MAKING SENSE OF THE DATA: USING A MEDICAL DATA DICTIONARY TO INTEGRATE, SHARE, AND UNDERSTAND CLINICAL DATA

Abstract

Today, a vast array of information is available to those seeking answers to medical questions. Consumer-oriented health Web sites, computer-based patient record systems (CPRs), electronic medical record systems (EMRs), and clinical bibliographic databases offer unprecedented access to medical knowledge. However, accurately and efficiently retrieving the desired information from these sources remains a challenge, whether you are a physician pursuing a difficult diagnosis or a researcher exploring a clinical protocol.

The sheer volume of information—combined with the idiosyncrasies of medical nomenclature—can make finding the right information a frustrating and time-consuming experience. At the same time, the importance of accurate medical information increases. In today's competitive environment, healthcare enterprises must excel in user/patient satisfaction, quality management, outcomes analysis, clinical guideline development, and fact-based reimbursement, all in the face of constantly changing rules and regulations.

Is there a "right" medical vocabulary that can describe, access, and make sense of all this clinical data? Unfortunately, no single healthcare vocabulary or terminology can meet all the needs of all the people who use healthcare information. Each terminology in use today (e.g., ICD, CPT®, SNOMED®, LOINC, etc.) has been designed for different purposes by different healthcare constituencies. What is needed is an integrated, structured terminology system that can mediate these differences and easily be incorporated into existing clinical information systems.

Over three decades ago, the concept of a **medical data dictionary** began as a means of supporting clinical information systems and the CPR. This paper describes how a well-designed, intelligent medical data dictionary with terminology services can enable a healthcare organization to integrate, share, and understand clinical data, received in all of its various formats and terminologies from diverse systems and sources. When applied in real-time in an organization's information system infrastructure, such a dictionary can—

- Describe clinical data in all its possible forms, providing a road map to the content and structure of the patient database.
- Support encoding of clinical data to remove ambiguities.
- Support exchange and comparison of data between independent computer systems.
- Provide structure and content for decision support across care encounters.
- Enable users to effectively query and report on the database.
- Support standardization of clinical data across enterprises by incorporating industry-standard **controlled medical vocabularies (CMVs)** as its foundation.

The key to accomplishing these tasks is the incorporation of *de facto* CMVs (e.g., LOINC) into the medical data dictionary itself. Adding and "mapping" (cross-referencing) industry-standard CMVs to a dictionary are labor-intensive efforts for the dictionary authors, but the resulting product can help support standardization across enterprises and enable clinical data to be compared and aggregated at national or international levels. ■

What's not making sense in clinical data today?

What does the word "cold" mean? Consider these possibilities:

- An accident victim brought into the ER tells the attending physician, "I feel cold."
- A pulmonologist tells a 58-year-old male patient that he is suffering from COLD (Chronic Obstructive Lung Disease).
- You call your family practitioner for an appointment and tell the receptionist, "I have a bad cold that's not getting better."

If the word "cold" is recorded on a patient's chart during any of these three scenarios, how does that word become an accurate and meaningful part of the computerized patient record (CPR)? How can the term be translated so that outcomes can be understood and managed and the medical events themselves accurately described and interpreted?

What is the "correct code" for a glucose result? Consider these possibilities:
An organization currently receives laboratory test results from two different external vendors, a government lab, and several internal STAT labs. Each external system uses its own distinct set of test result codes, which differ from the internal set of codes used by the organization's STAT labs. Patient John Doe has glucose test results from one or more of these labs—how can the organization's order entry and results review applications "reconcile and reunite" all of these different test codes and results in the patient's record? How can clinicians "see" all of Doe's glucose results together for effective comparison and comprehensive analysis?

How can a healthcare organization make well-informed decisions regarding future expansion and resource deployment? Consider this possibility:
As part of the strategic planning process, a healthcare organization's team of risk managers, care providers, and financial officers wants to examine population studies by facility to determine how best to deploy an expensive special service like cardiology. How can the team get an accurate report of all cardiology-specific services and treatments delivered across the organization?

In all of these situations, people are defining the particular term, code, or information query according to a particular context. In the first example, "cold" can be a sensory perception, a pulmonary diagnosis, or an upper respiratory viral infection. The human mind easily resolves the ambiguity of a word like "cold," especially when the term is considered in its context. But how can a computer track and correctly interpret "cold," especially when it can appear in various forms and many different contexts?

Computers easily process enormous volumes of data, but computers also "expect" unambiguous data in a specific form. Using computerized clinical data means more than simply storing numbers and words. Terms must be clearly defined and placed in a context. The vocabulary must be carefully monitored to avoid duplication, support synonyms, and completely describe terms from all areas of medicine—these characteristics lie at the heart of a **medical data dictionary.**

What is a medical data dictionary?

A medical data dictionary is a database that describes the organization and logical structure of the medical data found in a clinical database. It contains "metadata"—or "data about data"—that describes the content, structure, and relationships between clinical data. In short, a medical data dictionary "translates," precisely defines, and effectively accesses the contents of the **computerized patient record, or "CPR."**

What information does today's CPR contain?

A look at the information stored in today's typical CPR uncovers the Tower of Babel that exists in healthcare:

- Data comes from and resides in many different *sources* (systems and databases)—
  - Laboratory systems
  - Pharmacy systems
  - Medical record coding systems
  - Radiology systems
  - Hospital information systems
  - Billing systems

- Data from these diverse sources exists in many different *formats*—
  - Coding and classification system formats
  - Controlled medical vocabulary (CMV) formats (it is likely that a typical patient CPR would contain ICD-9-CM codes, CPT® codes, LOINC codes, etc.)
  - Proprietary formats from "home-grown" and commercial laboratory, hospital information, billing, and pharmacy systems and databases

Who is trying to use the information in the CPR?

