2

Healthcare Data, Information, and Knowledge

ELMER V. BERNSTAM • TODD R. JOHNSON • TREVOR COHEN

"...current efforts aimed at the nationwide deployment of health care IT will not be sufficient to achieve the vision of 21st century health care, and may even set back the cause if these efforts continue wholly without change from their present course."[1]

LEARNING OBJECTIVES

After reading this chapter the reader should be able to:

- Define Data, Information, and Knowledge
- Describe how vocabularies convert data to information
- Describe methods that convert information to knowledge
- Distinguish informatics from other computational disciplines, particularly computer science
- Describe the differences between data-centric and information-centric technology

INTRODUCTION

This chapter, will present a framework for understanding informatics. The definitions of data, information, and knowledge were presented in chapter 1 and this chapter will build upon these definitions to answer fundamental questions regarding health informatics. What makes informatics different from other computational disciplines? Why is informatics difficult? Why do some health IT projects fail?

In chapter 1, the fundamental mismatch between available technology (i.e., traditional computers, paper forms) and problems faced by informaticians was mentioned. In this chapter, these ideas are expanded to understand why many health IT (HIT) projects fail. To help organizations appropriately apply HIT, informaticians must understand the limitations of HIT as well as the potential of HIT to improve health.

To illustrate several points, this chapter will begin with a real-world example of challenges at the information level. (See case study on next page.)

DEFINITIONS AND CONCEPTS

Data, Information and Knowledge

In chapter 1, data, information and knowledge (see Figure 1.1) were defined.[4-5] Recall that **data** are observations reflecting differences in the world (e.g., "C34.9"). Note that "data" is the plural of "datum." Thus, "data are" is grammatically correct; "data is" is not correct. **Information** is meaningful data or facts from which conclusions can be drawn (e.g., ICD-10-CM code C34.9 = "Malignant neoplasm of unspecified part of bronchus or lung"). **Knowledge** is information that is justifiably believed to be true (e.g., "Smokers are more likely to develop lung cancer compared to non-smokers"). This relationship is shown in Figure 2.1 and readers will be referred to this diagram later in the chapter.

Data

To understand the relationship between data, information and knowledge in health informatics, readers must understand the relationship between what happens in a computer and the real world. Computers do not represent meaning. They input, store, process and output zero

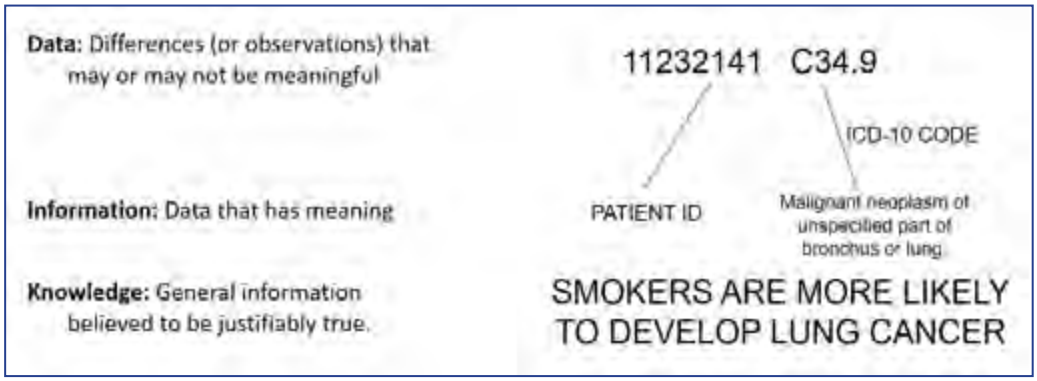

Figure 2.1: Data, information and knowledge

Case Study: The Story of E-patient Dave

In January 2007, Dave deBronkart was diagnosed with a kidney cancer that had spread to both lungs, bone and muscles. His prognosis was grim. He was treated at Beth Israel Deaconess Medical Center in Boston with surgery and enrolled in a clinical trial of High Dosage Interleukin-2 (HDIL-2) therapy. That combination did the trick and by July 2007, it was clear that Dave had beaten the cancer. He is now a blogger and an advocate and activist for patient empowerment.

In March 2009, Dave decided to copy his medical record from the Beth Israel Deaconess EHR to Google Health, a personally-controlled health record or PHR. He was motivated by a desire to contribute to a collection of clinical data that could be used for research. Beth Israel Deaconess had worked with Google to create an interface (or conduit) between their medical record and Google Health. Thus, copying the data was automated. Dave clicked all the options to copy his complete record and pushed the big red button. The data flowed smoothly between computers and the copy process completed in only few moments.

What happened next vividly illustrated the difference between data and information. Multiple urgent warnings immediately appeared, including a warning concerning the prescription of one of his medications in the presence of low potassium levels (hypokalemia) (Figure 2.2). Dave was taking hydrochlorothiazide, a common blood pressure medication, but had not had a low potassium level since he had been hospitalized nearly two years earlier.

Worse, the new record contained a long list of deadly diseases (Figure 2.3). Everything that Dave had ever had was transmitted, but with no dates attached. When the dates were attached, they were wrong. Worse, Dave had never had some of the conditions listed in the new record. He was understandably distressed to learn that he had an aortic aneurysm, a potentially deadly expansion of the aorta, the largest artery in the human body.

