Review ■

From Data to Knowledge through Concept-oriented Terminologies:

Experience with the Medical Entities Dictionary

JAMES J. CIMINO, MD

Abstract Knowledge representation involves enumeration of conceptual symbols and arrangement of these symbols into some meaningful structure. Medical knowledge representation has traditionally focused more on the structure than the symbols. Several significant efforts are under way, at local, national, and international levels, to address the representation of the symbols though the creation of high-quality terminologies that are themselves knowledge based. This paper reviews these efforts, including the Medical Entities Dictionary (MED) in use at Columbia University and the New York Presbyterian Hospital. A decade's experience with the MED is summarized to serve as a proof-of-concept that knowledge-based terminologies can support the use of coded patient data for a variety of knowledge-based activities, including the improved understanding of patient data, the access of information sources relevant to specific patient care problems, the application of expert systems directly to the care of patients, and the discovery of new medical knowledge. The terminological knowledge in the MED has also been used successfully to support clinical application development and maintenance, including that of the MED itself. On the basis of this experience, current efforts to create standard knowledge-based terminologies appear to be justified.

■ **J Am Med Inform Assoc.** 2000;7:288–297.

The first step on the path to knowledge is getting things by their right names. —CHINESE SAYING

A basic tenet of medical informatics is that if computers are to help us with the process of health care, they must be able to manipulate information symbolically rather than simply store and regurgitate it. If we can represent data about the patient and knowledge about health care appropriately, our computer systems can accomplish many tasks that will enhance our ability to care for specific patients and learn more about biomedicine in general.

One approach to such representation is *knowledge representation*, a collection of techniques drawn from computer science. There are many such techniques, but they all share a common approach of using symbols (usually represented with terms from a controlled terminology) and structures for arranging the symbols. In this paper, I review some of these techniques and examine how medical informaticians are applying them to the task of representing knowledge about the symbols (that is, the terminologies) themselves. I illustrate the advantages of this approach with examples drawn from the work of my colleagues and myself at Columbia University, to show how a knowledge-based terminology can help us get raw patient data "by the right names" and set us on the path to knowledge, to:

- Gain a better understanding of our patients
- Access knowledge relevant to the care of our patients
- Enable the use of smart systems to apply knowl-

Affiliation of the author: Columbia University College of Physicians and Surgeons, New York, New York.

This paper is based on a presentation by Dr. Cimino that was part of the Cornerstone on Representing Knowledge, one of four Cornerstone sessions included in the program of the AMIA Annual Symposium, Washington, D.C., Nov. 6–8, 1999.

Correspondence and reprints: James J. Cimino, MD, Columbia-Presbyterian Medical Center, Atchley Pavilion, Room 1310, 161 Fort Washington Avenue, New York, NY 10032; e-mail: ⟨ciminoj@flux.cpmc.columbia.edu⟩.

Received for publication: 11/15/99; accepted for publication: 11/18/99.

edge to the care of our patients

- Discover new knowledge from health data

Such knowledge can also be used, it turns out, to manage complex clinical applications, including the maintenance of the terminological knowledge itself.*

Knowledge Representation in Medicine

Representation of medical knowledge was one of the first tasks addressed at the advent of medical informatics, starting with Ledley and Lusted's landmark paper[2] describing the use of punch cards for indicating relationships between diseases and their manifestations. Since then, informaticians have drawn on computer science for a variety of techniques. Occasionally the influence has flowed in the opposite direction, as with Shortliffe and colleagues' MYCIN project.[3] A full review of knowledge representation methods is beyond the scope of this paper; however, one comparative study will serve to illustrate some of the general principles.

Starren and Xie[4] examined a guideline for cholesterol management and represented it using three different formalisms: PROLOG (a first-order logic-based system), CLASSIC (a frame-based system), and CLIPS (a production rule-based system). The authors concluded that "all three representations proved adequate for encoding the guideline." Despite the differences in notation, the underlying symbols used in the schemes were essentially the same. This suggests that while the structure chosen for representing knowledge may be important for practical considerations such as execution efficiency, the real heart of the knowledge lies in the symbols themselves. In fact, van der Lei and Musen[5] have argued that typical rule-based systems do not encode true medical knowledge.

Knowledge-based Terminology Representation

Knowledge-based systems, and medical computing systems in general, have traditionally treated the coded terms they use as simple placeholders for concepts that are understood by the users of the systems but not by the systems themselves.[6,7] As systems have become more sophisticated, their terminology needs have grown. At first, it was sufficient to turn to large, existing, standard terminologies to avoid the need to create them for each application. These terminologies offered little in the way for formal representation, beyond simple strict hierarchies. Eventually, these schemes were found to be inadequate, and informatics researchers began seeking ways to use knowledge to represent the terminologies themselves in order to support better comprehension, use, and maintenance.[1]

Like other knowledge representation tasks, the choice of formats for terminological knowledge differed from application to application. My colleagues and I[1,8] chose a frame-based representation for terminology, as did Masarie et al.[9] Bernauer[10,11] used an object-oriented approach expressed with conceptual graphs. These two approaches, shown in Figure 1, and their variations have become the most widely used representation schemes.

