Methods of Information in Medicine

J. J. Cimino

Formal Descriptions and Adaptive Mechanisms for Changes in Controlled Medical Vocabularies

Department of Medical Informatics, Columbia University, New York, NY, USA

Abstract: Standard controlled medical vocabularies are typically based on a coding scheme, while medical informatics applications tend to have a more formal conceptual foundation. When such applications attempt to use data coded with standard vocabularies, problems can arise when the standard vocabulary changes over time. A formal taxonomy is presented for describing the semantic changes which can occur in a vocabulary, such as simple addition, refinement, precoordination, disambiguation, redundancy, obsolescence, discovered redundancy, major name changes, minor name changes, code reuse, and changed codes. The taxonomy is described that used to effect change in one concept-based vocabulary (the Medical Entities Dictionary), and the utility of the approach is demonstrated by applying it to the changes appearing in the 1994 release of the *International Classification of Diseases, Ninth Edition, with Clinical Modifications* (ICD-9-CM).

Keywords: Controlled Medical Vocabulary, Nomenclature, Taxonomy, Electronic Patient Records, Medical Record Coding

1. Introduction

The use of standard, controlled medical vocabularies for coding patient information is a well-established procedure for U.S. health care providers. The most familiar of these vocabularies is the *International Classification of Disease, 9th Revision, with Clinical Modifications* (ICD-9-CM) [1]. The information coded is typically secondary in nature, being derived from data collected for patient care, and is used to meet government reporting requirements and to obtain reimbursement from third-party payers. However, since the data are in a controlled and, usually, computer-readable form, there is increasing interest in attempting to use such information for additional purposes such as clinical research [2], automated decision support [3], and linking clinical systems to bibliographic resources [4]. These applications require more than knowing the codes and associated text strings found in the database; they are concerned with the *meaning* of the data.

Standard vocabularies like ICD-9-CM are generally created and maintained by a central authority which is naturally resistant to adding the detail which may be relevant to those who record data with the vocabulary [5]. The users are often at the mercy of the central authority with respect to content and updates of the vocabulary. Updates are especially troublesome because data previously recorded using a particular code may be unreliable if the meaning of the code changes. The data may become completely unusable if the code is dropped from the vocabulary. Users typically cope with vocabulary updates by maintaining an historical file of the vocabulary changes, so that the meaning of a given code, on a given date, can be determined. This approach is satisfactory for purposes such as issuing bills and reporting mortality statistics. However, for purposes such as decision support and clinical research, this is less than ideal. In these applications, the task is to search the database for a particular code and extract it; if the meaning of the code has changed over time, then simply finding the code in the database is insufficient; comparison to the historical vocabulary is also required. This comparison may be prohibitive in terms of complexity and speed, particularly for applications such as automated decision support.

In contrast to most common standard medical vocabularies, the controlled vocabularies used in medical informatics applications are typically *concept-based,* in the sense that each code is associated with a particular meaning. In a concept-based vocabulary, changing the preferred name of a concept is permissible as long as the original meaning is preserved; however, if the meaning is changed, a new concept, with a new code, is created. The National Library of Medicine's Unified Medical Language System (UMLS) [6], for example, takes a concept-based approach [7]. The central clinical information system at the Columbia-Presbyterian Medical Center (CPMC) also uses a concept-based vocabulary, called

the Medical Entities Dictionary (MED) [8].

When changes occur in a standard vocabulary, they are generally reported in some *syntactic* form which conveys the surface differences but fails to represent the meaning of the changes. For example, the name of a term associated with a particular code may change, but there is no explicit statement about the *semantics* of the name change; that is, how has the meaning of the term changed, if at all? When concept-based vocabularies such as the UMLS and the MED include terms from standard sources, the persons responsible for the updating process must understand explicitly the semantics of the changes in order for proper editing to be carried out [9].

Our center has developed a taxonomy for describing the kinds of changes encountered from year to year in standard controlled medical vocabularies. Previously, we have described this taxonomy briefly and used the 1993 changes in ICD-9-CM to illustrate the different types of changes [10]. In this paper, we provide additional details about our taxonomy, examine the changes introduced in the October 1994 release of ICD-9-CM, and describe how we cope with them in the concept-based approach.

2. The Syntax and Semantics of Changes in Controlled Vocabularies

Published changes to standard vocabularies are typically presented in four different syntactic forms: new terms are *additions,* removed terms are *deletions,* there can be *name changes,* and there can be *code changes**. However, this simple classification belies the semantic complexity that may be present and its impact on applications which are concerned with the meaning of the terms. For example, the addition of a new term will have an impact on the way data are coded in the future, but it may also alter the interpretation of data which have been stored in the past. Our analysis of coded vocabularies in general, and ICD-9-CM in particular, reveal eleven different types of semantic changes, summarized in Table 1.