Why did this happen? In part, it was because the system transmitted billing codes, rather than doctors' diagnoses. Thus, if a doctor ordered a computed tomography (CT) scan, perhaps to track the size of a tumor, but did not provide a reason for the test, a clerk may have added a billing code to ensure proper billing (e.g., rule out aortic aneurysm). This billing code became permanently associated with the record. To put it another way, the data were transmitted from Beth Israel's computer system to Google's computer system quickly and accurately. However, the meaning of what was transmitted was mangled. In this case, the context (e.g., aortic aneurysm was a billing concept, not a diagnosis) was altered or lost. According to the definitions presented in chapter 1 (and reiterated later in this chapter), meaning is the defining characteristic of information as opposed to data.

After Dave described what happened in his online blog[2] (http://epatientdave.com/), the story was picked up by a number of newspapers including the front page of the Boston Globe.[3] It also brought international attention to the problem of preserving the meaning of data. It became very clear that transmitting data from system to system is not enough to ensure a usable result. To be useful, systems must not mangle the meaning as they input, store, manipulate and transmit information. Unfortunately, as this story illustrates, even when standard codes are stored electronically, their meaning may not be clearly represented.

(off) and one (on). Each zero or one is known as a **bit.** A series of eight bits is called a **byte**. Note that these bits and bytes have no intrinsic meaning. They can represent anything or nothing at all (e.g., random sequences of zeroes and ones).

Bits within computers are aggregated into a variety of **data types**. Some of the most common data types are listed below.

- *Integers* such as 32767, 15 and -20
- *Floating point numbers* (or floats) such as 3.14159, -12.014, and 14.01; the floating point refers to the decimal point
- *Characters* "a," and "z"
- *(Character) Strings* such as "hello" or "ball"

Note that these data types do not define meaning. A computer does not "know" whether 3.14159 is a random number or the ratio of the circumference to the diameter of a circle (known as Pi or π).

Data can be aggregated into a variety of file formats. These file formats specify the way that data are organized within the file. For example, the file header may contain the colors used in an image file (known as the palette) and the compression method used to minimize storage requirements. Common or standardized file formats allow sharing of files between computers and between applications. For example, as long as your digital camera stores photos as JPG files, you can use any program that can read JPG files to view your photos.

- Image files such as JPG, GIF and PNG.
- Text files
- Sound files such as WAV and MP3
- Video files such as MPG

Again, it is important to recognize that neither data types nor file formats define the meaning of the data, except for the purpose of storing or display on a computer. For example, photographs of balloons and microscopes can be stored in JPG files. Nothing about the file format helps us recognize the subject of the photograph.

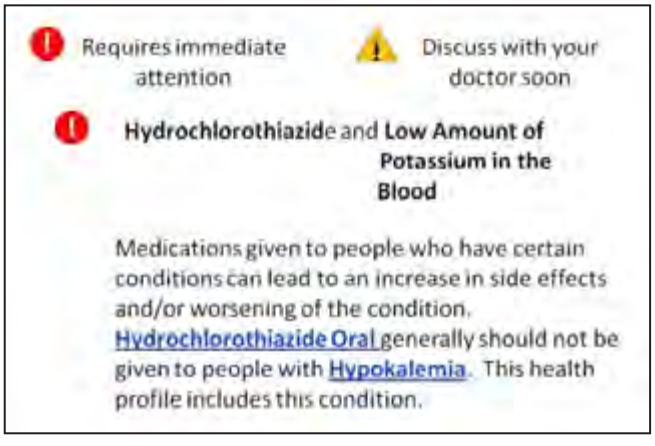

Figure 2.2: Urgent warning in e-patient Dave's record

Informatics vs. Information Technology and Computer Science

Data are largely the domain of information technology (IT) professionals and computer scientists. As computers become increasingly important in biomedicine, biomedical researchers are starting to collaborate with computer scientists. IT professionals and computer scientists concentrate on technology, including computing systems composed of hardware and software as well as the algorithms implemented in such systems. For example, computer scientists develop algorithms to search or sort data more efficiently. Note that *what* is being sorted or searched is largely irrelevant. In other words, the meaning of the data is of secondary importance. It does not matter whether the strings that are being sorted represent proper names, email addresses, weights, names of cars or heights of buildings.

Though they may be motivated by specific applications, computer scientists typically develop general-purpose approaches to classes of problems that involve computation. For example, a computer scientist may design a memory architecture that efficiently stores and retrieves large data sets. The computer science contribution is the