Over the past decade, knowledge-based representation of terminologies has accelerated. These techniques have been applied to existing terminologies in order to make them more understandable and, hence, usable. Campbell and Musen[12] demonstrated that the Systematized Nomenclature of Medicine (SNOMED) could be represented using conceptual graphs in a way that offered the potential for more consistent SNOMED coding. This theoretic approach has been applied to a large project to expand SNOMED content with the Convergent Medical Terminology (CMT) project between Kaiser Permanente and the Mayo Clinic.[13] More recently, Spackman et al.[14] have described significant efforts by the College of American Pathologists to represent SNOMED terms with logic-based descriptions. Bakken et al.[15] have used a similar approach to represent several nursing terminologies.

In contrast, some researchers have addressed the knowledge representation issue *before* creating actual content. Rector et al.[16] have undertaken the GALEN project to provide a foundation for representing terminologies that can span the multiple languages encompassed by the European Community. Using a representation language called GRAIL, they have developed a coding reference (CORE) model for defining ways of assembling medical terms. A number of experiments are under way to test the usefulness of their formalisms. For example, Brown et al.[17] have described the efforts of the National Health Service in the United Kingdom to represent definitional knowledge of the Read Codes using the GALEN model. Hardiker and Rector[18] have also used GRAIL to represent terms from nursing terminologies.

As new, special-purpose terminologies have arisen, their creators have begun turning to knowledge-based

*Knowledge-based systems typically reason about some part of the world outside their knowledge base but not about the information contained in their knowledge bases; that is, they are usually not introspective.[1]

```
Serum Glucose Test
        is-a:              Laboratory Test
        has-specimen:      Serum Specimen
        measures:          Glucose

[Serum Glucose Test] –
        (is-a)  -> [Laboratory Test]
        (measures) -> [Glucose]
        (specimen) -> [Serum]
```

Figure 1 Two representations of the medical concept "Serum Glucose Test," using frame-based (*top*) and conceptual graph (*bottom*) formalisms. In each case, the other terms ("Laboratory Test," "Serum Specimen," and "Glucose") are also controlled terms represented with their own knowledge.

representations. The Logical Observations, Identifiers, Names, and Codes (LOINC) project, described by Huff et al.,[19] started with a formal representation of the definitions of laboratory tests, using an approach that is similar to (but much richer than) the examples given in Figure 1. This approach has facilitated the adoption and use of LOINC across multiple institutions.[20,21] In the domain of drug terminologies, used by pharmacy systems, commercial efforts have focused on representing knowledge about pharmaceutical products that includes definitional information about ingredients and formulation (T. McNamara, C. Broverman, K. Eckert, M. Moore: personal communications, 1998, 1999).

The creation of terminological knowledge bases has led to development of knowledge-based tools for supporting their development and use. A vocabulary server called VOSER has been described by Rocha et al.[22] for use at the LDS Hospital, Rector et al.[23] have described the GALEN server, and Chute et al.[24] have recently enumerated the minimum desirable characteristics for vocabulary servers. Knowledge-based editing tools have been developed for terminology construction by Mays et al.[25] and have been adapted as part of the Gálapagos tools set by Campbell et al.[13] for use on the CMT (convergent medical terminology) project.

No description of terminological efforts would be complete without inclusion of the Unified Medical Language System (UMLS). Originally envisioned as a way "to improve the ability of computer programs to 'understand' the biomedical meaning in user inquiries and to use this understanding to retrieve and integrate relevant machine-readable information for users,"[26] it has initially been focused on the more modest goal of supporting "the development of user-friendly systems that can effectively retrieve and integrate relevant information from disparate machine-readable sources."[27] The UMLS provides a knowledge base not about the meanings of terms, per se, but about models used by existing source terminologies and how they relate to one another. So, for example, the information the UMLS provides about a laboratory test term will include which source terminologies it comes from, which terms it is related to in the hierarchies of those source terminologies, what its synonyms and lexical forms are, and which other terms it is related to in some source terminology. It does not, however, strive to provide definitional information (such as what the test measures are or what its specimen is) unless that information is available from a source terminology.

The Medical Entities Dictionary

The knowledge-based terminology effort at Columbia University and the New York Presbyterian Hospital[1] has grown into a repository called the Medical Entities Dictionary (MED).[28] It currently contains some 60,000 concepts organized into a semantic network of frame-based term descriptions. Terms are drawn from those used in laboratory, pharmacy, radiology, and billing systems. It includes 208,000 synonyms, 84,000 hierarchic relations, 114,000 other semantic relations, and 66,000 mappings to other terminologies, includ-

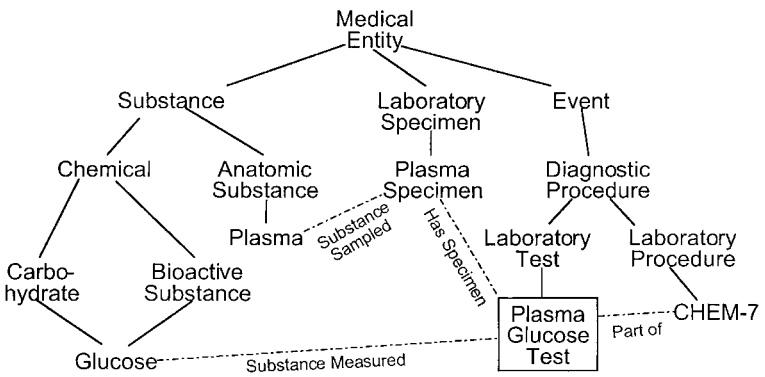


Figure 2 Example from the Medical Entities Dictionary. The term in the box (Plasma Glucose Test) is shown in relation to its parents in the is-a hierarchy (solid lines) and by nonhierarchic semantic links (broken lines) to other terms in the network.

ing the UMLS and LOINC. The relationships in the network provide definitional knowledge about the individual terms. For example, laboratory test terms are linked (via ''substance-measured'' relationships) to chemicals they measure, medication terms are linked (via ''has-ingredient'' relationships) to their chemical ingredients, and diseases terms are related (via ''has-location'' relationships) to their body locations. Figure 2 provides some examples of this knowledge.