Table 1 Summary of syntactic and semantic changes. When a semantic change occurs, it may require a change to an existing concept, an addition of a new concept, or both. Changes to existing concepts include changing the concept name or adding and/or deleting codes associated with the concept. If a concept no longer has a valid code, it is said to be retired; however, it is never deleted. N/C indicates No Change. Refer to text for definition of the semantic changes.

Change in Source Vocabulary		Change to	Creation of
Syntax	Semantics	Existing Concept	New Concept
Addition	Simple Addition	N/C	Add
	Refinement	N/C	Add as child of existing concept
	Precoordination	N/C	Add, related to existing concepts
	Disambiguation	Change name or retire concept	Add
	Redundancy	Add code to existing concept	N/C
Deletion	Obsolescence	Retire concept	N/C
	Discovered Redundancy	N/C	Add class of redundant concepts
Name Change	Minor Name Change	Change name	N/C
	Major Name Change	Retire	Add
	Code Reuse	Retire	Add
Code Change	Code Change	Assign new code	N/C

Additions

The vocabulary of medicine will continue to grow as long as new knowledge is added to the field. In some cases, an entirely new concept, requiring a new coded term, is introduced (e.g., discovery of a new organism, description of a new disease, or development of a new technology). Since no previous term refers to this new concept (by definition), then the new term has no impact on the pre-existing terms nor any data encoded with them. We refer to this as a *simple addition.*

A more frequent occurrence is the addition of one or more terms which provide *refinement,* or a greater level of detail than was present in a previous term. For example, a general term for some disease might be introduced as a simple addition in one year and then, in a subsequent year, several refining terms are added which subclassify the disease by body site or severity. The addition of refined terms has implications for querying patient databases; before the refinement, those interested in instances of a particular concept would search for that concept. After the introduction of refining terms, the search must be done for both the original term and any terms that are refinements (this is often referred to as "retrieval by class" or "retrieval by explosion").

A related type of addition is *precoordination,* in which a new, complex term is created by combining two or more pre-existing, simpler terms. This kind of addition may be used to allow specific coding for the co-occurrence of a disease and one of its complications, or to create a code for a surgical procedure which is really a combination of several, separately coded procedures. Like refinement, the addition of precoordinated terms has implications for database queries: after the introduction of a precoordinated term, searches must be done both for the original term of interest and for any precoordinated terms that include it.

Sometimes, new terms are added to clarify or rectify the intended meaning of a pre-existing term. If an existing term is ambiguous (that is, having more than one meaning), new terms may be added, each of which expresses one of the meanings, thereby providing *disambiguation.* Once disambiguation has taken place, decisions about how to use old data must be made on an individual basis. It may be that the previous term

* Note that code change is semantically equivalent to the simultaneous deletion of the term with the old code and an addition of the term under the new code.

Table 2 A Sample of the MED hierarchy. Integers are the unique identifiers for the concepts (the MED Codes). The numbers enclosed in [ ] are the ICD-9-CM codes for the terms. Note that the top node in this tree (PNEUMOTHORAX) does not have an associated ICD-9-CM code, since the corresponding code (512) is nonterminal and therefore not usable for coding purposes. Also note that many additional ICD-9-CM terms have been collected in this tree, even thought they are in disparate places in the ICD-9-CM hierarchy. Only one strict hierarchy is shown; in fact, some of the terms have additional parents. For example, HYDROPNEUMOTHORAX is a also a child of HYDROTHORAX (not shown).

10078 PNEUMOTHORAX

- 5781 TUBERCULOUS PNEUMOTHORAX - UNSPECIFIED [011.70]
- 5782 TUBERCULOUS PNEUMOTHORAX - NO BACTERIOLOGY/HISTOLOGY EXAM [011.71]
- 5783 TUBERCULOUS PNEUMOTHORAX - BACTERIOLOGY/HISTOLOGY EXAM UNKNOWN [011.72]
- 5784 TUBERCULOUS PNEUMOTHORAX - FOUND IN SPUTUM MICROSCOPY [011.73]
- 5785 TUBERCULOUS PNEUMOTHORAX - BY CULTURE NOT MICROSCOPY [011.74]
- 5786 TUBERCULOUS PNEUMOTHORAX - BY HISTOLOGY NOT BACTERIOLOGY [011.75]
- 5787 TUBERCULOUS PNEUMOTHORAX - NON-BACTERIOLOGY/HISTOLOGY EXAM [011.76]
- 10072 HYDROPNEUMOTHORAX
- 10074 SPONTANEOUS TENSION PNEUMOTHORAX [512.0]
- 10075 SPONTANEOUS PNEUMOTHORAX, OTHER (ICD9) [512.8]
- 10076 ACUTE PNEUMOTHORAX
- 10077 CHRONIC PNEUMOTHORAX
- 16388 TRAUMATIC PNEUMOTHORAX
- 16389 TRAUMATIC PNEUMOTHORAX WITH OPEN WOUND INTO THORAX [860.1]
- 16392 PNEUMOHEMOTHORAX
- 16393 PNEUMOHEMOTHORAX WITH OPEN WOUND INTO THORAX [860.5]
- 41516 IATROGENIC PNEUMOTHORAX [512.1]