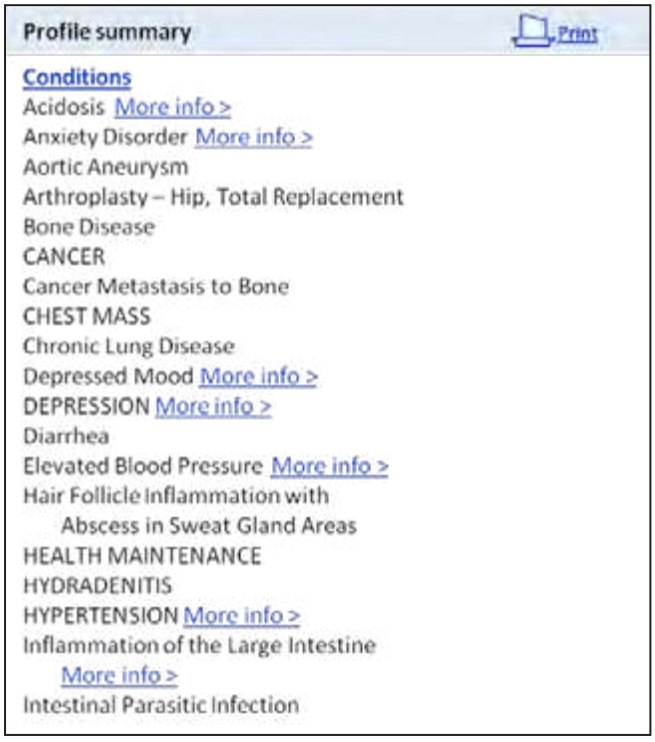

Figure 2.3: e-patient Dave's conditions as reflected in the newly-created personal health record (PHR)

development of the better memory architecture for large data sets; while the memory architecture is not a direct improvement of an EHR per se, it is nonetheless critical to its advancement.

Information and knowledge, on the other hand, are addressed by informatics. To an informatician computers are tools for manipulating information. Indeed, there are many other useful information tools, such as pens, paper and reminder cards. There are significant advantages to manipulating digitized data, including the ability to display the same data in a variety of ways and to communicate with remote collaborators. From an informatics perspective however, one should choose the optimal tool for the information task – often, but not always, the best tool for the task is computer-based.[4,6]

There are areas of overlap between computer science and informatics. For example, information retrieval is widely viewed as a sub-field of computer science and information retrieval researchers often reside in computer science departments. However, we would argue that information retrieval draws on both disciplines. Information retrieval is "*finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need by retrieving documents from large collections (usually stored on computers).*"[7]

Note that information retrieval is concerned with retrieval of information, not data. For example, finding documents that describe the relationship between aspirin and heart attack (myocardial infarction) is an example of an information retrieval task. The central problem is identifying documents that contain certain meaning. In contrast, efficient retrieval of documents (or records) that contain the string "aspirin" can be posed as a database problem (an area of computer science). Importantly, informatics and computer science differ in the problems that they address (see Figure 2.4). It should not be implied that computer science is easier or less intellectually challenging compared to informatics (or vice versa).

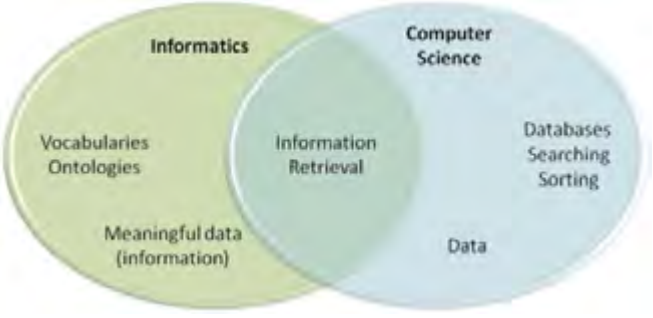

Figure 2.4: Relationship between informatics and computer science in area of information retrieval.

Increasingly, the term "data scientist" is used to refer to professionals engaged in the retrospective analysis of incidentally collected data (such as the online activity of users of a website). With biomedical data, effective analysis often requires the attribution of meaning to such data. So, we would argue, the biomedical data scientist must take both information and raw data into account (i.e., engage in informatics and not solely data analysis).

Artificial Intelligence (AI)

AI is generally considered to be a sub-field of computer science. This is arguably appropriate (particularly for the current crop of AI systems) since the focus is usually on the development of generalizable methods through which a computer can exhibit behavior that appears intelligent. AI is concerned with the development of systems that can do something that previously required human intelligence, such as driving a vehicle in city traffic, winning a game of chess, or solving a logic puzzle. Originally, AI developed in parallel with studies of human expert cognition, and a prevailing notion was that the simulation of intelligent behavior required systems based on knowledge of how expert humans solve problems in a given domain. However, the focus subsequently shifted away from the design of systems that use human-like processes, to the development of systems that can attain human-like performance regardless of how this performance is obtained (i.e., without simulating human cognition or expertise). Recent advances in statistical AI or "*machine learning*" have enabled computers to solve problems that have previously resisted automation. Specifically, "*deep learning*" refers to the use of multi-layer neural networks to learn patterns such as the features of objects in an image. A particularly prominent success in the biomedical domain has been in the field of dermatologic (skin) lesion categorization. Researchers at Stanford University were able to use a very large set of labelled images (129,450) showing various types of skin lesions to train a computer to distinguish specific kinds of malignant (cancerous) lesions from similar benign (non-cancerous) lesions. Importantly, the system was provided only pixel-level data and labels, no attempt was made to provide the system with any knowledge about how to recognize any dermatologic disease. System performance was compared against 21 board-certified dermatologists. The system performed comparably to the dermatologists.[8] This was an impressive and potentially clinically-useful application. This project also illustrates two limitations of deep learning. First, that it requires large sets of labelled data to "train" the system. Second, that the system cannot explain "why" it does something

to a human. Other applications of deep learning that may impact biomedicine include natural language processing and speech recognition.

