Review Paper:
Coding Systems in Health Care

J. J. Cimino

Department of Medical Informatics
Columbia University College of Physicians and Surgeons
New York, NY, USA

Abstract: Computer-based patient data which are represented in a coded form have a variety of uses, including direct patient care, statistical reporting, automated decision support, and clinical research. No standard exists which supports all of these functions. Abstracting coding systems, such as ICD, CPT, DRGs and MeSH fail to provide adequate detail, forcing application developers to create their own coding schemes for systems. Some of these schemes have been put forward as possible standards, but they have not been widely accepted. This paper reviews existing schemes used for abstracting, electronic record systems, and comprehensive coding. It also discusses the remaining impediments to acceptance of standards and the current efforts to overcome them, including SNOMED, the Gabrieli Medical Nomenclature, the Read Clinical Codes, GALEN, and the Unified Medical Language System (UMLS).

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

1. Introduction

One of the challenges facing health care computing is the representation of patient data in a usable form. The typical approach is to encode the information using some standard terms taken from a controlled vocabulary. Applications such as order entry, summary reporting, automated decision support, and data aggregation for clinical research all require recording the data in standard ways [1, 2]. This need for controlled vocabulary to support clinical applications has been recognized for decades (see, for example, [3-5]). Understandably, health care providers, educators, researchers and policy makers often take for granted the existence of an appropriate standard terminology and assume that it is in routine use. In reality, the lack of a standard for representing patient data is one of the greatest impediments to medical computing today [6, 7]. The importance of patient data encoding to the medical informatics community is reflected in the recent increase in published literature on the subject. For example, in the newly established Journal of the American Medical Informatics Association, 18 of the 51 papers in the first 8 issues deal with coding of clinical data. A survey of medical informatics conference proceedings, spanning the years 1974 to 1992, showed 8.4% were primarily about coding issues [8]; in the most recent Symposium on Computer Applications in Medical Care (SCAMC), of the 182 papers, 24 dealt specifically with controlled medical vocabularies, and an additional 65 dealt with applications requiring coded patient data [9].

In this paper, I review the current state of the coding schemes with general suitability for health care applications. First, I will survey the coding schemes which are used for abstracting patient data, as is done for health statistics reporting and reimbursement. Next, I will review the controlled vocabularies which are intended to support coding of detailed patient data, as in comprehensive electronic medical records and automated decision support. I will then report on current efforts to develop comprehensive clinical coding schemes that seek to serve both purposes. Finally, I will close with a summary of the research issues which remain to be addressed.

2. Coding for Medical Record Abstraction

The coding of patient information has been carried out long before the advent of computers. This coding has always been directed at simplifying the data, converting it to a general form which is easier to manipulate. For example, while a patient may have pneumonia that may be caused by any of a variety of organisms, involve different sites in the lungs, be accompanied by any of several different symptoms, and be of varying severity, coding the patient's diagnosis as simply "bacterial pneumonia" allows it to be aggregated with other cases for statistical purposes. If finer granularity is needed, more specific terms can be added to the coding scheme (such as "gram negative bacterial pneumonia", "lobar bacterial pneumonia", and "bacterial pneumonia requiring mechanical ventilation"). A set of patient records can be classified with such codes and then retrieved when cases of certain types are needed. Because the coding represents only a simplified synopsis of information extracted from the record, this kind of coding is referred to as abstraction. Record
abstraction has been performed since the advent of formal medical records, to allow assessment of incidence of a disease, mortality of a surgical procedure or (in the era of prospective payment) costs for a hospital stay.

The archetypical coding system for medical record abstraction is the International Classification of Diseases (ICD). Other major coding schemes are usually presented in terms of their compatibility with ICD and their ability to resolve some of ICD's problems with granularity or coverage of a particular domain. ICD was first published in 1893. It has been revised at roughly 10-year intervals, first by the Statistical International Institute and later by the World Health Organization (WHO). The Ninth Edition (ICD-9) was published in 1977 [10], and the Tenth Edition (ICD-10) in 1992 [11]. The coding system consists of a "core" classification of three-digit codes which are the minimum required for reporting mortality statistics to WHO. A fourth digit (in the first decimal place) provides an additional level of detail; usually 0 to .7 are used for more specific forms of the core term, .8 is usually used for an "other" category and .9 for "unspecified". Terms are arranged in a strict hierarchy, based on the digits in the code. For example, bacterial pneumonias are classified as shown in Table 1. While ICD proper is limited to disease terminology, WHO also provides a set of expansions for different "families" of terms for medical specialty diagnoses, health status, disabilities, procedures and reasons for contact with health care providers.

The publication of ICD-9 was immediately followed by publication of criticisms regarding its inadequacy for general coding and specific specialty coverage [12-14]. In order to address these and other perceived problems with ICD-9, the United States National Center for Health Statistics published a set of "clinical modifications" to ICD-9, known as ICD-9-CM [15]. While completely compatible with ICD-9, the additions provided an additional level of detail in many places by adding a fifth digit to the code, corresponding to another level in the hierarchy (see Table 2).

Another American creation for the purpose of abstracting medical records has been the Diagnosis-Related Groups (DRGs), developed initially at Yale University for use in prospective payment in the Medicare program [16]. In this case, the coding system is an abstraction of an abstraction: it is applied to lists of ICD-9-CM codes which are themselves derived from medical records. The purpose of DRG coding is to provide a relatively small number of codes for classifying patient hospitalizations while at the same time providing some separation of cases based on severity of illness. The principle motivations for the groupings are factors which affect cost and length of stay. Thus, a medical record containing the ICD-9-CM primary diagnosis of Pneumococcal Pneumonia (481) might be coded with one of eighteen codes (see Table 3) depending on associated conditions and procedures; additional codes are possible if the pneumonia is a secondary diagnosis.