was used for one of the meanings implicitly, or it may have been used for both meanings. To take a simple example, suppose that some code 1234 was used for the term "diabetes". In a subsequent version of the vocabulary, a new code 1235 "diabetes insipidis" is added. This addition suddenly makes clear that, previously, 1234 might have been used to mean "diabetes mellitus" implicitly, or it might have been used to mean either disorder. Perhaps, since the introduction of 1235, the meaning of 1234 will always be diabetes mellitus; however, data coded with 1234 prior to the introduction of 1235 may be ambiguous.

The above types of term additions can actually affect the meanings of other terms in the vocabulary. In particular, a catch-all term that is used to code all the data which do not fit one of several existing specific terms has its meaning changed when a new specific term is added. For example, suppose a vocabulary has nine specific "Pneumonia" terms plus the term "Pneumonia, Not Elsewhere Classified (NEC)". Next, suppose that a tenth specific pneumonia term is added. The meaning of the NEC term is altered since patients now classified with the tenth pneumonia would previously have been classified with the NEC term. Similarly, if a term is ambiguous, and the addition of a new term takes on one of the possible meanings, then the original term, by default, takes on the remaining meaning (or meanings!). Sometimes, the name of the original term is altered (see Name Changes, below), but even if the name does not change from a syntactic point of view, the meaning is still altered.

The above additions are all legitimate changes in response to the need for increasing the coding capabilities of the vocabulary. We also consider one addition that is not only unnecessary but detrimental: *redundancy*. The addition of a new term which has the same meaning as a previous term can only add confusion to both the coding and retrieval processes. Coders with a choice of codes for a particular meaning may use these codes inconsistently. Those who seek to query a database may choose a code which then only retrieves a subset of the desired data. Of course, the searcher may use both codes, but only if the redundancy is a known one.

Deletions

The deletion of a term from a coded vocabulary, when that term may be used for coding and storing patient information, is problematic. Deletion may occur due to *obsolescence*. The authority controlling the vocabulary may be no longer interested in particular data, or changes in medical knowledge may make a term antiquated. The addition of refining, precoordinated and disambiguating terms can also render an existing term obsolete. Obsolescence of the term notwithstanding, the data previously coded with the term are not so easily discarded. The fact remains that a patient may have had some disease or undergone some procedure. Users of the data interested in such events will continue to be interested in them and will continue to need a method for retrieving them.

A more challenging situation in which term deletion occurs is *discovered redundancy;* that is, the deleted term has been found to duplicate some other term in the vocabulary. Although future use of the redundant term is prevented, the appropriate use of previous data requires special consideration. Suppose, for example, that two terms A and B exist in the vocabulary and that at some point they are noted to be synonymous. Having two separate codes for the same concept is redundant, and so it is decided that B should be eliminated from the vocabulary and only A will be used. Users interested in B will discover that their term is no longer present and will use A instead. They might also replace all occurrences of B within the previous data if that is practical. If not, they will at least look for both A and B when searching for previous data. Unfortunately, users of A may be unaware of the discovered redundancy and will continue to search pevious data for A only.

Name Changes

Name changes are a frequent occurrence in vocabulary updates. They are often *minor name changes*, in which the meaning of the term remains unchanged. Some changes are made to reflect a new naming convention or to correct a spelling error. They may also be made to make some previously implicit meaning more explicit to avoid confusion. The change may be a "retronym", to differentiate the previous default meaning from some new term. For example, when echocardiography became possible through transesophageal route, previous echocardiography terms took on the adjective "transthoracic", although no change in their meaning occurred.

Alternatively, *a major name change* occurs when the new name differs sufficiently from the old one to involve a change in the meaning of the term. Such changes typically occur when a term is disambiguated by the alteration of the existing term to identify a specific meaning and the addition of a new term to take up the deleted meaning. With a major name change, not only would it be inappropriate to display the old data with the new name, but it may be inappropriate to aggregate past and current data with the same code.

Sometimes, the name associated with a code changes so dramatically that it is really a case of *code reuse*. From a conceptual point of view, it is clear that the existing term has been deleted and that a new term has been added; however, from a syntactic standpoint, it appears to be not different from any other name change.

Code Changes

Although the code for a term is intended to be the identifier for that term, it is possible for a vocabulary update to include *changed codes*. Depending on how the change is reported, the code change may be explicitly stated or it may appear as two unrelated changes: a deletion and an addition. In the latter case, careful examination of the two changes will reveal that the term has simply been "moved". This might happen, for example, in a hierarchical coding scheme (such as ICD-9-CM) where the code of a term determines its position in a classification hierarchy. If a term is moved from one class of terms to another, its code would need to be changed to reflect the new location.