CONVERTING DATA TO INFORMATION TO KNOWLEDGE

We live in the real world that contains physical objects (e.g., *aspirin tablet*), people (e.g., *John Smith*), things that can be done (e.g., John Smith *took* an aspirin tablet) and other concepts. To do useful computation in this context, one must segregate some part of the physical world and create a **conceptual model**. The conceptual model contains only the parts of the physical world that are relevant to the computation. Importantly, everything that is not in the conceptual model is excluded from the computation and assumed to be irrelevant.

The conceptual model is used to design and implement a **computational model**. In Figure 2.5, the real world contains a person, John Smith. There are many other things in the real world including other people, physical objects, etc. There are many things that we can say about this person, they have a name, height, weight, parents, thoughts, feelings, etc. The conceptual model defines what is relevant; everything that is not in the conceptual model is therefore assumed to be not relevant. In our example (Figure 2.5), name and age are chosen. Thus, the height, weight and all other things about John Smith are assumed to be irrelevant. For example, given our conceptual and computational models, one would not be able to answer questions about height. Next a **representation** must be defined. (Figure 2.5). A simple example is that of whole numbers. A representation has three components. The **represented world** is the *information* that one wants to represent (e.g., whole numbers: 0, 1, 2, 3, …). The **representing world** contains the data that represent the information (e.g., symbols "0", "1", "2", "3", …). There must be a **mapping** between the represented world and the representing world. In our example, the mapping is the correspondence between whole numbers and symbols that are used to represent them. Note that the data are, in and of themselves, meaningless.

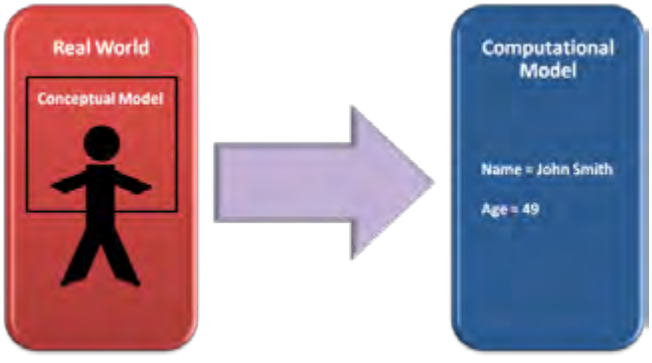

Figure 2.5: Computational framework

To do anything useful, one must also have rules regarding the mapping (i.e., relationship between the symbols and the real world), and what can be done with the symbols. In our example, these rules are the rules governing the manipulation of whole numbers systems (e.g., addition, multiplication, division, etc.).

The data part of a representational system may also be called its "form", in which case meaning is called its' "content." The word "form" is significant because of its relationship to **formal methods**, which are methods that manipulate data using systematic rules that depend only on form, not content (meaning). These formal methods, including computer programs, depend only on systematic manipulation of data without regard for meaning. Thus, only a human can ensure that the input and output of a formal method (e.g., computer program) correctly capture and preserve meaning. In the skin cancer example, deep learning network described above, the humans who designed or used the network know that the input represents digitized images of skin lesions and that the output represents whether the lesion is cancerous or not. However, the trained network knows nothing about lesions or cancer, it is simply a complex non-linear mathematical function mapping input (digitized images) to output (cancerous vs. non-cancerous). Recent research on deep neural nets has shown that they can be reliably fooled into confidently misclassifying images by adding noise that humans cannot perceive. For instance, researchers can add noise to an image of a panda to create a second image that to humans is indistinguishable from the first, but that causes a deep net to confidently classify the first image as a panda and the second as a gibbon.[10] Of course, there's by now a far larger literature on situations in which human diagnosticians reliably make mistakes.[11] That human and machine diagnosticians reach their conclusions through different processes suggests that they will make different sorts of errors, and that the safest system may be one that takes both of their perspectives into account.

In spite of the fact that formal methods manipulate only form (or data), not meaning, they can be very useful. If the formal method does not violate the rules of the physical world, one can apply the method to solve problems in the real world. For example, a whole number representation can be used to determine how many 8-person boats are needed to transport 256 people across the Nile river (i.e., 256 people divided by 8 people/boat = 32 boats).

However, one must be careful because the formal method (division) can easily violate the rules of the real world. For example, suppose that 250 people are in Cairo and six people are in Khartoum (1,000 miles away) and they must cross at the same time. In this case, 32 boats is the wrong answer since 32 boats are needed in Cairo and another boat is needed at Khartoum. In this example, the real world includes location (Cairo vs. Khartoum), but the conceptual model includes only the number of people; location and distance are ignored. Thus, the computational model (based on the conceptual model) gives an inappropriate answer. It can't be said that the answer is "wrong." Clearly 256/8 = 32; the computer did not malfunction. However, in the case where location is important, the numerical answer is not useful.

The distinction between the real (represented) world, the conceptual model (representing world) and the computational model (that which the computer manipulates) is fundamental to informatics.