The MED was constructed to serve the primary purpose of a repository for codes and terms used by clinical applications to represent data in the clinical data repository.[29] The knowledge included in the MED was originally intended to support intelligent vocabulary management tools. However, as the repository grew and the data in it were reused in a variety of ways, the MED knowledge was reused as well. In many cases, the MED served as a convenient repository for additional knowledge used by various applications, and so it grew to serve as a tight link between clinical applications and the terminologies used by them.

Application of Knowledge–based Terminology: Proof of Concepts

Over the years, application developers, researchers, and students have shown great creativity in exploiting the MED model and its content. For this paper, I have collected their work and attempted to summarize the kinds of roles the MED has played in bridging between the data encoded with its terms and the advancement of some aspects of human knowledge. Much of this work is anecdotal, so far as the MED is concerned; there are undoubtedly other terminological models that could have supported the various projects described here. However, taken in aggregate, I believe they provide substantial evidence that knowledge-based terminologies have great potential for furthering the goals of medical informatics.

Merging Data and Application Knowledge

Knowledge about the operation of clinical applications may be stated in written documentation, but is only occasionally described using formal modeling tools. Although the MED was not intended for application modeling, developers of the Decision-supported Outpatient Practice (DOP) application found it useful to include the various laboratory data spreadsheets as concepts in the MED.[30] Because each spreadsheet was linked in the MED to the appropriate test classes (each of which corresponded to a row in the spreadsheet), DOP was able to display test results dynamically, such that the addition of new tests and spreadsheets to the MED could be handled without modification to the program. When a new Web-based application (called WebCIS) was developed to replace DOP, the same knowledge in the MED was reused to support display of laboratory test results.[31] Figure 3 shows sample displays from both applications.

While the use of the MED knowledge was automated and dynamic, its maintenance was manual and tedious. Elhanan,[32] who was charged with this duty, viewed the task as a knowledge engineering problem and sought to find ways to use the knowledge to support the acquisition of new knowledge about the problem domain (i.e., the relations between test terms and spreadsheets). The result was an expert system that could be used to automatically audit the application knowledge in the MED, support its maintenance, and ultimately drive the performance of the clinical applications. It would, for example, identify tests that could not be displayed by any spreadsheet and make suggestions about how to link them to existing spreadsheets.[32]

Smarter Retrievals from the Record

Specific knowledge about patients is crucial to their care. Although the aggregation of data in the clinical record holds much of this knowledge, the amount and organization of the data can render the knowledge obscure. Because the MED contains knowledge about how data are coded in the record, Zeng[33] sought to supplement the MED with knowledge about how these data might be aggregated into concept-oriented views of the medical record—for example, with respect to a particular patient problem. She was able to extract information from other existing knowledge bases and reuse it in the MED. From this information, she was able to generate queries automatically to extract problem-specific data from the record. She was then able to assemble them into views that were demonstrably better than the more traditional time-oriented views. For example, if a user specifies interest in the problem ''Pulmonary Heart Disease,'' the application will identify test results that measure relevant substances (such as oxygen and carbon dioxide), reports on examinations of relevant body parts (such as cardiograms and chest x-rays), and medication orders (Figure 4) that are relevant to the treatment of the condition.

"Just-in-Time" Education

When an information need arises during the course of caring for a patient, an opportunity arises to supply specific knowledge to meet that need and, in the process, educate the clinician. Referred to as ''just-in-time

Chem-20	C7 + 22:50 24 Mar 99	Misc.Che+ 10:00 25 Jan 99	Misc.Chem 23:30 24 Jan 99	C7 + 17:15 24 Jan 99	C7 13:35 14 Oct 97	C7 9:15 12 Apr 97	C20 + 18:20 10 Apr 97	C7 10:20 07 Apr 97	C7 10:10 06 Apr 97	C7 10:00 05 Apr 97
Na	138	138		140	142	141		139	137	141
K	4.0	4.0		4.4	4.6	3.7		4.0	3.8	3.7
Cl	107	105		107	108	114		108	107	107
BUN	16	11		11	14	17		18	13	11
Creat	1.7	1.2		1.3	1.3	1.1		1.0	1.0	1.0
Gluc	133	87		94	85	76		82	88	87
Ca	8.0						8.0			
Phos	2.8									
Chol				105 *						
Alb				3.6						
TBili				0.9						
DBili				0.2						
Tot Alk P				70						
AST				47						
ALT				22						
CK		165	155	152						