The data in the CPR is valuable and needs to be available to many different people, ranging from clinicians and administrators to researchers and government regulatory agencies. The data should be retrieved and returned to these people in the appropriate format and with the correct degree of granularity for the audience. Data needs to be accessed in various ways and for many different purposes: patient clinical reports, statistical studies, *ad hoc* reporting, regulatory requirements, etc.

Not only must all of this data in the CPR be integrated, it must also be "normalized" into a form that can easily be shared by all audiences. For example, if users are compiling organization-wide reports on cases of acute myocardial infarctions, they want to easily find all cases in the database, regardless of data format or terminology (e.g., "MI," "myocardial infarction," ICD code 410, etc.). By the same token, users should be able to retrieve all those cases—and even the associated patient records—for further study in the data formats of their choice. A well-designed and intelligent medical data dictionary defines the connections between "MI," "myocardial infarction," and ICD code 410 within a CPR, and also makes it possible to retrieve and return the term most appropriate for a given audience.

What are the distinctions between a medical data dictionary, a CMV, and a coding and classification system?

As medical informaticists, vocabulary experts, and the medical community have refined their definitions of the CPR and a medical data dictionary, they have also refined the meaning and scope of a CMV. A working definition of CMV is "a standard code set and an associated semantic network that represents the information within a major domain of medicine."[1] A CMV has the primary purpose of describing data and follows what medical informaticists refer to as "vocabulary

[1] Hieb B, Rishel W, GartnerGroup. Defining "Controlled Medical Vocabulary." Research Note, Tutorials TU-10-9951, 8 May 2000.

principles." According to GartnerGroup, a CMV is characterized by these elements:

- *Concept*—an idea that has relevance to a discipline. It is a logical abstraction that represents the idea independently of any expression or description (e.g., "chest," "lung," or "x-ray").
- *Term*—a string of characters that represents a concept. A term is a linguistic expression of a concept and can consist of more than one word (e.g., "subacute bacterial endocarditis").
- *Compound concept*—a set of concepts combined into a more complex or more granular concept (e.g., "greenstick fracture of left ulna").
- *Code*—a machine-usable term that represents a concept. A code can be mapped one-to-one with a corresponding term.
- *Standard (or defined) code set*—a list of terms and their related codes that has gained acceptance in the healthcare industry and is administered by a specific organization.
- *Relationship*—a named and directed (one way) association between two concepts (e.g., "is a component of," "is a parent of"). A relationship establishes a semantic link between the two concepts.
- *Semantic network*—a set of concepts and a defined set of named relationships that expresses the known or relevant relationships between concepts for a specific domain of knowledge.[2]

2 Hieb.

Under GartnerGroup's definition, First Data Bank's pharmacy database, which contains a comprehensive and up-to-date list of National Drug Codes (NDCs), can be considered a CMV in the domain of pharmaceuticals. Another well-known example of a commercially available CMV is the **College of American Pathologists' Systematized Nomenclature of Medicine (SNOMED®).** The current version, **SNOMED® Clinical Terms, or SNOMED CT®,** is a full-fledged reference terminology that contains over 366,170 concepts, more than 993,420 synonyms, and approximately 1.46 million semantic relationships.[3] An intelligent medical data dictionary shares the same characteristics as a CMV, but has the important advantage of encompassing multiple CMVs.

3 SNOMED Clinical Terms® (SNOMED CT®) Core Content as of July 2005. Fact sheet from College of American Pathologists, July 6, 2005.

Coding and classification systems

A medical data dictionary can also include multiple **standard coding and classification systems and code sets**, such as the **International Classification of Diseases (ICD) coding sets** and the **American Medical Association's Current Procedural Terminology (CPT®).** Because it is required information and a standard format, a classification system like ICD-9-CM ("CM" indicates "Clinical Modifications") continues to be part of the CPR, despite the fact that classification systems generally have the primary purpose of grouping data and do *not* follow vocabulary principles.

Obviously, a CMV and a reference terminology can provide much more information about a patient than a basic classification system. But, consider the healthcare industry's current information exchange needs:

- Internationally, there are over 100 healthcare-related vocabulary sets, making comparisons of healthcare data between nations virtually impossible
- In the U.S., the average institution must handle between five and ten such vocabulary sets just to meet reporting and reimbursement requirements.[4]

4 Hammond, E. Call for a standard clinical vocabulary, J Am Med Informatics Assoc. 1997;4:254–255.

At present (and for the foreseeable future), these information exchange needs and requirements cannot be served by any single coding and classification system or any single CMV.

Yet another type of "language" in the CPR: Interface terminologies (organization- and user-specific terms)

But there is another harsh fact of life for today's typical healthcare organization: it must continue to use its legacy information systems until economic conditions allow them to be replaced or upgraded. Such information systems transmit data among themselves by means of interfaces, and each interface system has its own "language" or terminology. Many legacy systems use their own proprietary codes to describe clinical data, and these codes are not "understood" by any other system. The number and types of interfaces and proprietary coding systems are unique and specific to each healthcare organization.

The limits of Health Level 7 (HL7)

Moreover, interfaces do not evaluate the *content* of the messages they transmit. The healthcare industry-standard interface protocol **HL7** has standardized the message structure (syntax) of healthcare information transmissions. Thus, every organization that is HL7-compliant can send messages to other HL7-compliant organizations; as long as the message syntax is correct, HL7 considers the transmission successful. However, in HL7 version 2, while the syntax may be correct, the actual content of the transmission can be gibberish. Organizations can send messages to one another only to find that because the content of the message is not standardized, they still are not successfully "talking" to one another. Thus, HL7 version 3 is focusing on vocabulary standardization in its development.