When the real world, the conceptual model and the computational model match, it is possible to get useful answers from the computer. When they do not match, such as the case when a critical constraint was left out of the conceptual model, the answers obtained from the computer are not useful. This is what happened in the case of e-patient Dave. Formal methods (computer programs) were developed that linked fields in the Beth Israel Deaconess EHR to fields in Google Health. Data from one were dutifully transferred to the other. However, the meaning (i.e., that the data being transmitted were billing codes, not actual diagnoses) was lost. Further, there was a flaw in the conceptual model, the computational model or both models that prevented dates from being maintained correctly; perhaps because the dates reflected billing dates, rather than the date when a diagnosis was made.

Data to Information

The next step is to convert data into information. Consider the example in Figure 2.1. "C34.9" is, in and of itself, meaningless (i.e., it is a data item or datum). However, ICD-10-CM gives us a way to interpret C34.9 as "*Lung neoplasm, not otherwise specified*." Thus, the vocabulary ICD-10-CM turns the datum into a unit of information.

The computer still stores only data, not information. Thus, only a human can determine whether the meaning is preserved or not. In the case of e-patient Dave, all the computer systems functioned as they were designed. There were no "*computer errors*," but upon human review, the meaning was mangled.

However, associating ICD-10-CM C34.9 with a patient record labels the patient record (and thus the patient) as having "*Lung neoplasm, not otherwise specified*." Of course, one could design systems that turn data into information without using vocabularies. For example, patient records could be designed that include a bit for each possible diagnosis. Thus, setting the bit corresponding to lung cancer to 1 would be **semantically equivalent** to associating ICD-10-CM C34.9 with the patient's record. Semantically equivalent is simply another way of stating that the meanings are the same.

Transmission of information between computer systems, often referred to as **interoperability**, requires consistency of interpretation in the context of a particular task or set of tasks.[12] The source system (Beth Israel Deaconess EHR for e-patient Dave) and the receiving system (Google Health for e-patient Dave) must share a common way of transforming data into information. However, this is not sufficient. Note that in the case of e-patient Dave, both systems used ICD codes. However, associated information such as dates and most importantly the context: billing code vs. actual diagnosis, was not shared correctly.

Information to Knowledge

Multiple methods have been developed to extract knowledge from information. Note that it would not make sense to directly convert data (which by definition are not meaningful) to knowledge (justified, true belief). Thus, information is required to produce knowledge. Transformation of information (meaningful data) into knowledge (justified, true belief) is a core goal of science.

In the clinical world, most available knowledge is best described as justified (i.e., evidence exists that it is true), rather than proven fact (i.e., it must be true). This is an important distinction from traditional hard sciences such as physics or mathematics.

In this chapter, there is a focus on informatics techniques that are designed to convert clinical information into knowledge. Thus, clinical data warehouses (CDWs) are described that are often the basis for attempts to turn clinical information into knowledge, as well as methods for transforming information into knowledge.

Clinical research informatics is recognized as a distinct sub-field within informatics (see separate chapter on e-research for further information). Clinical research informaticians leverage informatics to enable and transform clinical research.[13-14] By "enable," what is meant is helping researchers accomplish their goals faster and cheaper than is possible using existing methods. For example, searching electronic clinical data may be

faster than manually reviewing paper clinical charts. "Transform" means developing methods that allow researchers to do things that they could not do using existing methods. For example, it is not possible to use aggregated clinical data contained in paper records to help clinicians make decisions in real time. One cannot ask, in real-time or near real-time, "*what happened to patients like me, at your institution, who chose treatment A vs. treatment B?*" Although the information required to answer this question is found in the clinical records, a manual chart review cannot be performed in real time. However, to derive knowledge from information and realize the benefits of computerized information, we must ensure that meaning is preserved.

CLINICAL DATA WAREHOUSES (CDWS)

The enterprise data warehouse was introduced in chapter 1. In this chapter, the focus will be on clinical, rather than administrative data, hence the reference to a **clinical data warehouse** or **CDW**.

Increasingly, clinical data are collected via electronic health records (EHRs). Clinical records within EHRs are composed of both **structured data** and **unstructured or (free text)**. Structured data may include billing codes, lab results (e.g., Sodium = 140 mg/dl), problem lists (e.g., Problem #1 = ICD-10-CM C34.9 = "*Lung Neoplasm, Not Otherwise Specified*"), medication lists, etc. In contrast, free text is similar to this chapter – simply human language such as English, called **natural language**. Although templates are often used, key portions of clinical notes are still often dictated and are represented in records as free text.

From an informatics perspective, structured data are much easier to manage – structured data are computationally tractable. Ideally, but not always, these data are encoded using a standard such as ICD-10-CM (previously ICD-10-CM in the United States, see chapter on data standards). Thus, retrieving patients with a particular problem is, theoretically, simply a matter of identifying all records that are tagged with a particular code. As one will see later in this chapter, in practice this does not always work. Further, nuances (e.g., similarity to a previous case) or vague concepts (e.g., light-colored lesion, tall man) may be difficult to convey with a "*one size fits all*" vocabulary.

Similarly, computerized physician order entry (see chapter on electronic health records) can be difficult to implement. If designers allow only structured data, they must anticipate what will be ordered and make choices that constrain the possible inputs. For example, they may choose to use a particular vocabulary for medication orders, allow specific dosing frequencies, etc. Inevitably, however, physicians will want to write unusual orders that will be difficult to accommodate.