3. ICD-9-CM

ICD-9-CM is a strict hierarchy of over 19,000 terms which provides additions to the World Health Organisation's *International Classification of Diseases, 9th Edition* [11]. Created and maintained by the National Center for Health Statistics and the Health Care Financing Administration in the USA, ICD-9-CM is used by virtually every health-care provider and third-party payer in the USA for encoding a variety of patient data, particularly diagnoses. Top-level terms (such as "Pneumothorax") are given codes consisting of integers (such as 512 for Pneumothorax) or letter-integer combinations. Lower level codes have additional decimal digits (such as 512.0 Spontaneous Tension Pneumothorax), to a maximum of two levels in the hierarchy. Leaf-node terms in the hierarchy (including top-level terms having no descendants) are valid for use in coding patient data and are referred to as "complete codes". Non-terminal nodes in the hierarchy, which are not valid for use in coding patient data, are referred to as "incomplete terms".

ICD-9-CM is distributed by HCIA (Ann Arbor, Michigan), which provides an annual release of the entire vocabulary in a variety of printed and computer-readable formats. For this study, we have used the Codes and Full Titles files from HCIA. Accompanying documentation describes in general terms the changes reflected in the new version, but does not give explicit detail about the meaning of, or reasons for, the changes. An explicit set of syntactic changes can be determined by comparing the current file with the file from the prior year. This process readily reveals additions, deletions and changes. However, some special considerations are needed when attempting to classify the modifications that occur.

Since we are interested only in the terms used for recording patient data, we are concerned here only with complete codes. The addition of children to a previously complete code renders that term incomplete and the removal of all children from a previously incomplete code renders that term complete. The HCIA documentation considers that a term which changes from complete to incomplete is, in effect, a deletion, while a term which changes from incomplete to complete is an addition. In general, we agree with these semantics; however, in some cases we classify these changes as code changes. When a code becomes incomplete, a new complete term is usually added which has the same meaning as the pre-existing term – in effect, the addition of a right-most digit to the code of the pre-existing term so that it can continue to be used for coding patient data.* Similarly, when a term becomes complete through the deletion of previous descendant terms, we consider it to be a code change if one of the deleted terms has the same or similar name.

4. The Medical Entities Dictionary

As mentioned above, the CPMC clinical information system is coded using the CPMC Medical Entities Dictionary (MED) [8]. The MED consists of a semantic network of concepts, initially based on the Unified Medical Language System (UMLS) [12], with a directional acyclic graph to provide for multiple, co-existing hierarchies. Each concept in the MED has a preferred name and some number of slots (depending on which classes the concept is in, some concepts have as few as 10 slots and others have as many as 39) for holding information about the concept. This information may include pointers to other concepts (semantic links) and literal data, such as alternative names and codes. The concepts in the MED are used to code clinical data such as laboratory tests, medications, patient prob-

* The addition of such a term might also be considered to represent the introduction of redundancy. However, since the addition of a more specific term precludes the use of the higher-level term, we do not consider them to co-exist in the vocabulary.

lems and radiographic findings. As of this writing, the MED contains 43,319 concepts, 150 different slots (62 paired semantic slots and 88 literal-valued slots), 845,012 assignments of slots to concepts, and 801,693 slot values. In the case of concepts corresponding to ICD-9-CM terms, slots include:
ICD-9 Name – The term associated with the concept (usually the same as the concept name),
ICD-9 Code – The code or codes for ICD-9-CM terms associated with the concept, and
Old ICD-9 Code – Previous codes for terms associated with the concept; date of change is included with each.

Table 2 shows a sample of the MED hierarchy, including some concepts derived from ICD-9-CM.

The MED grows in a monotonic manner; that is, concepts are incrementally added and, once added, cannot be removed nor have their *inherent* meaning altered. Concepts may only change in ways which clarify or improve their meaning *explicitly*. For example, if a concept has the name "glucose test", it might be given the name "serum glucose test" if and only if the change reflects its true meaning. If the concept was previously used to code data which were actually serum test results, then the name change would be allowed. If, on the other hand, the concept was used to code data which could reflect either serum or plasma test results, then the name change would be invalid. In this latter case, the original concept would be left unchanged (or perhaps changed to "serum or plasma glucose tests") and two new terms ("serum glucose test" and "plasma glucose test") would be added to the MED (perhaps as descendants of the original concept).

The MED currently contains over 42,000 concepts drawn from a number of sources, including the UMLS, local departmental systems, and ICD-9-CM. As each of these sources undergoes changes, the MED must be modified to reflect those changes. On the surface, these changes consist of name changes in existing codes, term deletions, and term additions. If the MED were only used to look up terms in current vocabularies, as one might do with the UMLS when retrieving on-line information or with ICD-9-CM when filling out a coding form, simply reflecting these changes in the MED would be adequate. However, when patient data are to be encoded and stored for later retrieval and reconstitution, close attention must be paid as to how alterations affect the meaning of the terms. These changes are incorporated into the MED in a systematic way, depending on the type of change involved. The taxonomy of changes, described above, provides the basis for the mechanism by which changes in a source vocabulary are reflected in the MED (see Table 1).