In addition, the people within a healthcare organization will continue using their accustomed "languages" and vocabularies. If they are familiar with a particular legacy system and its idiosyncrasies, they will continue to use that system. If site-specific pharmacy formularies are in place, those will continue to be used in addition to any other globally standard formulary. In the larger scheme of things, all of these "interface terminologies" and site-specific entities are legitimate components of a healthcare organization's "working vocabulary" and should be a part of an intelligent medical data dictionary.

As will be demonstrated, the power of the medical data dictionary is its ability to encompass and create the links between industry standard CMVs, coding sets, and organization-specific vocabularies. By integrating both global and site-specific vocabularies, the medical data dictionary allows **data in a CPR to be cross-referenced to standards everyone can understand, rather than locking the data up into yet another proprietary language that only a select few can speak.** When a dictionary can seamlessly translate universal and organization-specific data "behind the scenes," the healthcare organization can save time, money, and resources by avoiding the re-tooling or replacement of legacy systems and re-training of its people.

Technology and industry drivers—why does a comprehensive medical data dictionary matter now more than ever?

A variety of technology and industry drivers continually exert pressure on healthcare organizations—they always have and they always will. To a surprising degree, an intelligent medical data dictionary that encompasses the major CMVs can address many of the challenges posed by many of today's drivers. For example—

Weak information infrastructure. In an effort to achieve hoped-for economies of scale, healthcare providers in the recent past consolidated and created integrated delivery networks (IDNs). This consolidation brought together disparate legacy systems, and the IDNs had an economic need to keep using those systems, despite the fact that sharing data between them was difficult and often impossible. Today, as IDNs break up and move apart, they must still overcome weak information infrastructures to stay competitive. Even modest-sized healthcare organizations have multiple legacy systems that they cannot afford to replace and still need to unite.

What a medical data dictionary can do: When a dictionary has been designed to be both extensible and flexible, it can encompass all of the diverse "languages" spoken by an organization and its legacy systems. Rather than replace legacy systems, the organization can use the dictionary with terminology services to translate between such systems and normalize the data they send into the CPR.

Government and industry regulations and requirements. The **Healthcare Insurance Portability and Accountability Act of 1996 (HIPAA)** is only one of the latest regulatory drivers. This legislation is moving the healthcare industry toward requirements of a uniform medical "language" and standardized code sets. It also mandates **electronic data interchange (EDI) standards** for electronic transactions that involve healthcare information.

What a medical data dictionary can do: A well-designed medical data dictionary can seamlessly integrate classification systems such as ICD or CPT® along with other standard vocabularies and classification systems. Whatever HIPAA's medical "language of choice" turns out to be, it can be enveloped in a medical data dictionary as another CMV and consequently cross-referenced to the other CMVs still in use within a healthcare organization.

Patient safety. The **1999 Institute of Medicine (IOM) study** on patient safety raised awareness in federal regulators and in the public of serious problems with the provider system in the United States, citing studies that claimed between 44,000 and 98,000 Americans die each year because of medical errors.[5] The IOM report also indicates that 7,000 deaths annually are attributable to medication errors alone. As a result, many experts have since cited the need to implement computerized drug ordering systems.[6]

What a medical data dictionary can do: As will be shown later on in the discussion of the "ideal medical data dictionary," two of the touchstones of an intelligent medical data dictionary is its insistence on unambiguous data and its use of industry-standard CMVs to help reduce the opportunities for misinterpreted, inaccurate, or imprecise

5 Kohn LT, et al. To Err is Human: Building a Safer Health System. Report of the Committee on Quality of Health Care in America, Institute of Medicine, Washington, DC: National Academy Press, 2000.

6 Kohn.

data to become part of the patient's record. For example, to facilitate an automated pharmacy ordering system, a medical data dictionary can and should include a standardized pharmacy CMV (such as First Data Bank) that includes an accurate list of National Drug Codes (NDCs). When a dictionary can "enforce" accurate drug codes and positively identify the drug(s) being ordered, human errors and mistakes can be reduced.

Consumer activism. The language "gap" that exists between physicians and patients is currently being addressed by healthcare thought leaders and software companies alike. "CHT" ("Consumer Health Terminology") refers to an interface terminology that can:

- Translate a lay person's *terminology* into meaningful, clinically valid terminology (and vice versa)
- Allow professional *clinical codes* to be interpreted by the average lay person
- Enable *Web-based content* to be indexed
- Help deliver *personalized health information* to the people who need it[7]

One impetus behind CHT comes from reports that indicate 45 percent of American Internet users go online to find health information, seeking information not only on general health care and preventive measures, but also on very specific diseases. With an estimated 134 million Internet users in the United States, the significance of CHT increases.[8]

What a medical data dictionary can do: Consumers are expecting physicians and providers to communicate with them in terms they understand. A medical data dictionary can accommodate this need by including and cross-referencing a CHT as yet another type of vocabulary.

Increasing number of e-health relationships in the healthcare industry. According to the American Medical Association's Web site, "the Internet is altering the landscape of medicine,"[9] and the concepts of "e-health" and "e-health relationships" are at the center of much of the discussion. These terms refer to every aspect of communicating health-related information through electronic means (e.g., e-mail, Web sites, etc.), and such communication potentially—*or already*—exists not only between healthcare consumers and their providers and caregivers, but also between healthcare organizations, payers, fiscal intermediaries, and government. The GartnerGroup believes that "healthcare organizations that have not implemented standard vocabularies for internal data exchanges and integration with trading partners will be unable to compete for e-health relationships (0.8 probability)."[10]

What a medical data dictionary can do: The challenge here is for an organization to achieve comparable data with its business partners and move toward standardized code sets. A dictionary that adheres to sound vocabulary principles can identify a concept even when it is identified by numerous codes; moreover, these codes may even represent both past and present usage. For example, an ICD-9-CM code that is no longer acceptable to a claims payer must still be part of the dictionary as support for historical data.