Free text, on the other hand, has the advantage of being able to express anything that can be expressed using natural language. On the other hand, it is difficult for computers to process. Indeed, the field of **natural language processing** (NLP) is an active area of research in both computer science and informatics. Within clinical records, the free text notes are critically important. Indeed, as in the case of e-patient Dave, structured data (such as billing codes) may not accurately reflect clinical reality. This is not necessarily anyone's fault. Billing codes were assigned for billing, not for clinical care. Thus, it should not be surprising that using billing codes for a different purpose does not yield the desired result. Over 20 years ago, van der Lei warned:

> *...under the assumption that laws of medical informatics exist, I would like to nominate the first law: Data shall be used only for the purpose for which they were collected. This law has a collateral: If no purpose was defined prior to the collection of the data, then the data should not be used.*[15]

To make sense of clinical records, both structured data and free text must be leveraged. This remains an active area of informatics research.

A clinical data warehouse is a database system that collects, integrates and stores clinical data from a variety of sources including electronic health records, radiology and other information systems. EHRs are designed to support real-time updating and retrieval of individual data (e.g., Joan Smith's age). The general process is shown in Figure 2.6. Data from multiple sources including one or more EHRs are copied into a staging database, cleaned and loaded into a common database where they are associated with **meta-data**. Meta-data are data that describe other data. For example, the notation that a data item is an ICD-10-CM term represents meta-data.

Once loaded into a CDW, a variety of analytics can be applied, and the results presented to the user via a user interface. Examples of simple analytics include summary statistics such as counts, means, medians and standard deviations. More sophisticated analytics include associations (e.g., does A co-occur with B) and similarity determinations (e.g., is A similar to B).

In contrast to EHRs, CDWs are designed to support queries about groups (e.g., average age of patients with breast cancer). Although in principle an EHR may contain the same data as a CDW, databases that support EHRs are designed for efficient real-time updating and retrieval

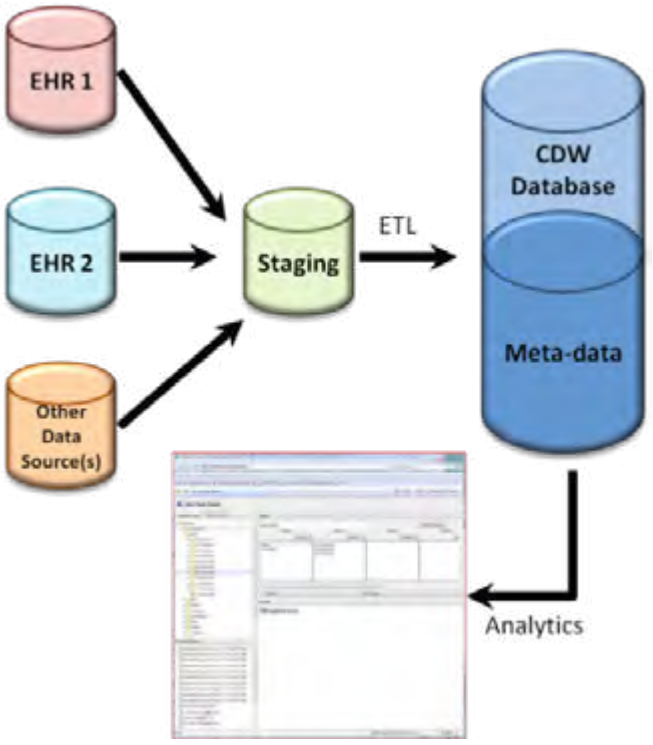

Figure 2.6: Overview of clinical data warehousing (ETL = Extract, transform and load)

of individual data. Thus, a query across patients rather than regarding an individual may take much more time. Further, since EHRs support patient care, queries about groups may be restricted to ensure adequate performance for clinicians. Another important distinction is that CDWs are usually not updated in real-time. Although update schedules differ, daily or weekly updates of the institutional CDW are typical.

CDWs are rapidly becoming critical resources. They enable organizations to monitor quality by allowing users to query for specific quality measures (see chapter on quality improvement strategies) in specific patient populations (e.g., retrieve all women who are 40 years old or older who have not had a mammogram in the past year). Similarly, clinical and translational researchers use CDWs to identify trends (e.g., did screening mammograms detect breast cancer at an early stage?).[16] Comparative effectiveness research (CER) or, more broadly, practice-based research, are increasingly important fields that attempt to link research with clinical practice using CDWs. They complement traditional clinical trials that ask very focused questions. For example, a clinical trial might be designed to compare treatment A vs. treatment B in a particular population of patients. In contrast, CER practitioners ask what happened in practice. For example, treatment A has been found to be more effective than treatment B in a clinical trial. What actually happened in practice?

Hospital infection control specialists use CDWs to track pathogens within hospitals. Public health agencies traditionally rely on reporting to conduct surveillance for natural or man-made illnesses (see chapter on public health informatics). However, reporting introduces a delay. Accessing aggregated data at the institutional level can be done much faster using a CDW.