Chem-20

	Na	K	Cl	HCO3	BUN	Creat	Gluc	Ca	Phos	Urate	Chol	Tot Prot	Alb	TBili	DBili	Tot Alk P	AST	ALT	LDH	CK
24Mar99 22:50	138	4.0	107	27	16	1.7	133	8.0	2.8											
25Jan99 10:00	138	4.0	105	27	11	1.2	87													165
24Jan99 23:30																				155
24Jan99 17:15	140	4.4	107	30	11	1.3	94				105 *		3.6	0.9	0.2	70	47	22		152
14Oct97 13:35	142	4.6	108	28	14	1.3	85													
12Apr97 09:15	141	3.7	114	21	17	1.1	76													
10Apr97 18:20	138 *	3.8 *	111 *	26 *	28 *	1.2 *	91 *	8.0												
07Apr97 10:20	139	4.0	108	25	18	1.0	82													
06Apr97 10:10	137	3.8	107	27	13	1.0	88													
05Apr97 10:00	141	3.7	107	26	11	1.0	87													
	Na	K	Cl	HCO3	BUN	Creat	Gluc	Ca	Phos	Urate	Chol	Tot Prot	Alb	TBili	DBili	Tot Alk P	AST	ALT	LDH	CK
04Apr97 13:08	144	4.1	112	26			107													
24Mar97 15:12	143	4.0	109	32	12	1.2	74					7.0	3.5	0.4	0.1	74	26	25		
14Dec96 21:13	SPE	SPE	SPE	SPE	SPE	SPE	SPE													
13Dec96 12:50	138	3.5	110	28	7	1.0	80													
12Dec96 12:10	140 *	3.9 *	108 *	25 *	6 *	1.0 *	82 *	8.3	1.6											
11Dec96 04:40	144 *	3.7 *	114 *	24 *	11 *	0.9 *	99 *	8.1	2.1				3.2			41	22	9		94 *
10Dec96 21:20	145	3.8	113	24	13	0.9	84													

Figure 3 Screens from two applications that use the Medical Entities Dictionary (MED) knowledge about spreadsheets. *Top*, A display from the Decision-supported Outpatient Practice application, an X-Window application, showing the Chem-20 spreadsheet. Each row corresponds to a class of laboratory test terms in the MED. *Bottom*, Use of the same information by WebCIS to create a Chem-20 display for the Web.

education'' (G. O. Barnett, personal communication, 1997), computer systems can assist with this task if they have sufficient knowledge about the context of care to anticipate the information need and if they have enough information about potential resources to help direct the clinician. They can also facilitate retrieval of the specific relevant information by using data about the patient to seed the search strategy. We have used the MED to support such tasks through a variety of applications that we refer to collectively as ''infobuttons.''

The first such application used the MED, together with the UMLS, to provide translations from diagnosis and procedure codes in a patient's record to MeSH terms for searching the medical literature.[34] Al-

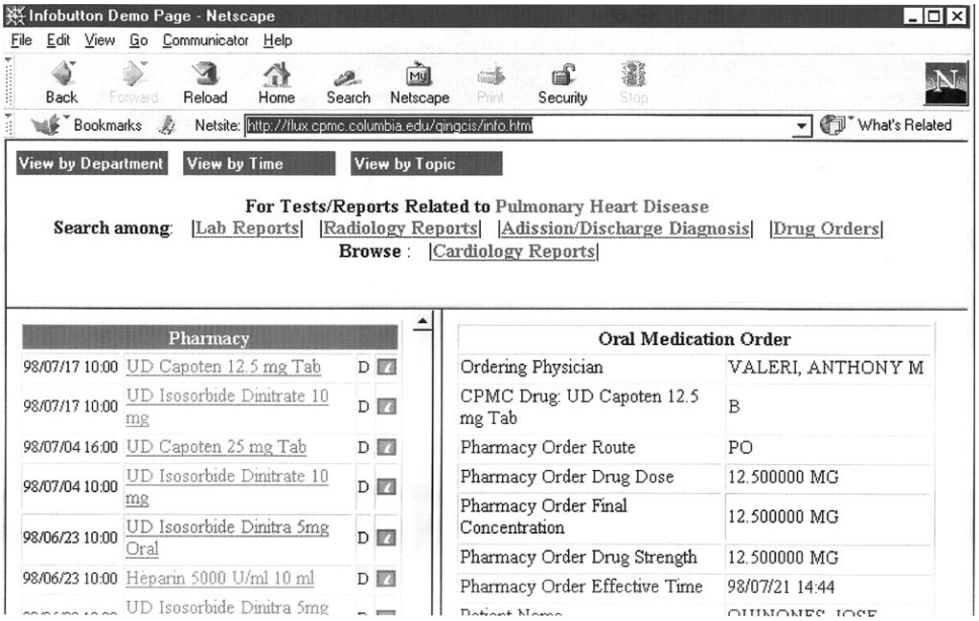


Figure 4 A concept-oriented view of a patient's medical record, generated from Medical Entities Dictionary knowledge. In this example, the problem of interest is pulmonary heart disease, and the data displayed are a subset of medication orders.

though the application did manage to carry out automatic MEDLINE searches, the technical process was awkward and unreliable. The advent of the World Wide Web has greatly simplified our ability to integrate online resources with our clinical information system. As a result, several different infobuttons have been created to link coded data[35] and text reports[36] to relevant resources such as PubMed, pharmacology reference books, and library materials in Utah, Wisconsin, and England.