A more international response to perceived deficiencies in ICD-9 came in the form of the International Classification of Primary Care (ICPC) from the World Organization of National Colleges, Academies and Academic Associations of General Practitioners/Family Physicians (WONCA) [17]. ICPC provides seven axes of terms and a structure to combine them to represent clinical encounters. While the granularity of the terms is generally less than that of other classifications schemes (e.g., all pneumonias are coded as R81), the ability to represent the interactions of the concepts found in a medical record is much greater through the postcoordination of atomic terms (see Table 4). In postcoordination, the coding is accomplished through the use of multiple codes as needed to describe the data. So, for example, a case of bacterial pneumonia would be coded in...
Table 3  DRG codes assigned to cases of bacterial pneumonia depending on co-occurring conditions and/or procedures (mycobacterial disease is not shown except as a co-occurring condition). "Simple Pneumonia" codes are used when the primary bacterial pneumonia corresponds to ICD-9 codes 481, 482.2, 482.3 or 482.9 (refer to Tables 1 and 2) and there are only minor or no complications. The remaining ICD-9 bacterial pneumonias [482.0, 482.1, 482.2, 482.4, 482.5, 484, and various other codes such as 093.22 (refer to Table 2)] are coded as "Respiratory Disease" or "Respiratory Infection". Cases in which pneumonia is a secondary diagnosis may also be assigned other codes (such as 789), depending on the primary condition.

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>Respiratory disease w/ major chest operating room procedure, no major complication or comorbidity</td>
</tr>
<tr>
<td>76</td>
<td>Respiratory disease w/ major chest operating room procedure, minor complication or comorbidity</td>
</tr>
<tr>
<td>77</td>
<td>Respiratory disease w/ other respiratory system operating procedure, no complication or comorbidity</td>
</tr>
<tr>
<td>79</td>
<td>Respiratory infection w/ minor complication, age greater than 17</td>
</tr>
<tr>
<td>80</td>
<td>Respiratory infection w/ no minor complication, age greater than 17</td>
</tr>
<tr>
<td>81</td>
<td>Simple Pneumonia w/ minor complication, age greater than 17</td>
</tr>
<tr>
<td>82</td>
<td>Simple Pneumonia w/ no minor complication, age greater than 17</td>
</tr>
<tr>
<td>87</td>
<td>Respiratory disease w/ ventilator support</td>
</tr>
<tr>
<td>478</td>
<td>Respiratory disease w/ major chest operating room procedure and major complication or comorbidity</td>
</tr>
<tr>
<td>538</td>
<td>Respiratory disease, other respiratory system operating procedure and major complication</td>
</tr>
<tr>
<td>539</td>
<td>Respiratory infection w/ major complication or comorbidity</td>
</tr>
<tr>
<td>631</td>
<td>Respiratory infection w/ secondary diagnosis of bronchopulmonary dysplasia</td>
</tr>
<tr>
<td>740</td>
<td>Respiratory infection w/ secondary diagnosis of cystic fibrosis</td>
</tr>
<tr>
<td>770</td>
<td>Respiratory infection w/ minor complication, age not greater than 17</td>
</tr>
<tr>
<td>771</td>
<td>Respiratory infection w/ no minor complication, age not greater than 17</td>
</tr>
<tr>
<td>772</td>
<td>Simple Pneumonia w/ minor complication, age not greater than 17</td>
</tr>
<tr>
<td>773</td>
<td>Simple Pneumonia w/ no minor complication, age not greater than 17</td>
</tr>
<tr>
<td>798</td>
<td>Respiratory infection w/ primary diagnosis of tuberculosis</td>
</tr>
</tbody>
</table>

ICPC as a combination of the code R81 and the code for the particular test result which identifies the causative agent. This is in contrast to the precoordination approach, in which every type of pneumonia is assigned its own code (as in Table 1).

Professional specialty groups find that general coding schemes are of little use for their purposes and often resort to developing their own coding schemes for medical record abstraction. For example, the American Medical Association developed the Current Procedural Terminology (CPT) in 1966 [18] to provide a precoordinated coding scheme for diagnostic and therapeutic procedures which has since been adopted in the US for billing and reimbursement. Like the DRG codes, CPT codes specify information about the codes which differentiates them based on their cost. For example, there are different codes for pacemaker insertions, depending on whether the leads are “epicardial, by thoracotomy” (33200), “epicardial, by xiphoid approach” (33201), “transvenous, atrial” (33201), “transvenous, ventricular” (33207), or “transvenous, AV sequential” (33208). CPT also provides information about the reasons for a procedure. For example, there are codes for arterial punctures for “withdrawal of blood for diagnosis” (36600), “monitoring” (36620), “infusion therapy” (36640), and “occlusion therapy” (75894).

Another successful specialty coding scheme is the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, published in 1987 in its Revised Third Edition (DSM-III-R) [19]. Publication of the Fourth Edition (DSM-IV) has been coordinated with the development of psychiatric diagnoses in ICD-10 [20]. The DSM nomenclature provides definitions of the disorders including diagnostic criteria. Thus it is used not only for coding patient data but as a tool for actually assigning diagnoses. Each edition of DSM has been coordinated with corresponding editions of ICD. Compatibility between ICD-9 and DSM-III-R was found to be reasonably good [21]; a number of studies have shown that compatibility between ICD-10 and DSM-IV is variable across its different sections.