Additions

New terms appearing in the source vocabulary will usually result in the addition of a new concept to the MED. If the new term is a refinement of an existing concept, it will be added as a child of that concept in the MED hierarchy. If the term is a precoordination, it will be added with specific relationships (including hierarchical) to the original "atomic" concepts. For example, when the concept "Chronic Hepatitis C with Hepatic Coma" (070.44) was added, it was included as a descendant of both the terms "Chronic Hepatitis C" (070.54) and "Viral Hepatitis with Hepatic Coma" (070.6).

Sometimes the addition of a new term has an impact on pre-existing concepts. In disambiguation, the existing term may now be recognized as having a more specific meaning, in which case its name may be changed. However, if the existing term could have had either meaning, then new concepts are added for each meaning and the existing concept is marked to indicate that it is retired from further use in recording patient data. In the case of ICD-9-CM terms, the code is removed from the ICD-9 Code slot and added to the Old ICD-9 Code slot, along with the date of change. The concept is still in the MED, so that old patient data can be interpreted, and the ICD-9-CM code for the old data can be retrieved.

When the addition of a new term is determined to be an instance of redundancy, then no new concept is added to the MED. Instead, the code for the new term is simply added to the appropriate slot in the existing concept. Thus, it is possible for one concept to have two codes from the same source vocabulary. At present, this does not occur for ICD-9-CM terms in the MED. However, the UMLS Metathesaurus, which includes incomplete ICD-9-CM terms, has many examples of this many-to-one code-to-concept relationship.

Deletions

In the case of deletions from a source vocabulary, the MED concept must be deactivated as described above. Term deletion on the basis of obsolescence is handled much the same way as disambiguation. However, when a deletion is made to correct a redundancy, the retired MED concept must be reconciled in some way with the remaining active term. This is usually accomplished by making the retired version of the concept a descendant of the active version.

Name and Code Changes

Changes to source terms are reflected in the MED differently, depending on what changes to the concepts are implied. For minor name changes, the concept name is simply changed. In the case of a major name change or code reuse, the existing concept is retired and a new concept is created. When a change occurs to the code, rather than the name, then the old code is moved from the ICD-9 Code slot to the Old ICD-9 Code slot and the new code is placed in the ICD-9 Code slot.*

5. Methods

In order to test the ability of our taxonomy to cover the different kinds of changes occurring in ICD-9-CM, and to test our ability to cope with these changes within the MED model, we compared the October 1993 and October 1994 releases of the Codes and Full Titles. The HCIA documentation provides an overview of the changes from year to year, however few details are included. Therefore, we determined the

* Note that our handling of code changes ends up being identical to the handling of its semantic equivalent: a deletion of a term and the addition of a synonymous term with a different code.

Table 3 Example of code change. The left-hand column shows ICD-9-CM codes and terms and the right-hand column shows MED codes and concepts. The first and second [ ] associated with each MED concept, correspond to the concept's values for the ICD-9 Code and Old ICD-9 Code slots, respectively. In the 1993 version of ICD-9-CM, a single term for cardiac pacemaker adjustment was present with the code V53.3 and a single MED concept (22630) corresponded to it. In 1994, V53.3 was used to subsume all types of cardiac devices, including pacemakers. The pacemaker-specific term is now assigned the code V53.31. In the MED, the only changes required were to place V53.31 in the ICD-9 Code slot and move V53.3 to the Old ICD-9 Code slot, along with the date of the move. For other concepts (V53.32 and V53.39), new MED concepts were created.

ICD-9-CM	MED
Old:	
V53.3 FITTING AND ADJUSTMENT OF CARDIAC PACEMAKER	22630 FITTING AND ADJUSTMENT OF CARDIAC PACEMAKER [V53.3] [ ]
New:	
V53.3 FITTING AND ADJUSTMENT OF CARDIAC DEVICE	
V53.31 FITTING AND ADJUSTMENT OF CARDIAC PACEMAKER	22630 FITTING AND ADJUSTMENT OF CARDIAC PACEMAKER [V53.31] [10/01/94^V53.3]
V53.32 FITTING AND ADJUSTMENT OF AUTOMATIC IMPLANTABLE CARDIAC DEFIBRILLATOR	41650 FITTING AND ADJUSTMENT OF AUTOMATIC IMPLANTABLE CARDIAC DEFIBRILLATOR [V53.32] [ ]
V53.39 FITTING AND ADJUSTMENT OF OTHER CARDIAC DEVICE	41651 FITTING AND ADJUSTMENT OF OTHER CARDIAC DEVICE [V53.39] [ ]

differences by processing the two files using the Unix "diff". The resulting file of differences was then examined. Lines dealing with changes in file header information were excluded. Since the MED does not include ICD-9-CM incomplete (nonterminal) terms, any changes involving incomplete codes were included in the analysis only when their incompleteness status changed as well (i.e., changed from a complete term to an incomplete term through refinement or became a complete term through deletion of children). After the above exclusions, the remaining lines in the diff file were examined through a manual process to determine which change type was applicable.