Expert Systems

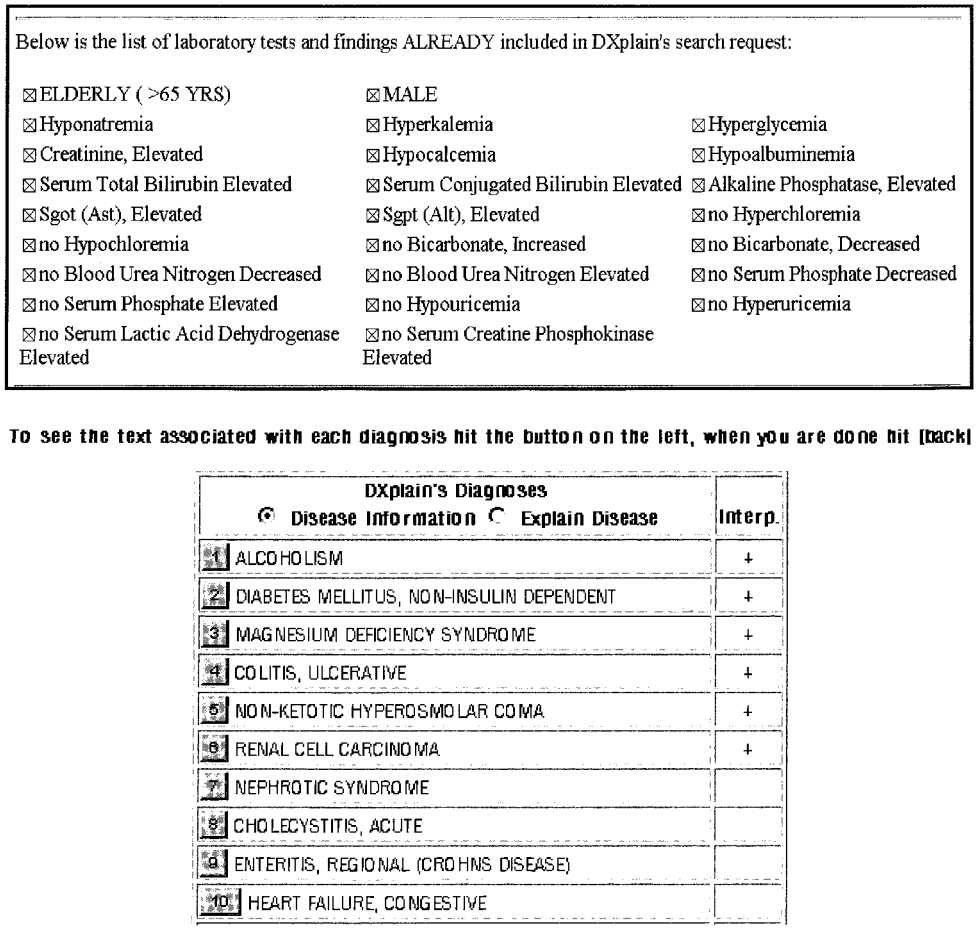
A centerpiece of our clinical information system has been the clinical decision support system.[37] The MED supports this system by integrating the relatively high-level terms used in decision rules (such as "Blood Sugar") with the relatively low-level terms used in the clinical record (such as "Stat Whole Blood Glucose Test"). Through this integration, the task of rule authoring is insulated from the occasional, and even day-to-day, changes that occur in the terminologies used to record patient data. The MED also plays a role in the end-to-end process of parsing and coding test reports for evaluation by rules searching for clinical conditions. In one study, by Hripcsak et al.,[38] the system reliably detected the evidence of six conditions of interest in 200 reports at a rate that was indistinguishable from expert human reviewers.

A long-held model for applying knowledge to patient care has been the expert system, in which expert knowledge is encoded in a system and brought to bear on specific patient problems when relevant data are supplied to the system. Elhanan et al.[39] used the MED to convert laboratory results into findings recognized by a diagnostic expert system called DXplain.[40] The terms were converted by linking test terms (such as "Serum Sodium Test") to measurable substances (such as "Sodium"), which were, in turn, linked to findings (such as "Hyponatremia" and "Hypernatremia"). In this way, a panel of test results could be converted to a patient description and passed to DXplain to obtain a differential diagnosis (Figure 5).

Automated guidelines are another form of expert system that has been successfully integrated with our clinical systems. Applications that encode the guidelines for cholesterol management[41] and mammography recommendations[42] have been integrated into the PatCIS (Patient Clinical Information System) project.[43] Users of the system can have their data automatically retrieved from their records, converted to the appropriate forms, and passed to the guideline programs to obtain results with a minimum of interaction with the guideline logic.