Nursing organizations have been extremely active in the development of standard coding systems for abstracting patient records. One review counted a total of 13 separate projects world-wide [22]. Two recent reports analyze the current state of these classification systems (as well as the more general

Table 4  ICPC Coding for Pneumonia. Only one of seventeen chapters (Respiratory System) is shown. Coding a clinical encounter for a patient with pneumonia entails the assignment of the code R81 as the diagnosis and including codes in any of the other six components that can be used to describe the severity and etiology of the case.

<table>
<thead>
<tr>
<th>Components</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Symptoms and complaints</td>
<td>R - Respiratory</td>
</tr>
<tr>
<td>2. Diagnostic, screening, prevention</td>
<td></td>
</tr>
<tr>
<td>3. Treatment, procedures, medication</td>
<td></td>
</tr>
<tr>
<td>4. Test results</td>
<td></td>
</tr>
<tr>
<td>5. Administrative</td>
<td></td>
</tr>
<tr>
<td>6. Other</td>
<td></td>
</tr>
<tr>
<td>7. Diagnoses, disease</td>
<td>R81</td>
</tr>
</tbody>
</table>

Respiratory Tract Diseases
Lung Diseases
Pneumonia
Bronchopneumonia
Pneumonia, Aspiration
Pneumonia, Lipid
Pneumonia, Lobar
Pneumonia, Mycoplasma
Pneumonia, Pneumocystis Carinii
Pneumonia, Rickettsial
Pneumonia, Staphylococcal
Pneumonia, Viral
Lung Diseases, Fungal
Pneumonia, Pneumocystis Carinii
Respiratory Tract Infections
Pneumonia
Pneumonia, Lobar
Pneumonia, Mycoplasma
Pneumonia, Pneumocystis Carinii
Pneumonia, Rickettsial
Pneumonia, Staphylococcal
Pneumonia, Viral
Lung Diseases, Fungal
Pneumonia, Pneumocystis Carinii

Table 5  Partial tree structure for the Medical Subject Headings (MeSH) showing pneumonia terms. Note that terms can appear in multiple locations, although they may not always refer to the same children, implying that they have somewhat different meanings in different contexts. For example, Pneumonia means "lung inflammation" in one context and "lung infection" in another context.

3. Coding for Medical Record Systems

Abstracting systems are a fact of life for medical record keeping, both for health statistics reporting and, at least for the US, for reimbursement [33]. The relevant question here is: can these systems support computer-based health care systems? When an abstract system fails at its original task (reporting causes of mortality and morbidity) [34], it should not be surprising that it is inappropriate for more strenuous tasks, such as coding a research database [35]. An even more challenging task is the coding of data in a record in a way that retains sufficient detail for a care provider to use it directly in patient care. Treatment decisions, for example, require more detail than "Pneumonia Due to Other Specified Bacteria" in order to select an appropriate antibiotic.

At the same time, coding of detailed data must consider the additional uses for the data, such as case review, summary review, decision support, research, quality assurance and, of course, reporting of mortality and morbidity.

Electronic medical record (EMR) systems typically have the greatest vocabulary requirements, assuming that the data in the record are to be encoded. In general, developers of health care applications have difficulty using existing coding systems. For example, the developers of TMR (The Medical Record) at Duke University have explicitly rejected standard vocabularies...
as inappropriate for use in an EMR [36]. They, and others, have resorted to developing their own controlled vocabularies. In some cases, they are created in an ad hoc manner, adding coded terms as needed. In other cases, developers have applied a deliberate methodology to vocabulary development.

One of the most comprehensive EMRs is the HELP System in use at the LDS Hospital in Salt Lake City, Utah [37]. The data in HELP are drawn from most of the hospital departments, cover a wide range of functional types, and are used for a variety of purposes [38]. Almost all of the data in HELP are encoded with the PTXT data dictionary. This dictionary is structured as a strict hierarchy with each term having an eight-byte code in which the first three bytes specify general information about the type of data being stored and the last five define the term’s position in the PTXT hierarchy. The system is now commercially available and the PTXT vocabulary is common across the various HELP installations; however, as of this writing PTXT has not been implemented in any other EMRs. Furthermore, while PTXT is used successfully by the on-line decision support capabilities of the HELP system, it has proven difficult to use for a diagnostic expert system developed by the same research group [39].

COSTAR (Computer-Stored Ambulatory Record) [40], developed at the Massachusetts General Hospital, also makes extensive use of a formal, albeit “home grown” controlled vocabulary called the Directory. Like PTXT, the COSTAR Directory is a strict hierarchy with a coding system (in this case, three alphabetic digits, plus a check digit and optional modifiers) which provides terms for coding a wide range of information in the record. COSTAR is available from commercial vendors, but can also be obtained in a public domain form that is available from the COSTAR Users Group. A standard Directory is supplied with the software; however, it only specifies the uppermost levels in the hierarchy. It is left to each installation site to flesh out the hierarchy with specific terms for their own institution. There has been no attempt to standardize these individual development efforts.

The Regenstrief Medical Record System (RMRS) at the University of Indiana [41] also uses a coded vocabulary for representing a portion of its data. This particular vocabulary construction task was complicated by the need to coordinate terminologies from four different hospitals. Despite the effort expended to make RMRS inter-institutional, it remains institution-dependent and has not been adopted for use in other systems.