7 Marshall P. Bridging the Terminology Gap between Physicians and Patients—The Consumer Health Terminology. AMIA conference panel discussion, 7 November 2000.

8 Marshall.

9 E-health Initiatives: A Matter of Trust. www.ama-assn.org/ama/pub/article/3216-4915.html

10 Rishel W, GartnerGroup. Moving to Standard Medical Vocabularies. Research Note, Tactical Guidelines TG-10-5546, 6 March 2000.

What characterizes the "perfect" CMV—*and medical data dictionary*—for the twenty-first century?

In 1988, informaticists at the Columbia-Presbyterian Medical Center in New York set out to develop a knowledge-based representation for a controlled terminology that could describe the clinical information coming from their ancillary departmental systems.[11] Their project, called the **Medical Entities Dictionary (MED),** is based on the **Unified Medical Language System (UMLS),** begun by the National Library of Medicine in 1987.

[11] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based Approaches to the Maintenance of a Large Controlled Medical Terminology. J Am Med Informatics Assoc. 1994;1:35-50.

James J. Cimino, a well-known and respected informaticist and long-time scholar of medical vocabularies, was the principal investigator on the research grant awarded to Columbia University by the Library. After reviewing the literature of an almost 40-year period, Cimino presented a paper entitled, "Desiderata for Controlled Medical Vocabularies in the Twenty-first Century."[12] The paper compiles its list from a large body of studies done in the last decade of the twentieth century, when medical informaticists really began to be specific in their ideas of what constituted a true CMV. Though not exhaustive, Cimino's list is nevertheless an excellent starting point for healthcare organizations currently evaluating available vocabularies. Very briefly summarized and paraphrased,[13] Cimino's desiderata are:

[12] Cimino JJ. Desiderata for Controlled Medical vocabularies in the Twenty-First Century. IMIA WG6 Conference, Jacksonville, Florida, 19–22 January 1997.

[13] All desiderata are paraphrased and quoted from Cimino's article, unless otherwise noted.

- *Vocabulary content—more is better.* Content is of utmost importance, with the dictionary including as many available standards as possible.
- *Concept orientation.* Vocabularies must use the concept as their "symbolic processing unit," meaning that—
  - Terms correspond to at least one meaning (non-vagueness)
  - Terms have no more than one meaning (non-ambiguity)
  - Meanings correspond to no more than one term (non-redundancy)
- *Concept permanence.* Once created, a concept is "inviolate." Its preferred name may change or it may become inactive or archaic, but its meaning remains.
- *Non-semantic concept identifier.* Each concept must have a unique identifier that is itself "free of hierarchical or other implicit meaning."
- *Polyhierarchy.* "General consensus seems to favor allowing multiple hierarchies to coexist in a vocabulary without arguing about which particular tree is the essential one."
- *Formal definitions.* These describe the "relationships" between concepts. For example, a formal definition for the concept of "Pneumococcal Pneumonia" can include an "is-a" link to the concept of "Pneumonia" and a "caused-by" link to the concept "Streptococcus pneumoniae."
- *Refusal to use "Not Elsewhere Classified" terms.* Vocabularies definitely should neither tolerate nor contain such "catch-all" types of terms.
- *Multiple granularities.* Multipurpose vocabularies need multiple granularities, because one level of granularity is inadequate for all the audiences within a healthcare organization. For example, consider the possible granularities for a diagnosis of "diabetes," a term so coarse-grained as to be considered vague:
  - Diabetes Mellitus (coarse grained—a general practitioner may be most comfortable with this lowest level of granularity for a diagnosis)
  - Type II Diabetes Mellitus (more granular)
  - Insulin-dependent Type II Diabetes Mellitus (of the three terms, this is the most granular—an endocrinologist may insist on this level of granularity)
- *Multiple consistent views.* Since a vocabulary serves multiple functions and

requires different granularities, it must provide multiple consistent views into itself that are suitable for different purposes.

- *Beyond medical concepts: representing context.* Formal, explicit information about how concepts are used and how they can be arranged to make sense.
- *Graceful evolution.* The vocabulary must be able to change over time to keep pace with medical knowledge.
- *Recognition of redundancy.* In the context of CMVs, redundancy is the "condition in which the same information can be stated in two different ways." For example, the existence of synonyms in the vocabulary is a good thing, because users can recognize the terms they use, and all of the synonyms cross-reference the same concept—as a result, the *coding* of the information is *not* redundant.