Data Mining

The clinical record holds knowledge that has implications beyond the care of individual patients. By studying patterns and trends through a process known as data mining, it is possible to generate new medical knowledge from patient data. The MED has supported such efforts directly though its coding of the patient record. For example, Wilcox and Hripcsak[44] have used the MED, together with natural



Below is the list of laboratory tests and findings ALREADY included in DXplain's search request:

☒ELDERLY (>65 YRS)	☒MALE	
☒Hyponatremia	☒Hyperkalemia	☒Hyperglycemia
☒Creatinine, Elevated	☒Hypocalcemia	☒Hypoalbuminemia
☒Serum Total Bilirubin Elevated	☒Serum Conjugated Bilirubin Elevated	☒Alkaline Phosphatase, Elevated
☒Sgot (Ast), Elevated	☒Sgpt (Alt), Elevated	☒no Hyperchloremia
☒no Hypochloremia	☒no Bicarbonate, Increased	☒no Bicarbonate, Decreased
☒no Blood Urea Nitrogen Decreased	☒no Blood Urea Nitrogen Elevated	☒no Serum Phosphate Decreased
☒no Serum Phosphate Elevated	☒no Hypouricemia	☒no Hyperuricemia
☒no Serum Lactic Acid Dehydrogenase Elevated	☒no Serum Creatine Phosphokinase Elevated	

To see the text associated with each diagnosis hit the button on the left, when you are done hit [Back]

DXplain's Diagnoses ◉ Disease Information ○ Explain Disease	Interp.
1 ALCOHOLISM	+
2 DIABETES MELLITUS, NON-INSULIN DEPENDENT	+
3 MAGNESIUM DEFICIENCY SYNDROME	+
4 COLITIS, ULCERATIVE	+
5 NON-KETOTIC HYPEROSMOLAR COMA	+
6 RENAL CELL CARCINOMA	+
7 NEPHROTIC SYNDROME	
8 CHOLECYSTITIS, ACUTE	
9 ENTERITIS, REGIONAL (CROHNS DISEASE)	
10 HEART FAILURE, CONGESTIVE	

Figure 5 Integrations of DXplain with a clinical information system. *Top*, The numeric results from a chemistry panel have been converted to specific clinical findings to be passed to DXplain. *Bottom*, The differential diagnosis obtained from DXplain.

language processing, to identify patient records of interest for clinical studies. These researchers have also used knowledge in the MED to construct tools that use terminological knowledge to help would-be data miners understand what data are in the medical record, how they are coded, and how best to retrieve them.[45]

Terminology Maintenance and Use

The knowledge in the MED was originally included to support intelligent terminology maintenance tools. This knowledge has, indeed, been used for this purpose. Terminologies from disparate laboratory systems at Presbyterian Hospital were successfully merged and a tool was created to support automated update of the pharmacy terminology in the MED. This tool also proved useful for detecting discrepancies in the pharmacy's terminology, particularly with regard to missing allergy information.[46]

More recently, the MED has supported the integration of information from disparate systems as Presbyterian Hospital merged with New York Hospital (unpublished data). Other tools have been developed to facilitate the browsing and visualization tasks needed by terminology editors.[47] These tools have been used successfully to help correct errors and inconsistencies in the MED and to improve its comprehensibility.[48]

Discussion

Knowledge representation in medical informatics has a rich history. While much of the previous work has been focused on how to organize symbols into knowledge, the representation of the symbols themselves has turned out to be as important, if not more important, for supporting the use of knowledge in practical systems. One can theorize that the lackluster adoption of artificial intelligence systems in health care may be due in part to failure to ascribe proper importance to "getting things by their right name." In any case, many terminology developers today are committing extensive resources to the task of knowledge representation because they believe that this approach will serve them well in managing their products and serve their clients well in using their products. Time will tell whether the extra effort is worthwhile. The recently announced merger of the Read Codes and SNOMED[49]

will be particularly interesting to watch, since each includes definitional knowledge that is (theoretically) interchangeable and may support the merging process.

The MED at Columbia University and the New York Presbyterian Hospital is but one example of a knowledge-based terminology, and a rather modest one in comparison with current efforts elsewhere. However, it does serve as a proof-of-concept for the general approach, and we have had a decade of experience in building and using it. From that experience, we can offer anecdotal evidence that the effort to include knowledge in a terminology is, indeed, worth it. The knowledge in the terminology lets us take coded patient data and arrive at knew knowledge in several ways.

First, we can gain knowledge about the patient. Although clinicians may claim that they need to know all the data about a patient to make appropriate decisions, human memory is simply no match for the amount of information modern medicine is capable of generating about a complex patient. By using knowledge about the meaning of the data, we can retrieve, filter, and organize them in more intelligent ways, which are appropriate to the task at hand and reduce cognitive overload.[33]

Second, patient data are potentially useful for pointing us to relevant information resources; they are more likely to be useful if they can be translated or mapped to a form that can be used to search a resource. For example, a serum sodium test result of 120 cannot be used to retrieve useful MEDLINE citations by searching PubMed for "120" or even by searching for "serum sodium test." Knowing that the test result is related to a term that *is* recognized by PubMed, such as "hyponatremia," provides the necessary bridge between patient data and the knowledge available in the medical literature.

Third, we finally have an opportunity to bring expert systems to bear directly on the task of patient care. These systems are typically constructed using terms that are appropriate for medical decisions but not equivalent to those appearing in the medical record.[50] As a result, using expert systems requires human translation and transfer of the information from the medical record to the system.[51] The ability to translate terms as originally envisioned by the UMLS developers,[26] coupled with the ease of integrating applications on the Web, offers exciting potential for expert systems to find practical use in everyday patient care.

Fourth, the medical records of patients contain invaluable information about the human condition that can inform clinical research. However, to mine the gems from these data, we need to know where to look and how to recognize what we find. At least in the MED's case, knowledge-based approaches are helping support both these tasks.[44,45]

Finally, the development of complex medical applications to support patient care demands its own type of knowledge about how all the pieces fit together. In our case, this includes the task of maintaining the MED knowledge itself. The incorporation of such knowledge into the MED has implications beyond its simple symmetry; it facilitates the use of expert systems to audit the knowledge and applications to verify that they will function as intended. The example of discovering missing allergy information in the pharmacy system is of more than theoretic interest: it is a concrete example of how MED knowledge about itself can have a potentially life-saving impact on patient care.