There is one notable exception to the rule that abstracting systems have failed to support EMRs. Developed at Erasmus University in Rotterdam and now in use in a majority of private practitioners’ offices in the Netherlands, the ELIAS system makes use of the ICPC for coding diagnoses and reasons for encounters [42]. This adoption was not without cost, however. An extensive project was undertaken to translate ICPC to Dutch and to match the ICPC codes with the terms entered by users of the ELIAS system [43]. This project resulted in a greatly enhanced version of ICPC, with a significant addition of index terms and synonyms. Evaluations thus far have shown relatively good general acceptance. Similar success in other settings awaits further work to establish vocabulary standards [44].

All of the aforementioned EMRs make use of coding schemes which, while varying in their domain coverage and richness of detail, all share a fairly simple structure—that of a strict hierarchy. In some cases, synonyms are allowed and in some cases appropriate modifiers are specified. However, the depth of the representation of the vocabularies is generally shallow compared to that invested in other aspects of the systems. The approach at the University of Manchester has been quite different. In the PEN & PAD project (Practitioners Entering Notes and Practitioners Accessing Data), the vocabulary model is based on a semantic net formalism (Structured MetaKnowledge, or SMK) which allows for a variety of vocabulary-related information to be specified and allows multiple hierarchies [45]. System developers have found that the extra effort made in vocabulary development ultimately pays off in terms of the ability of the EMR to remain faithful to the description of the original patient care processes it records. The structure of the PEN & PAD vocabulary also provides the flexibility needed to support secondary uses of the data and to adapt the system for uses in a variety of patient care settings and populations.

The Medical Entities Dictionary (MED) used in the Columbia-Presbyterian clinical information system is also based on a semantic network model [46]. This vocabulary integrates terms from national coding schemes with those from local ancillary systems to produce a unified coding scheme that retains the fine granularity from the original coding schemes while accommodating the coarser granularity of a variety of applications making use of the patient data. The semantic network model is useful both for supporting the addition of new terms from ancillary systems [47] and for maintaining currency with changes in the national vocabularies [48].

### 4. Current Efforts to Develop Medical Coding Systems

The developers of each EMR have dealt with controlled vocabulary in a unique way. The results have been generally satisfactory for supporting the needs at each site; however, the ability to share the coding scheme for use at other sites has been limited, when it occurs at all. The implication is that other developers may enjoy the same successes but they will, essentially, be required to start from scratch. With several decades of experience in computer-based vocabulary requirements, researchers are now beginning to collaborate to apply their individual experiences to the task of developing general-purpose, comprehensive controlled vocabularies to support health care applications.

The first coding scheme which attempted to provide terms for a broad range of clinical domains was the Systematized Nomenclature of Medicine (SNOMED) from the College of American Pathologists. First published in 1975 and then revised as SNOMED II in 1979, it has recently been released...
Table 6  SNOMED international codes for pneumonia. The first set of terms are those from the Disease axis which are included under the Bacterial Infectious Disease hierarchy (excluding several veterinary diseases). "NOS" stands for "Not Otherwise Specified". The codes shown on the right are the SNOMED codes which, when taken together, are the equivalent of the precoordinated bacterial pneumonia terms. For example, "Pneumococcal pneumonia" (DE 13510) is the precoordination of the terms "Lung, NOS" (T-28000), "Inflammation, NOS" (M-40000), and "Streptococcus pneumoniae" (L-25116). The second set of terms shows some of the other pneumonia terms in SNOMED which could be coupled with specific Living Organism terms to allow postcoordinated coding of concepts not found explicitly in SNOMED.

<table>
<thead>
<tr>
<th>Code</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE-10000</td>
<td>Bacterial infectious disease, NOS</td>
</tr>
<tr>
<td>DE-11205</td>
<td>Pneumonia in anthrax</td>
</tr>
<tr>
<td>DE-13212</td>
<td>Pneumonia in pertussis</td>
</tr>
<tr>
<td>DE-13450</td>
<td>Pneumonic plague, NOS</td>
</tr>
<tr>
<td>DE-13431</td>
<td>Primary pneumonic plaque</td>
</tr>
<tr>
<td>DE-13432</td>
<td>Secondary pneumonic plaque</td>
</tr>
<tr>
<td>DE-13510</td>
<td>Pneumococcal pneumonia</td>
</tr>
<tr>
<td>DE-13954</td>
<td>Salmonella pneumonia</td>
</tr>
<tr>
<td>DE-14120</td>
<td>Staphylococcal pneumonia</td>
</tr>
<tr>
<td>DE-14213</td>
<td>Pneumonia due to Streptococcus</td>
</tr>
<tr>
<td>DE-14817</td>
<td>Tuberculosis pneumonia</td>
</tr>
<tr>
<td>DE-15104</td>
<td>Pneumonia in typhoid fever</td>
</tr>
<tr>
<td>DE-15613</td>
<td>Haemophilus influenzae pneumonia</td>
</tr>
<tr>
<td>DE-15716</td>
<td>Pittsburgh pneumonia</td>
</tr>
<tr>
<td>DE-15810</td>
<td>Mycoplasma pneumonia</td>
</tr>
<tr>
<td>DE-19110</td>
<td>Bacterial infection due to Klebsiella pneumonia</td>
</tr>
<tr>
<td>DE-19111</td>
<td>Pneumonia due to Klebsiella pneumonia</td>
</tr>
<tr>
<td>DE-19151</td>
<td>Pneumonia due to Pseudomonas</td>
</tr>
<tr>
<td>DE-19162</td>
<td>Pneumonia due to Proteus mirabilis</td>
</tr>
<tr>
<td>DE-19204</td>
<td>Pneumonia due to E. coli</td>
</tr>
<tr>
<td>DE-21611</td>
<td>Ornithosis with pneumonia</td>
</tr>
<tr>
<td>DE-21704</td>
<td>Pneumonia in Q fever</td>
</tr>
<tr>
<td>DE-3632A</td>
<td>AIDS with bacterial pneumonia</td>
</tr>
<tr>
<td>DE-3632B</td>
<td>AIDS with pneumococcal pneumonia</td>
</tr>
<tr>
<td>DE-3633B</td>
<td>AIDS with pneumonia, NOS</td>
</tr>
<tr>
<td>D2-50100</td>
<td>Bronchopneumonia, NOS</td>
</tr>
<tr>
<td>D2-50104</td>
<td>Parabronchial pneumonia</td>
</tr>
<tr>
<td>D2-50110</td>
<td>Hemorrhagic bronchopneumonia</td>
</tr>
<tr>
<td>D2-50120</td>
<td>Terminal bronchopneumonia</td>
</tr>
<tr>
<td>D2-50130</td>
<td>Pleuropneumonia</td>
</tr>
<tr>
<td>D2-50140</td>
<td>Pneumonia, NOS</td>
</tr>
<tr>
<td>D2-50142</td>
<td>Catarrhal pneumonia</td>
</tr>
<tr>
<td>D2-50150</td>
<td>Unresolved pneumonia</td>
</tr>
<tr>
<td>D2-50152</td>
<td>Unresolved lobar pneumonia</td>
</tr>
<tr>
<td>D2-50300</td>
<td>Aspiration pneumonia, NOS</td>
</tr>
<tr>
<td>D2-61020</td>
<td>Gangrenous pneumonia</td>
</tr>
<tr>
<td>D8-72532</td>
<td>Infective pneumonia acquired prenatally, NOS</td>
</tr>
</tbody>
</table>