A working application of the CMV desiderata

As they designed and developed the 3M™ Healthcare Data Dictionary, the researchers and developers at 3M Health Information Systems incorporated Cimino's recommended "desiderata" and design concepts. Several 3M researchers worked with Cimino on the MED/UMLS project and another important data dictionary development project known as "VOSER."[14] The 3M researchers also came to these dictionary projects with over 20 years of experience in coding and classification systems and development with Intermountain Health Care, Inc., of the 3M™ HELP System and its data dictionary, the "PTXT" ("Pointer-to-Text") file.[15]

The result of combining industry experience and professional and academic expertise is a very practical implementation of informatics principles and industry-standard vocabularies. Supporting such applications as data warehousing, order entry, and results review, the 3M Healthcare Data Dictionary is operational in the "real world," having been mapped in 16 commercial healthcare enterprises and all of the hospitals and clinics supported by 100 hosts in the Department of Defense's CHCS II project. It is a working product that is an integral part of the 3M™ Clinical Data Repository, the database component of 3M™ Care Innovation. 3M's strategy has been to partner with healthcare organizations to help them make optimal use of their patient data, with the goal of improving quality of care, outcomes, costs, and competitiveness. The strengths of 3M's dictionary lie in its:

- Structure and depth/breadth of content
- Ability to map "local extensions"
- Architecture
- Clinical foundations/expertise

3M Healthcare Data Dictionary's structure: depth and breadth of content

The 3M Healthcare Data Dictionary is comprised of an **information model, vocabulary,** and **knowledge base.**

Information model

The 3M Healthcare Data Dictionary uses an information model to accurately represent clinical data in the 3M Clinical Data Repository. An information model describes how the vocabulary concepts should be used and how data can be combined to create meaningful database records that represent clinical events. It can be thought of as a set of "grammar rules" that show how data interacts with other data. The information model establishes temporal and spatial contexts for patient data, so that clinical observations can be attributed to the correct patient, clinical observer, and time sequence, and describes the appropriate information domains and types of values that should be present (see *Figure 1,* next page).

[14] For more information on VOSER, see *Appendix A* and the following:
— Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a Controlled Medical Vocabulary Server: The VOSER Project. Comput Biomed Res. 1994;27:472-507.

[15] For more information on the 3M HELP System, see *Appendix A* and the following:
— Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP System. Journal of Medical Systems. 1983;7(2):87-101.
— Pryor TA. The HELP medical record system. MD Computing, 1988;5:22-33.
— Kuperman GJ, Gardner RM, Pryor TA. HELP: A Dynamic Hospital Information System. New York: Springer-Verlag, 1991.
— Warner HR. Computer-assisted medical decision-making. New York: Academic Press, 1979.

A medication order consists of:

Drug	*What drug is being ordered?*
Dose	*What dosage is being ordered?*
Route	*How will the drug be given?*
Frequency	*How often will the drug be given?*
Start time	*When (date/time) will the drug first be given?*
End time	*When (date/time) will the drug stop being given?*
Ordered by	*Who ordered this drug?*
Order #	*What is the order number?*

A medication order can be represented in the computer as:

```
MedicationOrder ::=SET {
    drug          Drug,
    dose          Decimal,
    route         Route,
    frequency     Frequency,
    startTime     DateTime,
    endTime       DateTime,
    orderedBy     Clinician,
    orderedNum    OrderNumber }
```

Figure 1. How concepts are placed in a 3M Healthcare Data Dictionary information model

Concept ID	Definition
Cold, #123	A sensory perception ("patient complains of feeling cold")
Cold, #569	A pulmonary diagnosis (Chronic Obstructive Lung Disease)
Cold, #784	An upper respiratory viral infection ("common cold," "cold," "flu," etc.)

Figure 2. Sample of 3M Healthcare Data Dictionary's concepts and concept IDs

3M™ Healthcare Data Dictionary vocabulary—enriched by its inclusion of industry-standard CMVs and coding and classification systems

The 3M Healthcare Data Dictionary's vocabulary component identifies and represents the various medical concepts found in clinical data, and it is organized to support synonyms and other lexical characteristics. As of August 2005, the 3M dictionary contains over 1.4 million active concepts, over 11 million representations, and nearly 11 million relationships. The source vocabularies are:

- SNOMED CT®
- Unified Medical Language System (UMLS)
- Logical Observation Identifiers Names and Codes (LOINC)
- National Drug Codes (NDCs) from the First Data Bank Pharmacy database
- ICD-9-CM
- Diagnostic Related Groups (DRGs)
- All Patient-DRGs (AP-DRGs)
- All Patient Refined-DRGs (APR-DRGs)
- CPT®
- HCPCS
- PTXT (from the 3M™ HELP System)
- Customer vocabularies (legacy systems, local and organization-specific terms)

In terms of content areas, the 3M Healthcare Data Dictionary encompasses:

- Encounter and demographics
- Laboratory
- Microbiology
- Pharmacy
- Diagnostic and procedural coding
- Findings, signs, and symptoms
- Problem lists and diagnoses

Concepts and concept IDs

In the 3M Healthcare Data Dictionary, a concept is a unique, definable idea or item that has a very specific, known meaning (e.g., cold, temperature, sensation, viral infection, infection, diagnosis) or a combination of concepts ("chest x-ray"). In the 3M dictionary, each concept is defined by both a human-readable text description and an assigned, unique numerical identification, referred to as an "NCID" ("Numerical Concept Identifier;" an NCID has no intrinsic meaning or significance in itself). No redundant concepts are allowed, since they defeat the purpose of a controlled vocabulary. To demonstrate the need for concept IDs and a controlled vocabulary, consider how the word "cold" can be used in medical language, as shown at left in *Figure 2*.