Additions were examined to see if they corresponded to existing concepts (redundancy) or if they were otherwise related concepts (through disambiguation, refinement or precoordination). Name changes were classified as major (including code reuse) or minor. Deletions were examined to see if they represented obsolescence or discovered redundancy. They were also compared with additions to determine if there were addition/deletion pairs that represented code changes. For each change encountered, the MED was then updated using the MED Editor [13] in the usual manner.

6. Results

The comparison of the 1993 and 1994 files showed 329 changes. Of these, three were changes to header records, 44 were changes to record fields other than name and code, five were changes to the names of incomplete code terms, two were deletions of incomplete code terms, and seven were additions of incomplete code terms. After exclusion of these 61 changes, we examined the syntactic and semantic aspects of the 268 remaining changes.

Syntactic Changes

A total of 202 new terms were added: 27 children of existing incomplete code terms, 133 children of 32 existing complete code terms (rendering them incomplete), 40 children of the seven new incomplete code terms (mentioned above), and two incomplete code terms which became complete, due to deletion of all of their children terms.

A total of 48 complete codes were deleted: seven were deleted along with their two parent incomplete code terms (mentioned above), eight were deleted such that their two parent incomplete code terms became complete code terms, one was deleted without any effect on its parent incomplete code term, and 32 were changed from complete code terms to incomplete code terms.

A total of 18 complete codes had name changes.

Semantic Changes

Semantic analysis was conducted by examining each of the 48 deletions to determine if any of the 202 additions were replacements, constituting a code change. Of the 32 terms being deleted due to refinement, three involved the addition of a term which was identical to the deleted term (Table 3), 25 involved additions which were basically replacements for the refined term but with minor name changes (Table 4) and three involved additions which were intended to replace the refined term but involved major name changes (Table 5). The remaining one term involved no additions which could serve as replacements: when the term "702.1 Seborrheic Keratosis" was refined, no new term was added to replace the now-incomplete term. Therefore, in its corresponding MED concept, the ICD-9 Code slot value was moved to the Old ICD-9 Code slot and two new MED Concepts were added to correspond to the new complete terms 702.11 and 702.19.

Of the other 16 deletions, 15 were related to the conversion of two in-

Table 4 Example of code change with minor name change. When the ICD-9-CM term Nausea *and* Vomiting was rendered "incomplete" by the addition of child terms, one of the children, Nausea *with* Vomiting was judged to have an equivalent meaning. Therefore, the name of the corresponding MED concept (22019) was changed to reflect the new ICD-9-CM name, the new code (787.01) was added and the old code (787.0) was moved to the Old ICD-9 Code slot. Note that the MED already contained concepts for the other two children, so the ICD-9-CM codes (787.02 and 787.03) were simply added to their ICD-9 Code slots.

ICD-9-CM	MED
Old:	
787.0 NAUSEA AND VOMITING	22019 NAUSEA AND VOMITING [787.0] [ ]
	21836 NAUSEA ALONE [ ] [ ]
	22020 VOMITING ALONE [ ] [ ]
New:	
787.0 NAUSEA AND VOMITING	
787.01 NAUSEA WITH VOMITING	22019 NAUSEA WITH VOMITING [787.01] [10/01/94^787.0]
787.02 NAUSEA ALONE	21836 NAUSEA ALONE [787.02] [ ]
787.03 VOMITING ALONE	22020 VOMITING ALONE [787.03] [ ]

complete code terms to complete code terms; in each of these, one of the deleted terms was effectively replaced by the converted term, albeit with minor name changes. For example, in 1994, the term "Tobacco Use Disorder" (305.1) became a complete term when its four child terms were deleted. In the MED, the ICD-9-CM codes for three of the deleted terms were moved from the ICD-9 Code slots to the Old ICD-9 Code slots. The fourth term, 305.10, actually had a meaning which corresponded to the new complete code. For this MED code (7849) a minor name change was made, the code 305.1 was added to the ICD-9 Code slot, and 305.10 was moved to the Old ICD-9 Code slot.

Thus, of the 48 deletions, 26 could be handled in the MED by changing the codes associated with existing concepts, with (21) or without (5) minor name changes (as in Tables 3 and 4), while the remaining 12 MED concepts were marked as old ICD-9-CM terms because either no new code was added that corresponded to the meaning, or the name change of the replacement term was too drastic (as in Table 5).