in a greatly expanded version: the Systematized Nomenclature of Human and Veterinary Medicine - SNOMED International [49]. SNOMED consists of a set of axes (now eleven), each of which serve as a taxonomy for a specific set of concepts (organisms, diseases, procedures, etc.), containing a total of over 150,000 terms. Coding patient information is accomplished through the postcoordination of terms from multiple axes to represent complex terms which may be desired but do not exist in SNOMED. For example, although many of the various bacterial pneumonia terms seen in other terminologies are in SNOMED (see Table 6), additional terms can be constructed by pairing a generic pneumonia term with a bacteria term taken from the Living Organism axis.

Despite its long history and extensive efforts to provide the codes needed for coding in EMRs, SNOMED has not been widely embraced. The latest version goes a long way toward addressing past complaints about missing terms; however, the structure of previous versions, also found to be an impediment to use, has persisted in SNOMED International. The main problem with using SNOMED for coding patient information is that it is too expressive. Because there are few rules about how the postcoordination coding should be done, the same expression might end up being represented differently by different coders. For example, "acute appendicitis" can be coded as a single disease term, as a combination of a modifier ("acute") and a disease term ("appendicitis"), or as a combination of a modi-
ifier ("acute"), a morphology term ("inflammation") and a topography term ("vermiform appendix"). Each of these codings is correct, yet there is no formal way, in SNOMED, to know they have equivalent meaning. Such freedom of expression may be welcome to those who must encode human utterances, but it is frustrating to system developers who must make sure that their applications can recognize medical concepts.

One proposed solution to this redundant coding problem is the representation of the semantics of SNOMED expressions in a formal way that would allow different surface forms to be recognizable as equivalent [50]. For example, if the disease term "acute appendicitis" was formally represented as equivalent to the combination of a modifier term and a disease term, and the disease term "appendicitis" was formally represented as a combination of a morphology term and a topography term, then the three coding schemes for "acute appendicitis" would be computationally equivalent. Such equivalence would permit the development of rules for consistent coding and/or sophisticated retrieval of patient data. The SNOMED developers have embraced this approach and work is now under way to formalize the semantics in SNOMED to make it meet the needs of EMRs [51].

The Read Clinical Codes are a set of codes designed specifically for use in coding electronic medical records. Developed privately in the 1980's [52, 53], the first version was adopted by the British National Health Service in 1990. Version 2 was developed to meet the needs of hospitals for cross-mapping their data to ICD-9. Version 3 [54] was developed to support not only medical record summarization, but to support patient care applications directly. While previous versions of the Read Codes were organized in a strict hierarchy, Version 3 made an important step by allowing terms to have multiple parents in the hierarchy; that is, the hierarchy became that of a directed acyclic graph. Table 7 shows the hierarchy for bacterial pneumonia. Version 3.1 added the ability to make use of term modifiers through a set of templates for combining terms in specific, controlled ways so that both precoordination and postcoordination is used.

Finally, the NHS has undertaken a series of "terms projects" which are expanding the content of the Read codes to assure that the terms needed by practitioners are represented in the Codes [55].

At about the same time, the Gabrieli Medical Nomenclature was described in the US [56]. This system, first developed at the University of Buffalo, was adopted for use in a proprietary system. It consists of a single, large hierarchy which contains successively more complex expressions as one moves down through the hierarchy. The aim of this system is to take precoordination to the extreme, providing a code for each utterance that might be found in a medical record (see Table 8). Although initially available as a commercial product, the developers have used it as the basis for nomenclature work under the American Society for Testing and Materials (ASTM - an international standards organization based in the US) [57]. The ASTM is currently working to move this nomenclature through the standards development process.