Synonyms—expanding the 3M Healthcare Data Dictionary vocabulary

Much of the richness of the 3M dictionary's vocabulary comes from its use of synonyms (see *Figure 3*, next page, for examples). In its use of synonymy, the vocabulary includes:

- Synonyms, homonyms, and eponyms (names derived from people or places)
- Different representations of the same concept, either in a natural language or other coded format
- Common variants of a term, such as acronyms or even common misspellings

Synonym examples:

Acute Sinusitis

ACUTE SINUSITIS

Acute sinusitis, NOS

Sinusitis, acute

Acute infection of nasal sinus, NOS

Acute inflammation of nasal sinus, NOS

C0149512 (UMLS)

621850 (hospital-specific interface ID)

Possible contexts for a synonym—

Synonym	Context
Acute Sinusitis	*Problem list display*
Acute infection of nasal sinus	*Explanation*
C0149512	*UMLS code*
621850	*Interface code*

Figure 3. Sample of how the 3M Healthcare Data Dictionary uses synonyms and can specify a context

- Foreign language equivalents (human languages—French, Spanish, Portuguese, etc.—can be added)
- The terms preferred in specific contexts (for example, "dyspnea" can be designated as the term for a cardiologist, while "shortness of breath" can be the preferred term for a lay person)

The 3M™ Healthcare Data Dictionary's knowledge base

The 3M dictionary's knowledge base consists of semantic networks and hierarchies that describe the complex relationships existing between concepts in the vocabulary. These relationships can be *hierarchical* (parent-child or "is-a") or *non-hierarchical* ("is-a-component-of"). *Figure 4* (below) is an example of how the knowledge base can describe the relationships between the components of a CHEM 4 laboratory test.

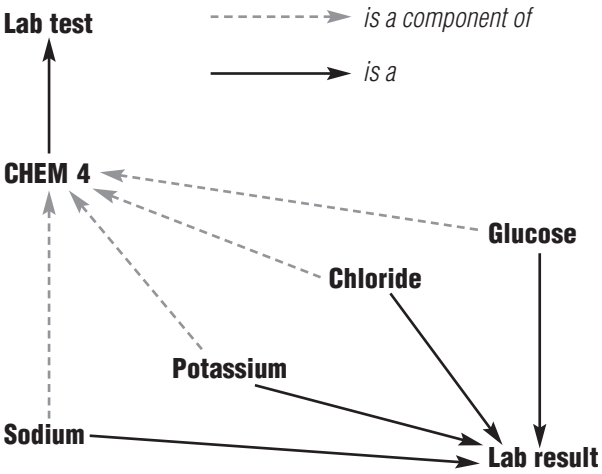

Figure 4. Sample of the 3M Healthcare Data Dictionary's knowledge base as applied to a CHEM 4 lab test

Figure 5 (below) shows how the 3M dictionary's **information model, vocabulary,** and **knowledge base** tie together to create a meaningful database record that represents a clinical event.

Information model

```
MedicationOrder ::=SET {
    drug          Drug,
    dose          Decimal,
    route         Route,
    frequency     Frequency,
    startTime     DateTime,
    endTime       DateTime,
    orderedBy     Clinician,
    orderedNum    OrderNumber }
```

Database record

```
MedicationOrder  {
    drug          Ampicillin,
    dose          500,
    route         Oral,
    frequency     Q6H,
    startTime     08/01/01 10:01,
    endTime       08/11/01 23:59,
    orderedBy     John Doe MD,
    orderedNum    A234567 }
```

Figure 5. Sample of a 3M Healthcare Data Dictionary information model (left) and the resultant database record (right) populated with vocabulary concepts and organized by the knowledge base

The 3M dictionary's ability to "map" source CMVs and "local vocabularies"

All industry-standard CMVs can coexist in the 3M™ Healthcare Data Dictionary because of a process referred to as "mapping," which cross-references elements in each CMV with a concrete and unambiguous concept in the 3M dictionary.

Is mapping really necessary?

The short answer to this question is "Yes," simply because—

- There is no single, universally accepted and applied standard vocabulary for the healthcare industry.
- Existing coding systems are incomplete.
- The HL7 version 2 standard specifies message *structure* only; it does not specify the actual data that is sent within the structure.
- Every healthcare organization uses different terms and codes.
- It is impractical and too expensive to replace all legacy systems.

The 3M dictionary's ability to map "local extensions"

Because of the mapping process, an organization's "local" terminologies can also be integrated with the standard CMVs, along with such interface requirements as interface codes, billing codes, etc. Site-specific mapping is time and resource intensive, requiring highly trained and experienced clinical personnel who—

- *Understand* each CMV's inherent characteristics (e.g., "molecular" combinations)
- *Consider* each CMV's limitations (e.g., the lack of explicit relationships, reuses codes, etc.)
- *Design* a mapping strategy that—
  - Follows vocabulary principles
  - Meets user needs
  - Is flexible and extensible

The advantages of mapping

The ability to create local extensions means that the healthcare organization's users can continue using their own terms, while the dictionary seamlessly handles the "translation" of such terms behind the scene. Mapping provides today's resource- and time-strapped healthcare enterprise with several advantages—

- People do not need to be "retrained" in the language of another computer system. They continue using the terms they know and understand.
- Existing information systems do not need to be replaced or redesigned.
- Individual entities within an organization can retain specific information structures or sets, such as preferred formularies.
- A specific form of the concept can be displayed in a specific context for the user.

For example, the 3M dictionary can map a serum sodium laboratory result as—

Concept ID	Representation	Context
123	1234-5	LOINC code
123	NAS	Interface code for Site #1
123	Serum Sodium	User display for Site #1
123	CL357	Interface code for Site #2
123	S. Sodium	User display for Site #2

The 3M™ Healthcare Data Dictionary's clinical foundations and staff expertise

3M researchers who support the product and provide mapping services include:

- A staff of clinicians and informaticists
- Members of both the laboratory and clinical committees of LOINC
- Members of the American Medical Informatics Association (AMIA)
- Members of HL7 Vocabulary Technical Committee
- Members of Healthcare Information and Management Systems Society (HIMSS)
- Members of American Health Information Management Association (AHIMA)

3M™ Care Innovation architecture and "platform independence"

The 3M Care Innovation components were designed to meet open architecture standards, allow for platform independence, and conform to these industry standards:

- Health Level 7 (HL7)
- CORBAmed Patient Identification Services (PIDS) to access the 3M™ Enterprise Master Person Index
- CORBAmed Terminology Query Service (TQS) to access the 3M Healthcare Data Dictionary
- CORBAmed Clinical Observation Access Service (COAS) to access the 3M™ Clinical Data Repository
- ASN.1 information model, translatable to XML
- Application Programming Interfaces (APIs) to access the 3M Healthcare Data Dictionary and 3M Clinical Data Repository

Finally, because the 3M dictionary is an independent entity, it can be updated with new medical knowledge without rewriting the application programs that use it.

