>Coronary vessel aneurysm repair (36.91)</td>
</tr>
<tr>
<td>Aortocoronary bypass</td>
<td>Aortocoronary bypass not otherwise specified (36.10)</td>
</tr>
<tr>
<td>Internal mammary-coronary artery</td>
<td>Single internal mammary-coronary artery bypass (36.15)</td>
</tr>
<tr>
<td>Splenorenal shunt, surgical</td>
<td>Intra-abdominal venous shunt (39.1)*</td>
</tr>
</tbody>
</table>

*Most of the matches are virtually exact; an asterisk indicates a term for which no better ICD-9 term exists.
strengthens their similarity, whereas the number by which they differ weakens it. When two procedures have the same property (because they are in the same class), the values for the properties also serve as a basis for comparison. Identical values (including the null value) provide the strongest possible similarity. In general, two terms will be similar if one is a descendant of the other, with the degree of similarity relative to their "generation gap"—the number of levels between the two terms. For example, a procedure with "left heart ventricle" as the site is closer to a procedure with "heart ventricle" as the site than it is to a procedure with "right heart ventricle" as the site.

The comparison algorithm used adds points for each superclass in which both terms are members, subtracts points for each superclass in which only one term is a member, and then subtracts the distance between values for each property that both terms have in common. The resulting score reflects the degree of similarity between the two terms and serves as a basis for ranking terms in the target vocabulary when they are compared with the source term.

RESULTS

Developing the Semantic Network
A review of the procedures in ICD-9 quickly revealed that terms can be described by two main features: the method applied (e.g., insertion, excision, destruction, replacement, or radiography) and the part of the body to which the method is applied. Under this general scheme, procedures were easily classified according to the methods used. For example, members of the class "insertion procedures" have the properties "insertion method" (e.g., intubation, cannulation, or implantation), "insertion site" (an anatomic location), and "material inserted" (a catheter or pacemaker). In all, our classification of ICD-9 procedures produced 27 subclasses.

In describing ICD-9 procedures, we used 13 general "methods" with 364 specific methods. Besides "methods," other types of terms needed to describe procedures included "anatomic sites," "body processes" ("blood pressure"), "chemicals," "diseases," "donors" ("autologous donor"), "materials," "organisms," "patient characteristics" ("newborn"), and "time descriptors" ("continuous").

Validation of the Semantic Definitions Format
The ICD-9 procedures are presented in 16 chapters; the first 13 deal with operative procedures of various organ systems, and the last three deals with all other procedures (e.g., diagnostic tests, examinations, and consultations). In order to feel confident that our method could be applied to all procedures, we manually created formal definitions of all procedures in the first 15 chapters ("Operative Procedures of the Cardiovascular System" (272 procedures) and "Miscellaneous Procedures" (845 procedures). We then applied our methods to the terms for cardiovascular procedure in MeSH (56 procedures), CPT-4 (400 procedures) and SNOMED (280 procedures). This process, though tedious, was not complicated. Each procedure was easily categorized into one of the 27 subclasses, thus showing that the semantic definition format can be used for describing procedures.

During the creation of formal definitions, many other vocabulary terms were required. There was no attempt to describe the terms semantically; they were simply arranged in a hierarchy. The generation of instances resulted in 350 anatomic sites, 110 body processes, 71 chemicals, 93 diseases, 6 donors, 157 materials, 2 organisms, 8 patient characteristics, and 9 time descriptors.

Testing Automatic Translation
We evaluated the ability of our representation of procedures to facilitate automated translation by applying the comparison algorithm to each of the 56 MeSH cardiovascular terms and each of the 1117 ICD-9 cardiovascular terms (62,552 pairwise comparisons). For each MeSH term, the most closely matching ICD-9 term (or terms in the case of a tie) was selected as the best translation. Sixteen different MeSH procedures were found to match closely to a single ICD-9 procedure (Table 1).