In Europe, a consortium of universities, agencies and vendors, with funding from the Advanced Informatics in Medicine initiative (AIM), has formed the GALEN project to develop standards for representing coded patient information [58]. GALEN is developing a reference model for medical concepts using a formalism based on the SMK of PEN & PAD. The reference model is intended to allow representation of patient information in a way that is independent of the language being recorded and independent of the data model used by an EMR system. The GALEN developers are working closely with the Technical Committee on Medical Informatics (TC251) of the Comité Européen de Normalisation (CEN) to develop the content that will populate the reference model with actual terms.

A collaborative effort is currently under way between ASTM (LOINC) [59] and CEN (EUCLIDES) [60] to develop the reference model and content for the domain of laboratory test names. The standard specifies structured coded semantic information about each test, such as the substance measured and the analytical method used. Rather than establish a vocabulary for use in laboratory systems, this standard is aimed at providing a vocabulary into which local laboratory terms can be mapped for exchange with other institutions.

The Canon Group [61] has experimented with the use of conceptual graphs as a form of concept representation. Using this approach, they have experimented with collaborative vocabulary development. The development work thus far has resulted in a reference model and content for the domain of
Table 8  Bacterial pneumonia coded in the Gabrieli (ASTM) Medical Nomenclature. Sixteen descendants of Mycobacterial pneumonia not shown. Some terms appear in multiple locations (e.g., Staphylococcal Pneumonia, which has additional descendants in one context). Note that Bacterial Pneumonia and Bacteriogenic Pneumonia are not considered synonymous and have different descendants. Similarly, Streptococcal Pneumonia (4-3-3-2-1-1.1-1.3.1.2) and Streptococcal Pneumonia (4-3-2-1-1.1) are not considered synonymous. Additional bacterial pneumonias can be found elsewhere in the hierarchy, such as Listerial Pneumonia (4-3-2-2-1-1.2-1.4), Staphylococcus Aureus Pneumonia in a Granulocytic Host (4-3-3-2-1-1-1.1-1-2), its child Staphylococcus Epidermidis Pneumonia in a Granulocytic Host, and Staphylococcus Pneumonia in Children (16-10-5-7-2-14-1-3).

4-3-3-2-1-1-7-1.1 Pneumonia
4-3-3-2-1-1-1-3 Causes of Pneumonia
4-3-3-2-1-1-3-1 Bacterial Pneumonia
4-3-3-2-1-1-3-1-1 Presumed Bacterial Pneumonia
4-3-3-2-1-1-3-1-2 Streptococcal Pneumonia
4-3-3-2-1-1-3-1-3 Staphylococcus Aureus Pneumonia
4-3-3-2-1-1-3-1-4 Staphylococcus Pyogenes Pneumonia
4-3-3-2-1-1-3-1-5 Neisseria Meningitidis Pneumonia
4-3-3-2-1-1-3-1-6 Branhamella Catarrhalis Pneumonia
4-3-3-2-1-1-3-1-7 Hemophilus Influenzae Pneumonia
4-3-3-2-1-1-3-1-8 Klebsiella Pneumonia
4-3-3-2-1-1-3-1-9 Escherichia Coli Pneumonia
4-3-3-2-1-1-3-1-10 Serratia Species Pneumonia
4-3-3-2-1-1-3-1-11 Enterobacteria Species Pneumonia
4-3-3-2-1-1-3-1-12 Proteus Species Pneumonia
4-3-3-2-1-1-3-1-13 Pseudomonas Aeruginosa Pneumonia
4-3-3-2-1-1-3-1-14 Pseudomonas Caplain Pneumonia
4-3-3-2-1-1-3-1-15 Pseudomonas Multiphila Pneumonia
4-3-3-2-1-1-3-1-16 Pseudomonas Pseudaetaalgens Pneun
4-3-3-2-1-1-3-1-17 Actinobacter Species Pneumonia
4-3-3-2-1-1-3-1-18 Legionella Species Pneumonia
4-3-3-2-1-1-3-1-19 Anaerobic Microbial Pneumonia
4-3-3-2-1-1-3-1-19-1 Fusobacterium Species Pneumonia

4-3-3-2-1-1-7-1.1-1-12  Bacteroides Species Pneumonia
4-3-3-2-1-1-1-3-1-1-9-3 Peptostreptococcus Species Pneumonia
4-3-3-2-1-1-1-3-1-1-9-4 Micrococcophilus Streptococcus Pneumonia
4-3-3-2-1-1-1-3-1-1-9-5 Actinomyces Pneumonia
4-3-3-2-1-1-1-3-1-1-9-6 Nocardia Species Pneumonia
4-3-3-2-1-1-1-3-1-1-9-7 Mycoplasma Pneumonia
4-3-3-2-1-1-1-3-1-1-9-8 Coxiella Burnetii Pneumonia
4-3-3-2-1-1-1-3-1-1-9-9 Chlamydia Pneumonia
4-3-3-2-1-1-1-3-1-1-9-10 Chlamydia Trachomatis Pneumonia
4-3-3-2-1-1-1-3-1-1-9-11 Pneumococcal Pneumonia
4-3-3-2-1-1-1-3-1-1-9-12 Pneumococcal Pneumonia
4-3-3-2-1-1-1-3-1-1-9-13 Pneumococcal Pneumonia
4-3-3-2-1-1-1-3-1-1-9-14 Pneumococcal Pneumonia
4-3-3-2-1-1-1-3-1-1-9-15 Pneumococcal Pneumonia
4-3-3-2-1-1-1-3-1-1-9-16 Pneumococcal Pneumonia
4-3-3-2-1-1-1-3-1-1-9-17 Pneumococcal Pneumonia
4-3-3-2-1-1-1-3-1-1-9-18 Pneumococcal Pneumonia
4-3-3-2-1-1-1-3-1-1-9-19 Pneumococcal Pneumonia

chest radiograph reports which can serve a variety of purposes, including natural language processing, predictive data entry and automated decision support [62].