After removing the 28 code changes, the remaining 174 additions (examples of which are shown in Tables 3-5) were examined to determine if any of these were redundant with each other or with concepts already existing in the MED and, if true additions, what type. This analysis showed that none of the additions were redundant with each other or with ICD-9-CM terms already in the MED. However, 14 of these news terms corresponded to concepts which already existed in the MED but were derived from other sources. For these terms, the ICD-9-CM code was simply added to the concept information. For the remaining 160 terms, new concepts were added to the MED (as shown, for example, in Table 5). These terms constituted 49 simple additions (such as "759.83 Fragile X Syndrome"), 96 refinements (such as "789.03 Abdominal Pain, Right Lower Quadrant", which refines "Abdominal Pain"), nine precoordinations (such as "342.90 Hemiplegia and Hemiparesis"), and six disambiguations (such as the retronym "414.01 Coronary Atherosclerosis of Native Coronary Vessel", which makes explicit the previously implied affected vessel).

Finally, with regard to the 18 name changes, all of these were judged to be simple in nature. Examples include the change of code 440.24 from "Atherosclerosis of the Extremities with Gangrene" to "Atherosclerosis of Native

Table 5 Example of refinement with replacement with major name change. In 1994, the ICD-9-CM term Flaccid Hemiplegia (342.0) underwent two changes: its name was modified to include "Hemiparesis", and it was rendered "incomplete" through the addition of child terms. Normally, the term with the code 342.00 would be considered to be the new code for the previous term. However, in this case there was an associated name change which changed the meaning of the term. Since the meaning changed, the corresponding MED Concept was changed by moving "342.0" from the ICD-9 Code slot to the Old ICD-9 Code slot. Since no term with the same meaning was added to ICD-9-CM, the MED Concept was given no new value for its ICD-9 Code slot. Instead, entirely new concepts were created in the MED to correspond to the new children terms.

ICD-9-CM	MED
Old:	
342.0 FLACCID HEMIPLEGIA	8363 FLACCID HEMIPLEGIA [342.0] [ ]
New:	
342.0 FLACCID HEMIPLEGIA AND HEMIPARESIS	8363 FLACCID HEMIPLEGIA [ ] [10/01/94^342.0]
342.00 FLACCID HEMIPLEGIA AND HEMIPARESIS AFFECTING UNSPECIFIED SIDE	41510 FLACCID HEMIPLEGIA AND HEMIPARESIS AFFECTING UNSPECIFIED SIDE [342.00] [ ]
342.01 FLACCID HEMIPLEGIA AND HEMIPARESIS AFFECTING DOMINANT SIDE	41571 FLACCID HEMIPLEGIA AND HEMIPARESIS AFFECTING DOMINANT SIDE [342.01] [ ]
342.02 FLACCID HEMIPLEGIA AND HEMIPARESIS AFFECTING NONDOMINANT SIDE	41572 FLACCID HEMIPLEGIA AND HEMIPARESIS AFFECTING NONDOMINANT SIDE [342.02] [ ]

Arteries of Extremities with Gangrene" and the change of code 770.1 from "Massive Aspiration Syndrome of Newborn" to "Meconium Aspiration Syndrome". No instances of code re-use were found in the 1994 ICD-9-CM file.

Effort Required

The diff function took seven seconds to produce a file of changed records. The review of the results took the author three hours, and the editing of the MED an additional three hours.

7. Discussion

Many health-care institutions must deal with the disparity between the detailed clinical information they collect and store in their systems and the more general standard forms required for reporting to regulatory agencies and third-party payers. When data are to be used in patient care, the coding scheme used must be precise and accurate. Modifications such as drastic changes in names, deletion of terms and re-use of codes cannot be tolerated. Standard vocabularies used for reporting purposes are not subject to the same rigorous demands. Our response to this apparent dilemma is to consider our coding system a vocabulary of concepts, rather than terms. In our approach, the meanings of the concepts remain unchanged. As the meanings of the terms in a standard vocabulary drift from year to year, we simply allow their codes to drift from concept to concept in a corresponding way. Our procedures for reassigning codes to concepts are based on our classification of vocabulary changes: simple addition, refinement, precoordination, disambiguation, redundancy, obsolescence, discovered redundancy, major name changes, minor name changes, code re-use, and changed codes. Our results demonstrate that this classification is adequate for coping with all of the changes encountered in the 1994 ICD-9-CM.

As with many principled methods, our classification is not always easy to apply. For example, the differentiation between major and minor name changes is admittedly arbitrary, and determining when redundancy is being inadvertently introduced cannot always be done reliably. We are attempting to address such problems through the use of deeper semantic representations of terms. For example, when a term is defined in a structured way, if a name change requires a corresponding change in the structure, we can assert that the name change is major. Similarly, the structured definition of a new term can be compared to existing definitions and, if a match is found, we can consider whether the new term is redundant. These methods have been demonstrated to have practical value in well-defined domains [14]; however, the additional work required for developing the structured definitions has precluded their use with large vocabularies such as ICD-9-CM.