One of the most popular clinical data warehousing platforms is the product of the Informatics for Integrating Biology and the Bedside (i2b2) project based at Harvard Medical School.[17] The open source and very modular i2b2 platform was designed to enable the reuse of clinical data for research but can also be very useful for non-research tasks such as quality monitoring. As of August 2017, i2b2 has been implemented at over 100 institutions including academic institutions and commercial entities in the US and abroad.[18]

I2b2 relies on a star schema composed of facts and dimensions (Figure 2.7.). *Facts* are pieces of information that are queried by users (e.g., diagnoses, demographics,

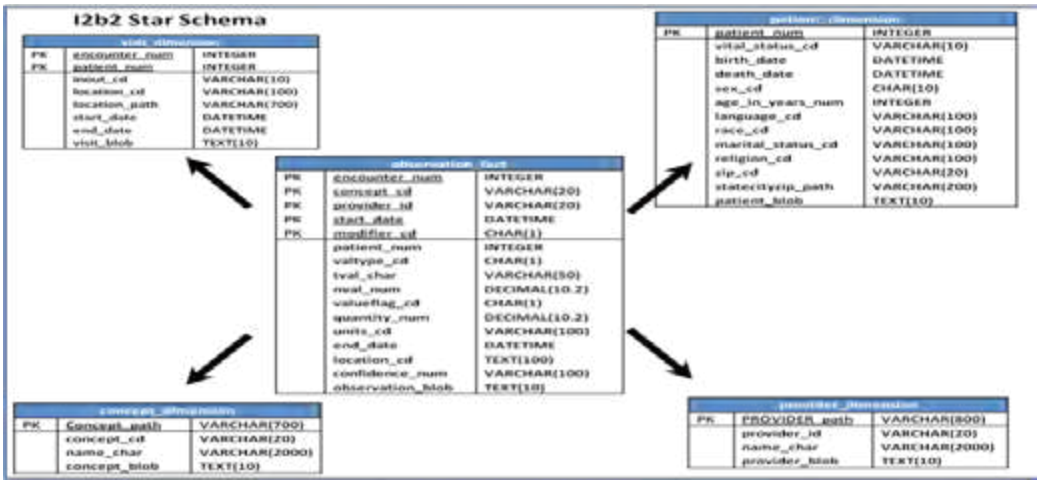

Figure 2.7: i2b2 data model 12

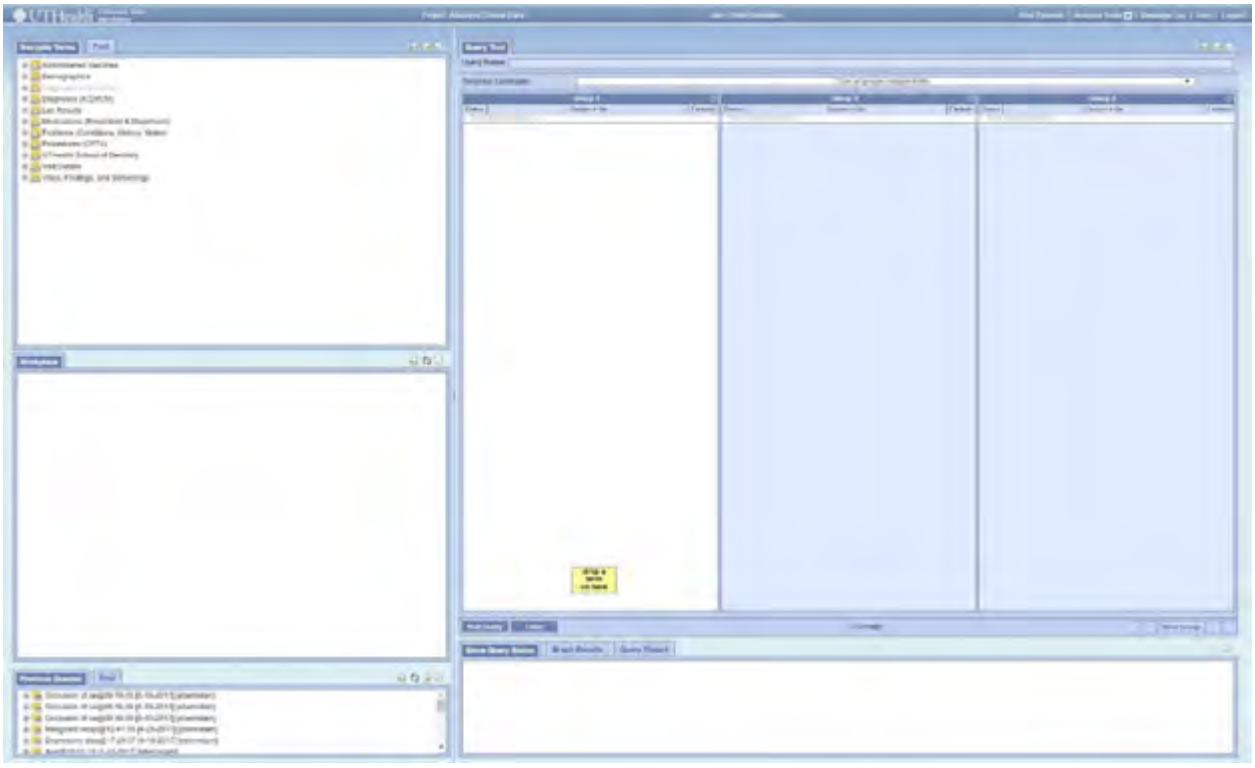

Figure 2.8: i2b2 screenshot showing the result (patient count) of a query for female patients ages 45-64 with ischemic heart disease

laboratory results, etc.) and *dimensions* describe the facts. Note that the data model is organized around facts, rather than individual patients, as would be the case for an EHR. Another benefit of organizing the CDW around observations is that data from multiple sources (e.g., different hospitals) can be aggregated into a common data model – new observations are simply added to the table of facts. Meta-data, such as the vocabulary that was used for encoding the fact, is an important component. Thus, the i2b2 data model by itself is not sufficient to ensure interoperability.