Another nine MeSH procedures matched to more than one ICD-9 procedure for which the ICD-9 terms are specific forms of the MeSH term and no general ICD-9 term exists. The MeSH term "heart catheterization" matched equally well to the ICD-9 terms "left heart catheterization" and "right heart catheterization"; there is no ICD-9 term for unspecified catheterization. In some cases, two or more ICD-9 terms have similar meanings that the semantic definitions are indistinguishable. For example, the MeSH term "phlebography" matched the ICD-9 terms "phlebography (88,68), "contrast phlebogram not otherwise specified (86,60)," and "contrast phlebogram not elsewhere classified (88,67)." The remaining seven cases were similar in that no one match could be considered truly superior to another.

The other 31 MeSH terms were relatively specific medical terms for which no similarly specific ICD-9 term exists and were therefore matched to one or more nonspecific ICD-9 terms. For example, there is no ICD-9 term equivalent to the MeSH term "blood circulation time"; the comparison algorithm matched it to the ICD-9 term "non-operative cardiovascular examination not elsewhere classified." Similarly, the MeSH term "diagnosis, cardiovascular" was matched to a total of 17 ICD-9 terms, ranging from "diagnostic radiology (87)" to "microscopic examination not otherwise specified (91,99)." This situation was due to the fact that ICD-9 terms for therapeutic procedures tend to be organ-specific, whereas those for diagnostic procedures are more method-specific and organ-independent. The remaining 29 cases in this category showed similar types of matches.

DISCUSSION
The compilation of semantic definitions of medical terms offers a method of translation that differs from the standard approach of assigning a specific translation for each term. We have described an approach to semantic representa-
SIXTEEN DIFFERENT MeSH PROCEDURES WERE FOUND TO MATCH CLOSELY TO A SINGLE ICD-9 PROCEDURE.

It is evident that the ICD-9 contained no appropriate term. For a true evaluation of our methods, we will need a "gold standard" for translation, so that we can determine when our system finds the correct match, when it fails, and when no correct match exists.

The exhaustive approach to translation between vocabularies (manual assignment of equivalence) requires an amount of work that increases geometrically with the number of vocabularies involved. For example, to provide for bidirectional translation among 10 vocabularies of 1,000 terms each would require 90,000 translations, performed by translators who were familiar with multiple vocabularies. Our approach would require 10,000 formal definitions of terms, each assigned by a translator with knowledge of only one vocabulary. Maintenance of the manual assignments becomes problematic when dynamic vocabularies (such as MeSH) are included. Adding a new term to a vocabulary requires translating it to each of the other vocabularies and determining whether each of the new terms is a better match than previous terms. With our approach, when a change occurs in a vocabulary, only the affected semantic description must be added, deleted, or modified; there are no hard-coded, predetermined links to be maintained.

Further efforts are needed to expand our semantic network to include other classes of medical terms, to standardize the methods used in creating formal definitions, and to improve the comparison algorithm. However, our limited study shows that the semantic description of medical terms can offer a method of automated translation among disparate medical vocabularies.

REFERENCES


JAMES J. CIMINO, M.D.

Dr. Cimino is an assistant professor of medicine and medical informatics at Columbia University. He works at the Columbia-Presbyterian Medical Center in New York, where he divides his time between internal medicine, teaching, and research. He has recently completed a National Library of Medicine fellowship in medical informatics, under Dr. Barnett at Massachusetts General Hospital, where he worked on the terminology and knowledge base of a medical diagnostic support program (DXplain) and on the Unified Medical Language System. His present research is on using the techniques of medical knowledge representation to create and maintain a controlled vocabulary for decision support.

G. OCTO BARNETT, M.D.

Dr. Barnett received a B.A. in chemistry from Vanderbilt University and his M.D. from Harvard Medical School. He is a professor of medicine at Harvard Medical School and director of the Laboratory of Computer Science at Massachusetts General Hospital. Members of this laboratory were the original developers of the MUMPS language, COSTAR (a computer-based ambulatory record system), the RxRx series of medical education programs, and DXplain (a diagnostic decision-support system). Dr. Barnett's professional interests include the use of information technology in medical education, coordinating a training program for postdoctoral fellows in medical informatics, and developing and implementing computer-based clinical decision-support systems and medical information systems.