The traditional approach of coding patient data with codes from a standard vocabulary and maintaining an historical vocabulary file is appealing because of its simplicity and its ability to support retrieval based on the codes. However, the ability to retrieve patient data based on codes whose meanings are subject to change is of limited usefulness for clinical information systems, where concept-based retrieval is required. The MED and the CPMC clinical information systems are not unique in this orientation; most medical informatics applications dealing with coded terminology are concept-based rather than code-based. Like the MED, many of them have some relationship to code-based standard vocabularies which are outside their control. For such applications, the formal approach to classification of change may be relevant to their maintenance.

It is our hope that the work presented here may also be deemed relevant to those charged with the maintenance of standard vocabularies. Perhaps they will be inspired to adopt principled approaches to their tasks and introduce changes in ways that are sensitive to the requirements of those who collect and use longitudinal patient data. Until then, the formalism for representing vocabulary changes provides a coping mechanism which works at CPMC and may be applicable at other institutions with similar approaches to coding patient information.

8. Conclusion

The syntactic changes which occur in standard controlled medical vocabularies can have serious implications for concept-based clinical information systems. One way of addressing this issue is through a formal representation of the corresponding semantic changes coupled with explicit methods for accommodating each type of change. The approach presented in this paper demonstrates one such strategy which is being applied successfully.

Addendum

Subsequent to the study described here, the October 1995 version of ICD9-CM was released. Syntactic analysis showed the following changes: 17 deletions (including 16 changes of complete terms to incomplete terms – some with name changes), 89 new complete terms, and 59 name changes (not all of which were described in the accompanying documentation). Semantic analysis of the "deletions" showed that 13 could be handled as code changes (10 with minor name changes), and four could be handled by marking their ICD9-CM codes as "old". Semantic analysis of the additions showed that 76 were true additions (64 simple additions and 12 disambiguations) and 13 were actually replacements for the "deleted" codes (i.e., code changes). Semantic analysis of the name changes showed 55 minor name changes and four major name changes. Thus, our approach remains valid and sufficient for applying ICD9-CM updates to a concept-based vocabulary.

Acknowledgments

This work was supported in part by the IBM Corporation and the National Library of Medicine.

REFERENCES

1. United States National Center for Health Statistics. *International Classification of Diseases, Ninth Revision,* with Clinical Modifications. Washington DC, 1980.
2. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. Ann Intern Med 1993; 119: 844-50.
3. Nelson BD, Gardner RM, Hedrick G, Gould P. Decision support for concurrent utiliza-

tion review using the HELP system. JAMIA 1994, 1: 339-52.
4. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The Medline Button. In: Frisse ME, ed. *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care.* New York: McGraw-Hill 1992: 81-5.
5. Gersenovic M. The ICD family of classifications. Meth Inform Med 1995; 34: 172-5.
6. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Meth Inform Med 1993; 32: 281-91.
7. Tuttle MS, Olson NE, Campbell KE, Sherertz DD, Nelson SJ, Cole WG. Formal Properties of the Metathesaurus. In: Ozbolt JG, ed. *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care.* New York: McGraw-Hill 1994: 145-9.
8. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA 1994; 1: 35-50.
9. Campbell KE. *Distributed Development of a Logic-based Controlled Medical Terminology.* (Dissertation proposal). Stanford CA: Stanford University 1994.
10. Cimino JJ, Clayton PD. Coping with changing controlled vocabularies. In: Ozbolt JG, ed. *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care.* New York: McGraw-Hill 1994: 135-9.
11. World Health Organisation. *International Classification of Diseases Index. Manual for the International Statistical Classification of Diseases (9th ed).* Geneva, 1977.
12. McCray AT, Aronson AR, Browne AC, Rindfleisch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. Bull Med Libr Assoc 1993; 81: 184-94.
13. Cimino JJ, Hripcsak G, Johnson SB, Friedman C, Clayton PD. Prototyping a vocabulary management system in an object-oriented environment. In: Timmers T, Blum BI, eds. *Proceedings of the International Medical Informatics Association Working Conference on Software Engineering in Medical Informatics.* Amsterdam: North Holland 1990: 429-39.
14. Cimino JJ, Johnson SB, Hripcsak G, Hill CL, Clayton PD. Managing vocabulary for a centralized clinical system. In: Kaihara S, Greenes RA, eds. *Proceedings of the World Congress on Medical Informatics, Medinfo 95.* Vancouver 1995: 117-20.

Address of the author:
Dr. James J. Cimino, M.D.,
Department of Medical Informatics,
Atchley Pavilion, Room 1310,
Columbia-Presbyterian Medical Center,
New York NY 10032,
USA