For some time, the NLM has been developing the Unified Medical Language System (UMLS) [63] to serve a number of controlled vocabulary needs [64]. Included in the UMLS is the Metathesaurus, which contains concepts, and the UMLS Semantic Net, which provides information about how the semantic classes of concepts can be interrelated. The concepts in the Metathesaurus are drawn from established controlled vocabularies, such as MeSH, ICD-9-CM, and SNOMED. Information about each concept includes the preferred form of the concept in the various source vocabularies, synonyms and lexical variants of the concepts, and information about relationships between specific concepts (Tables 9 and 10). Various uses for the UMLS have been described, including the coding of patient data. However, the NLM has acknowledged that the UMLS does not serve clinical encoding well. This is largely due to the fact that the source vocabularies do not themselves serve this function. The NLM is now developing ways in which the UMLS can be enhanced to support the coding of clinical data and has enlisted the help of a large number of researchers (including most of the Canon Group) to provide input for and evaluation of this UMLS expansion.

Finally, vocabulary servers have become a research issue in their own right. The servers are intended to provide open, distributed health care systems with information about up-to-date vocabulary content. Groups working on vocabulary servers include GALEN [65], the NLM [66], the University of Utah [67], and Stanford University [68].

Table 9  Pneumonia concepts in the Unified Language Systems (UMLS) Metathesaurus.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacterial pneumonia</td>
<td></td>
</tr>
<tr>
<td>Pneumonia, Lobar</td>
<td></td>
</tr>
<tr>
<td>Pneumonia, Staphylococcal</td>
<td></td>
</tr>
<tr>
<td>Pneumonia, Streptococcal</td>
<td></td>
</tr>
<tr>
<td>Pneumonia due to Streptococcus</td>
<td></td>
</tr>
<tr>
<td>Pneumonia in antrax</td>
<td></td>
</tr>
<tr>
<td>Pneumonia, antrax</td>
<td></td>
</tr>
<tr>
<td>Bronchopneumonia</td>
<td></td>
</tr>
<tr>
<td>Pasteurellosis, Pneumonic</td>
<td></td>
</tr>
<tr>
<td>Salmonella Pneumonia</td>
<td></td>
</tr>
<tr>
<td>Other bacterial Pneumonia</td>
<td></td>
</tr>
<tr>
<td>Pneumonia due to Klebsiella Pneumonia</td>
<td></td>
</tr>
<tr>
<td>Pneumonia due to other specified bacteria</td>
<td></td>
</tr>
<tr>
<td>Pneumonia in whooping cough</td>
<td></td>
</tr>
<tr>
<td>Pneumonia due to Pseudomonas</td>
<td></td>
</tr>
<tr>
<td>Pneumonia due to Hemophilus influenzae (H. influenzae)</td>
<td></td>
</tr>
</tbody>
</table>
5. Research Issues

The preceding discussions of standard codes for abstraction, codes for electronic medical records, and current research efforts supports my opening statement that no accepted standard exists for coding patient information. In the past, the tendency for developers to create their own coding schemes, rather than adopt an existing one, may have been due to the "Not Invented Here" phenomenon. However, as systems have become larger and begin to address more comprehensive domains (such as EMRs), developers are quite willing to take advantage of existing standards. Their continued inability to do so points to failure on the part of controlled vocabularies to meet their needs. The developers of the vocabularies, on the other hand, have continued to be surprised at this resistance to use. The source of the problem is that the vocabularies are created for specific purposes and often have characteristics which limit their usefulness for other purposes. The standards developers are investing considerable effort to address this problem. In the process, a variety of issues have come to light. Some of these are internal issues, dealing with the structure and content of the vocabularies themselves, and others are external issues, dealing with the relationships between vocabulary developers and vocabulary users.

Internal Issues

The first basis on which vocabularies are judged is their content. A user cannot adopt a coding scheme if it does not have the ability to express the necessary concepts. The vocabulary domains are moving targets as medical knowledge grows with new terms to add and old ones to discard. The developers of the comprehensive vocabularies devote substantial effort into expanding their content. This usually involves development of committees and interaction with professional specialty groups to provide input. As a result, the large vocabularies being built today seem to be coming close to having the content needed.

One place where vocabularies have run into trouble has been the codes they use to represent terms. In many cases, the codes are designed to reflect the position of the term in the hierarchy. There is a certain elegance to this approach; however, in the real world of medical terminology, this elegance breaks down. If the code has a limited number of positions or digits, then the depth of the hierarchy is limited. If the positions in the code are limited to a fixed number of characters, then the breadth of the hierarchy is limited. These limitations can adversely affect vocabulary content, since some domains become too full to allow additional terms, requiring the use of catch-all "Other" terms. In addition, multiple hierarchies (see below) cannot be accommodated with a single code.

Vocabulary developers are addressing the coding issue by diverging the unique identifiers for the terms from their hierarchical positions. Among the comprehensive coding systems, only SNOMED continues to use a hierarchy-based unique identifier. The remainder either provide hierarchical information as semantic links or they allow free addresses which can be of arbitrary length and breadth.