Although not originally intended as a data dictionary, information about the clinical repository's tables and columns, and their interdependencies, has been added to the MED. This knowledge has the potential to support database administrators and system developers in their understanding of how coded data relate to the database structure (S. B. Johnson, personal communication, 1999). The advantages of having the database model represented as a collection of concepts, integrated with the concepts stored in the database, seem likely. For example, subsets of the MED can be reused in different parts of the database. Also, if the data model is changed, the impact on the terminology should be apparent, and vice versa. However, it is too early to tell how this form of knowledge will prove most useful.

Terminology requirements, as stated by researchers in terminological work, were recently collected and summarized as set of "desiderata."[52] Two short years ago, I was unable to predict "whether the semantic, definitional information provided by [terminology] developers will be minimal, complete, or somewhere in between." Cautious optimism now suggests that current efforts are moving toward the "complete" end of the spectrum. Getting there will require change, compromise, and overcoming technical, epistemological, and political hurdles. As we move forward, we will do well to recall the namesake for the terminology desiderata:

> *Go placidly amid the noise and haste, and remember what peace there may be in silence. As far as possible, without surrender, be on good terms with all persons.*
>
> —Desiderata, MAX EHRMANN, 1927

The author thanks all his colleagues at the Columbia University Department of Medical Informatics for working with him to expand and exploit the knowledge in the MED in innovative and exciting ways, including Paul Clayton, George Hripcsak, Steve Johnson, Soumitra Sengupta, Carol Friedman, Bob Sideli, Justin Starren, Randy Barrows, Bruce Forman, Barry Allen, Robert Jenders, Gai Elhanan, Qing Zeng, Adam Wilcox, and Eneida Mendonça. He also thanks Sue Bakken for the inspiration to write on this topic, Nancy Lorenzi for the opportunity to present it at the 1999 AMIA Fall Symposium, and Andria Brummitt for editorial support.

References ■

1. Cimino JJ, Hripcsak G, Johnson SB and Clayton PD. Designing an introspective, controlled medical vocabulary. Proc Annu Symp Comput Appl Med Care. 1989:513–8.
2. Ledley R, Lusted L. Reasoning foundations of medical diagnosis. Science. 1959;130:9–21.
3. Shortliffe EH, Axline SG, Buchanan BG, Merigan TC, Cohen SN. An artificial intelligence program to advise physicians regarding antimicrobial therapy. Comput Biomed Res. 1973; 6(6):544–60.
4. Starren J, Xie Q. Comparison of three knowledge representation formalisms for encoding the NCEP cholesterol guidelines. Proc Annu Symp Comput Appl Med Care. 1994: 792–6.
5. van der Lei J, Musen MA. Separation of critiquing knowledge from medical knowledge: implications for the Arden Syntax. In Timmers T, Blum BI (eds). Proceedings of the International Medical Informatics Association Working Conference on Software Engineering in Medical Informatics. Amsterdam, The Netherlands: North-Holland Press, 1990:499–509.
6. Campbell KE, Das AK, Musen MA. A logical foundation for representation of clinical data. J Am Med Inform Assoc. 1994;1:218–32.
7. Chute CG. The Copernican era of healthcare terminology: a re-centering of health information systems. Proc AMIA Symp. 1998:68–73.
8. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. Proceedings of the American Association for Medical Systems and Informatics Congress. 1989:113–7. Also in: MD Comput. 1990;7(2):104–9.
9. Masarie FE Jr, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24(4):379–400.
10. Bernauer J. A controlled vocabulary framework for report generation in bone scintigraphy. Proc Annu Symp Comput Appl Med Care. 1990:195–9.
11. Bernauer J. Conceptual graphs as an operational model for descriptive findings. Proc Annu Symp Comput Med Care. 1991:214–8.
12. Campbell KE, Musen MA. Representation of clinical data using SNOMED III and conceptual graphs. Proc Annu Symp Comput Appl Med Care. 1993:354–8.
13. Campbell KE, Cohn SP, Chute CG, Rennels G, Shortliffe EH. Gálapagos: computer-based support for evolution of a convergent medical terminology. Proc AMIA Annu Fall Symp. 1996:269–73.
14. Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp. 1997:640–4.
15. Bakken S, Cashen MS, Mendonça EA, O'Brien A, Zieniewicz J. Representing nursing activities within a concept-oriented terminological system: evaluation of a type definition. J Am Med Inform Assoc. 2000;7:81–90.
16. Rector AL, Nowlan WA, Glowinski A. Goals for concept representation in the GALEN project. Proc Annu Symp Comput Appl Med Care. 1994:414–8.
17. Brown PJB, O'Neil M, Price C. Semantic definition of disorders in version 3 of the Read Codes. Methods Inf Med. 1998;37(4–5):415–9.
18. Hardiker NR, Rector AL. Modeling nursing terminology using the GRAIL representation language. J Am Med Inform Assoc. 1998;5(1):120–8.
19. Huff SM, Rocha RA, McDonald CJ, et al. Development of the Logical Observations Identifiers, Names, and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5(3): 276–92.
20. Baorto DM, Cimino JJ, Parvin CA, Kahn MG. Using Logical Observations, Identifiers, Names, and Codes (LOINC) to exchange laboratory data among three academic hospitals. Proc AMIA Annu Fall Symp. 1997:96–100.
21. Lau LM, Lam SH. Comparing the mapping of laboratory results to LOINC among different healthcare institutions using a common data dictionary. Proc AMIA Symp. 1999:1105.
22. Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a controlled medical vocabulary server: the VOSER project. Comput Biomed Res. 1994;27(6):472–507.
23. Rector AL, Solomon WD, Nowland WA, et al. A terminology server for medical language and medical information systems. Methods Inf Med. 1995;34(1–2):147–57.
24. Chute CG, Elkin PL, Sherertz DD, Tuttle MS. Desiderata for a clinical terminology server. Proc AMIA Symp. 1999:42–6.
25. Mays E, Weida R, Laker M, et al. Scalable and expressive medical terminologies. Proc AMIA Annu Fall Symp. 1996: 259–63.
26. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993;32(4): 281–91.
27. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998;5(1): 1–11.
28. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. 1994; 1(1):35–50.
29. Johnson S, Friedman C, Cimino JJ, Clark T, Hripcsak G, Clayton PD. Conceptual data model for a central patient database. Proc Annu Symp Comput Appl Med Care. 1991: 381–385.
30. Barrows RC, Allen BA, Sherman E, Smith K. A Decision-supported outpatient practice system. Proc AMIA Annu Fall Symp. 1996:792–6.
31. Hripcsak G, Cimino JJ, Sengupta S. WebCIS: large-scale deployment of a Web-based clinical information system. Proc AMIA Symp. 1999:804–8.
32. Elhanan G, Cimino JJ. Controlled vocabulary and design of laboratory results displays. Proc AMIA Annu Fall Symp. 1997:719–23.
33. Zeng Q, Cimino JJ. Evaluation of a system to identify relevant patient information and its impact on clinical information retrieval. Proc AMIA Symp. 1999:642–8.
34. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The MEDLINE button. Proc 16th Annu Symp Comput Appl Med Care. 1992:81–5.
35. Cimino JJ, Elhanan G, Zeng Q. Supporting Infobuttons with