Conclusion

Technology and industry drivers are quickly making a medical data dictionary mandatory for any healthcare organization that wishes to remain competitive—and even viable—in the twenty-first century. The potential power of a medical data dictionary lies in its ability to help healthcare organizations meet these challenges:

- Use the power of computers to deliver **decision support** to care providers in diagnostic, therapeutic, and management arenas.
- Significantly **improve outcomes and contribute to the quality and safety of patient care** by helping to reduce the number of medical errors resulting from misinterpreted, inaccurate, or imprecise patient data.
- Effectively **integrate information systems** to increase the healthcare facility's competitive strength in the community and **reduce the costs associated with poor information infrastructure.**
- Meet the challenges posed by both **industry and governmental mandates** for uniform data standards and standardized clinical vocabularies, coding, and coding sets.
- Bridge communication gaps between healthcare providers, business partners, and consumers—*even as technology and software evolve*—and allow expansion into **Internet-enabled technologies, e-health relationships, and other communication vehicles.**

These challenges are best met through a working medical data dictionary that is an integral part of a clinical data repository. In design and architecture, 3M's dictionary can support existing technologies and continued use of legacy systems, providing a flexible and extensible "vocabulary server" that addresses regulatory pressures and serves the needs of both business partners and healthcare consumers alike. ■

Appendix A: Developing a medical data dictionary—a brief history

Please note: *This appendix is not exhaustive of its subject, but provides significant background on the 3M™ Healthcare Data Dictionary.*

In the early 1960s, the **National Science Foundation** sponsored basic research into language and information as the basis for future information systems. It was thought at the time that "automated language analysis" would provide the bridge between users and stored information.[16] A **National Institute of Health (NIH)** survey conducted in 1973 also pointed out that the majority of data collected in the patient care process is not numerical, but rather is expressed in natural human language.[17] One challenge of developing a medical vocabulary involved the creation of a system that could handle syntactic and semantic constructions—one that could "understand" the meaning of the words (semantics) and also know how to combine them (syntax). In 1991, the **Institute of Medicine's Committee on Improving the Patient Record** cited a "large data dictionary" as the very first trait of its "ideal" CPR system, pointing out both the necessity and high priority of defining the contents of the CPR itself.[18] Independent researchers, clinicians, and the Board of Directors of the **American Medical Informatics Association (AMIA)** have also acknowledged that a medical data dictionary is a "key component of any clinical information system."[19] Attempts to create a dictionary raised awareness of the need to standardize medical vocabularies and create a "medical-concept-representation language" that enables medical data to be effectively integrated and exchanged between users and information systems.[20]

Early methods of encoding medical language: structured text

An early solution to "encoding" medical language was structured text, which is still used in information systems today. Structured text works well for very hierarchical and "predictable" medical data (e.g., medication lists), relying on a rudimentary information model that defines the order of data. For example, a lab result might be:
[patient ID] [test name] [result name] [result value] [units] …

As data is placed in this model, content is *not* evaluated. The second slot should contain a valid test name, but anything—even a misspelled word—can appear. For each structured text element (e.g., patient ID, test name), a relational table also exists to relate each possible entry to a number (hence the claim to "encoding"). These tables must contain every possible entry for each element (e.g., all the possible test names for a sodium result, such as "Sodium," "NA," "NA++").
Shortcomings of structured text. A major disadvantage of a structured text database is that all application programs used to access that database must "know" all of the possible data variations in order to process them. Every time new information (e.g., lab tests, drug names, etc.) is added to a structured text database, relational tables and all application programs must be modified to reflect the changes. Structured text also cannot handle more complex, "unpredictable" clinical data (e.g., microbiology results, physician problem lists, radiology findings, etc.), and it is not equipped to provide a controlled, standardized vocabulary that can be shared between facilities. Since each database is essentially "custom-made," a clinical decision support program must also be specifically written to operate on a particular database.

The 3M™ HELP System

3M Health Information Systems and Intermountain Health Care, Inc. developed the 3M HELP System and its data dictionary, the "PTXT" ("Pointer-to-Text") to provide:

- A comprehensive, centralized, "patient-focused" database
- Encoded and structured data that can be analyzed by medical decision support

16 Sager N, Lyman M, Bucknall C, Nhan N, Tick L. Natural Language Processing and the Representation of Clinical Data. J Am Med Informatics Assoc. 1994;1:142-160.

17 Sager.

18 Institute of Medicine Committee on Improving the Patient Record. The Computer-Based Patient Record: An Essential Technology for Health Care. Washington, DC: National Academy Press, 1991.

19 Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a Controlled Medical Vocabulary Server: The VOSER Project. Comput Biomed Res. 1994;27:472-507.

20 Rocha.

processes, which can generate alert messages to appropriate personnel as needed
- A database structure that provides integrated, timely clinical reporting

PTXT: flexible structure and organization. The PTXT file is a hierarchical data dictionary that contains encoded descriptions of all of the medical data gathered in a hospital environment; it contains around 170,000 unique codes, arranged hierarchically from general to specific. About 70,000 of these PTXT codes correlate to all ICD-9-CM codes and the test codes used by the laboratory's computer system. A cross-reference table easily maintains and matches PTXT codes with other external naming systems.