A related issue is the need for medical terms to be organized in multiple classes. If a vocabulary permits only a single hierarchy, it will invariably be the one that meets the developer's view of the world. When this view differs from the user's view, the user may look elsewhere for a coding scheme. For example, users may wish to be able to access patient diagnoses based on location or on etiology. This becomes awkward when the user, for example, wants to identify all patients with bacterial pneumonia but the coding scheme selects the codes as in ICD, with some in the Pneumonia class, and others in the various bacterial disease classes.

Most vocabulary developers have recognized the need to accommodate multiple classes and allow them. This has been simplified by the departure from the use of hierarchical codes. In systems such as Read, GALEN and UMLS, hierarchies are represented as links between parents and children, so multiple hierarchies are simply the result of multiple links. In systems which use tree addresses, such as MeSH and the Gabrieli Nomenclature, the solution is simply to allow terms to have multiple tree addresses. Still to be resolved are the issues of variation of meaning and variation of children across different hierarchical addresses for the same term.

Researchers are realizing, though, that allowing multiple classification was the easy part. As the structures of the vocabularies become more powerful and complex, the task of where to place a term becomes as important as what term to place [69]. New techniques are being explored by several groups to take advantage of the semantic information included about the terms, either as frames, semantic nets, or conceptual graphs. One of these techniques is automated term subsumption, long used in artificial intelligence research, in which the attributes of the term define its location. For example, if the ICD-9-CM term "Salmonella pneumonia" included attributes that identify it as being caused by Salmonella and occurring in the lung, it might be possible to automatically assign it as a child of both of the desired parents.

A continuing controversy in vocabulary development revolved around the choice between precoordination and postcoordination. On one hand, a precoordinated term like "Salmonella pneumonia" is probably a useful concept and more natural than the combination "Salmonella" + "Pneumonia". On the other hand, precoordination can easily lead to combinatorial explosion as all permutations of all modifiers are appended to terms in order to have a preassigned code for the composite. Attempting to choose one or the other approach is probably not feasible. Terms which seem reasonably atomic to one user of the vocabulary will seem to some other user to be a coordination of smaller concepts. Precoordinated terms will often be found to be missing some minute detail, requiring the addition of a modifier, turning it into a postcoordination. The reality is that vocabularies which do not allow postcoordination are usually too limiting, while those that allow postcoordination always have a healthy collection of precoordinated terms. The use of conceptual graphs, as described in the appendixes in SNOMED, may accommodate both approaches while allowing equivalence between a precoordinated term and a postcoordinated phrase to be recognized.

External Issues

Once vocabularies are created, continuity needs to be maintained. Besides the issues related to how to include new terms (described above), there are epistemologic issues related to identifying new terms for inclusion and marking old ones for deletion. Monitoring usage of terms, such as is done by the National Library of Medicine for MeSH [70] will be important for determining what users need. Changes will include the addition of new terms, the addition of new classes or aggregations of terms, the addition of an existing term to an existing class, identification of a particular type of (semantic) relationship between two terms, and the addition of entirely new types of relationships.

The development of mechanisms for responding to needs for additions will be crucial for the success of any controlled vocabulary, since the lack of necessary terms in a standard coding scheme will merely push system developers to create their own coded terminologies. Any vocabulary that is interested in meeting user needs would do well to follow the lead of the NLM, which requests UMLS users to submit suggestions for changes via electronic mail [71].

An important part of maintaining a vocabulary is the communication of changes to the users. The traditional method has been to convene a committee of experts periodically to review the current version of a vocabulary and prescribe changes. This approach seems to result in updates measured in years and decades. However, for many applications, this is inadequate. For example, if a new drug goes on the market, or a new test can be ordered from the laboratory, waiting a year is too long if the new term is encountered and needs to be coded immediately. Users need to get changes as soon as they are available. This issue is being addressed in the various projects to develop vocabulary servers. Such servers will facilitate the dissemination of changes from the central authority and also provide a link back to the authority to recommend changes when they are seen, rather than waiting for the next standards-setting group to meet.

6. Conclusion

The application of computers to medicine has accelerated the breadth of uses and depth of detail needed for the representation of patient data. Legacy abstracting systems were recognized as inadequate for applications such as electronic medical records and automated decision support, but simply expanding their content has not solved the problem. Today, research into medical data representation is livelier than ever, as formal computer science techniques are being applied to large, real-world domains. Local solutions have shown great promise for the application builders who have had the resources needed for vocabulary development. For those who do not have such resources, current efforts to develop thoughtful solutions at national and international levels are under way.
Addendum

This paper is adapted from a paper presented in the 1995 Edition of the 1995 IMA Yearbook of Medical Informatics [72]. Since that publication, a paper has been published presenting the results of an evaluation of content coverage of ICD-9-CM, ICD-10, CPT, SNOMED III, Read Version 2, and UMLS Version 1.3 [73]. The study concluded that these vocabularies are currently incomplete in their coverage of the content of patient records.

Acknowledgments

I thank my collaborators on the InterMed project, sponsored by the National Library of Medicine, and the members of the Cancer group for their stimulating discussions on vocabulary issues. I also thank Leslie Jaeseen and George Hripcsak for their contributions to the editing of the manuscript.

REFERENCES

45. Heathfield HA, Hardiker N, Kirby J, Tullis R, Gonsalkaran M. The PEn & FAD medi-


Address of the author:
James J. Cimino, M.D.,
Department of Medical Informatics,
Arts and Sciences Pavilion 13190,
Columbia-Presbyterian Medical Center,
New York, NY 10032,
USA
E-mail: James.Cimino@columbia.edu