terminological knowledge. Proc AMIA Annu Fall Symp. 1997:528–32.
36. Zeng Q, Cimino JJ. Linking a clinical system to heterogeneous information resources. Proc AMIA Annu Fall Symp. 1997:553–7.
37. Hripcsak G, Clayton PD, Jenders RA, Cimino JJ, Johnson SB. Design of a clinical event monitor. Comput Biomed Res. 1996;29(3):194–221.
38. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1955;122(9):681–8.
39. Elhanan G, Socratous DA, Cimino JJ. Integrating DXplain into a clinical information system using the World Wide Web. Proc AMIA Annu Fall Symp. 1996:348–52.
40. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: an evolving diagnostic decision-support system. J Am Med Assoc. 1987;258(1):67–74.
41. Cimino JJ, Socratous SA, Clayton PD. Automated guidelines implemented via the World Wide Web [poster]. Proc Annu Symp Comput Appl Med Care. 1995:941.
42. Chen Y, Wang SS, Cimino JJ. Linking guidelines for mammography to an electronic medical record for use by patients. Proc AMIA Symp. 1999:1041.
43. Cimino JJ, Sengupta A, Clayton PD, Patel VL, Kushniruk AW, Huang X. Architecture for a Web-based clinical information system that keeps the design open and the access closed. Proc AMIA Symp. 1998:121–5.
44. Wilcox AB, Hripcsak G. Classification algorithms applied to narrative reports. Proc AMIA Symp. 1998:455–9.
45. Wilcox AB, Hripcsak G, Chen C. Creating an environment for linking knowledge-based systems to a clinical database: a suite of tools. Proc AMIA Annu Fall Symp. 1997:303–7.
46. Cimino JJ, Johnson SB, Hripcsak G, Hill CL, Clayton PD. Managing vocabulary for a centralized clinical system. Medinfo. 1995:117–20.
47. Gu H, Halper M, Geller J, Cimino JJ, Perl Y. Utilizing OODB schema modeling for vocabulary management. Proc AMIA Annu Fall Symp. 1996:274–8.
48. Gu H, Halper M, Geller J, Perl Y. Benefits of an object-oriented database representation for controlled medical terminologies. J Am Med Inform Assoc. 1999;6:283–303.
49. College of American Pathologists. SNOMED® RT and READ Codes to Be Combined into an International Terminology of Health [press release]. May 19, 1999. CAP Web site. Available at: http://www.cap.org/html/public/snomed_intl.html. Accessed Mar 15, 2000.
50. Wong ET, Pryor TA, Huff SM, Hau PJ, Warner HR. Interfacing a stand-alone diagnostic expert system with a hospital information system. Comput Biomed Res. 1994;27:116–29.
51. Miller RA, McNeil MA, Challinor SM, Masarie FE, Myers JD. The Quick Medical Reference Project: status report. West J Med. 1986;145:816–22.
52. Cimino JJ. Desiderata for controlled medical vocabularies in the 21st century. Methods Inf Med. 1998;37(4–5):394–403.