PTXT limitations. The 3M™ HELP System and its data dictionary represented a strong start toward filling the need for computable data and the ability to perform clinical decision support. However—

- PTXT codes can only support five hierarchical levels, and many medical hierarchies are deeper than five levels.
- PTXT codes can be combined to express almost any medical concept, but the system has no rules governing how codes should be combined.
- There are no formal definitions for PTXT codes, and, hence, no way to allow for synonyms (other words that have the same or similar meanings) or homonyms (words that sound/are spelled the same but have different meanings—e.g., "cold").

The UMLS and Medical Entities Dictionary (MED) projects

The MED and the Unified Medical Language System (UMLS) project, described previously, made significant contributions to dictionary development, including:

Semantic networks, links, relationships. Both the UMLS and MED use a semantic network for their representational scheme. In a semantic network, each vocabulary concept is a node; if a concept has synonyms, they will have semantic links to the concept node. All the nodes are connected to each other by semantic relationships, which are phrases that describe the relationship between concepts (e.g., "is a," "is part of," "caused by," etc.). Since a concept may descend from any number of other concepts and also have any number of descendants, there is no limit to the breadth or depth of the hierarchy.

Encoded data and decision support. Terms from four ancillary hospital systems (laboratory, electrocardiography, medical records coding, and pharmacy) have been incorporated into the MED; additional knowledge, expressed as semantic links, has been added for laboratory specimens, tests, and medications. The result is that the MED now "provides medical knowledge that the ancillary systems do not include."[21] For instance, the pharmacy system lets a pharmacist enter a patient's allergy to aspirin, yet also allows aspirin to be ordered for that patient. The MED provides the decision support system with the data needed to flag such a conflict.

The Controlled Medical Vocabulary Server (VOSER) project

The "VOSER" (the Controlled Medical VOcabulary SERver) project contributed expertise to the UMLS effort. Many VOSER researchers had extensive backgrounds in data dictionary development gained from their experience with the 3M HELP System and its PTXT file.[22] However, VOSER far surpassed the PTXT file's capabilities. Two VOSER design principles are critical:

- *Medical event model*—the goal is to accurately code and represent in the database any event that "can occur in the real world."[23]

[21] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based Approaches to the Maintenance of a Large Controlled Medical Terminology. J Am Med Informatics Assoc. 1994;1:35-50.

[22] Rocha.

[23] Huff SM, Rocha RA, Bray BE, Warner HR, Haug PJ. An Event Model of Medical Information Representation. J Am Med Informatics Assoc. 1995;2:116-134.

- *Rule-based linguistic representation*—the premise is that each medical concept must be represented by one unambiguous form.

The medical event model. In VOSER's medical event model, the "key assumption … is that medical information can be represented as a series of linked events."[24] An event can be any one of many diverse activities—the placement of a medication order, the birth of a child, the reading and interpretation of a chest radiograph, or the experience of chest pain.

[24] Huff.

Events and their attributes. Each event is described by its own unique characteristics or attributes; for example, a medication order is characterized by the name of the drug, a dosage, a route of administration, and so on. Naturally, any event involving a patient also has two very important implied characteristics: the patient's identity and the date and time when the event occurred.[25]

[25] Huff.

Event "relationships." Individual event instances are linked to one another in the patient database by particular named relationships. For example, an event such as a chest x-ray performed at 11:00 a.m. on 09-23-01 can be linked in the database when it is compared with a previous chest x-ray performed at 1:00 p.m. on 09-1-01. The semantic relationship "compared to" becomes an attribute of the 09-23-01 chest x-ray. If any additional observations from the earlier x-ray exist, these too can become linked events in the patient database.[26]

[26] Huff.

Advantages of the event model. Thus, by using the medical event as a model both for storing and representing clinical data, the VOSER database logically sequences and stores patient information. At the same time, the event model preserves the time-oriented and associative relationships between medical events. The advantages for healthcare enterprises are obvious. Instead of having "groups" of data from each patient encounter stored separately as they come in from a department, data is organized using patient, source, date, and time tags that become part of the data's description. As a result, applications can address patient care activity as "episodes of care," including all treatment offered a patient for a particular problem.

Rule-based linguistic representation. In VOSER's data dictionary, formal descriptions exist for all medical concepts. For each medical concept, there must be one—*and only one*—unambiguous form representing it. This authoritative representation is achieved in part by identifying smaller concepts that, when combined, convey without ambiguity the meaning of the original concept. These individual components also mean that it is easier to both aggregate and separate concepts. For example, aggregating the concepts of "arm," "upper," and "left" results in the concept of "left upper arm." The VOSER data dictionary also supports these event and vocabulary relationships:

- *Taxonomies*—"is-a" relationships (e.g., penicillin is a drug)
- *Meronomies*—"is-a-part-of" relationships (e.g., a hand is a part of an upper limb)
- *Semantic networks*—causal and other non-hierarchical relationships (e.g., Streptococci causing rheumatic fever)
- *Synonyms*—words that have the same or similar meanings(e.g., "dyspnea" is synonymous with "shortness of breath" and UMLS CUI C0013404)
- *Homonyms*—words that are spelled and pronounced the same, but have different meanings (e.g., "cold") ■

3M

Health Information Systems

Division Headquarters
575 West Murray Boulevard
Salt Lake City, UT 84123
800-367-2447
www.3Mhis.com

Clinical Research Department
100 Barnes Road
Wallingford, CT 06492

Consulting Services
100 Ashford Center North, Suite 200
Atlanta, GA 30338

Printed in USA 09/02; updated 09/05

70-2009-4459-6