I2b2 also provides a very usable interface to an institutional CDW that can be used by non-informaticians (see Figure 2.8). Users click and drag concepts from the ontology window (upper left) into the query panes (upper right) and obtain results, such as the number of patients fulfilling certain criteria, in lower right. In addition to the basic i2b2 package, specialized modules have been developed for NLP and other tasks.

In short, clinical data are collected via EHRs and archived in CDWs. As EHRs are becoming increasingly common, CDWs are becoming increasingly important. However, to realize the potential of CDWs to improve health, we must do more than archive data. One must turn these data into information and knowledge. Users must be able to "*make sense*" of clinical data; to make clinical data meaningful (data → information) and then learn from aggregated clinical data (information → knowledge). In practice, many of the benefits of EHRs (see chapter 3) actually require a CDW. The transformation of data into information and knowledge is a core concern of informaticians.

Use of Aggregated Clinical Data

To make use of aggregated clinical information, we must be able to recognize records that belong to patients with specific conditions. For example, it is necessary to identify records belonging to patients who have been diagnosed with breast cancer. A simple answer is to rely on billing codes, one of the most common forms of structured data in clinical records. However, as we saw in the case of e-patient Dave, one cannot simply rely on billing codes. Sometimes other structured data are available, problem lists are particularly useful. Unfortunately, problem lists are often out of date or incomplete.[19] Thus, a great deal of interest has focused on extracting information from free text clinical notes.

Concept extraction refers to the problem of identifying concepts within unstructured data, such as discharge summaries or pathology reports. Usually, these concepts are mapped to a controlled vocabulary,

such as ICD-10-CM, SNOMED-CT and others. While this may on the surface appear to be a trivial problem, there are many ways in which a single concept might be expressed (for example "high blood pressure" and "hypertension"), and it is often the case that a single word or acronym may have multiple medically relevant meanings (for example "DM" may refer to "Diabetes Mellitus" or "Depressed Mood") that cannot be teased apart without considering contextual cues. Consequently, much effort has been devoted toward the development of systems that aim to map between terms or phrases and controlled vocabularies with accuracy.

Multiple biomedical concept extraction systems exist including MetaMap[20] and cTAKES.[21] Broad-purpose medical language processing systems such as MedLEE,[22] have also been adapted to this end. These systems can be tuned to perform well but require re-tuning when applied to different corpora (e.g., changing institutions) or clinical problems (e.g., breast cancer vs. diabetes mellitus). Table 2.1 summarizes the published performance of these three concept extraction systems; note that the results are not directly comparable to each other due to different tasks, experimental design (e.g., pre-processing), and gold standards (a common limitation).[23-24]

Table 2.1: Published performance of three notable biomedical systems

Concept Extractor	Gold Standard	Precision	Recall	F-score (F1)
cTAKES[21]	Mayo clinic	0.80	0.65	0.72
MetaMap (MMTx)[25]	Proprietary	0.74	0.76	0.748
MEDLEE[26]	Proprietary	0.86	0.77	0.81

Classification refers to the problem of categorizing data into two or more categories. For example, one might want to classify medical records as belonging to patients who have vs. have not been diagnosed with breast cancer. A variety of classification algorithms have been developed, most of which rely on statistical methods. These classification algorithms generally depend on the selection of a set of features, such as the presence or absence of particular terms, concepts or phrases. Once these features have been selected, either manually or through automated methods, medical records can be categorized based on these features. A commonly utilized approach is supervised machine learning, in which an algorithm is used to learn a representation of the features that characterize annotated positive (patients with breast cancer) and negative (patients without breast cancer) cases. New cases can then be categorized automatically based on the extent to which their features are characteristic of previously encountered positive or negative examples. Deep learning using multi-layer neural networks described above, is an example of supervised learning approaches.

WHAT MAKES INFORMATICS DIFFICULT?

Why are some domains highly computerized, while health care and biomedicine resist computerization? Consider the banking system.[4] It is clearly very complex and involves a vast quantities data and meaning. Why do all banks use computers? In contrast to health care, there are no arguments regarding the suitability of computers to track accounts. We argue that in the case of banking, there is a very narrow "*semantic gap*" between data and information. In other words, the correspondence between the data (numbers) and information (account balances) is very direct. As one manipulates the computational model, the meaning of these manipulations follows easily.

Consider the differences between banking data and health care data, such as an account at a bank versus a patient (Table 2.2). One difference is that concepts relevant to health are relatively poorly defined compared to banking concepts. The symbols require significant background knowledge to interpret properly. For example, there are multiple ways that a patient can be "sick" including derangements in vital signs (e.g., extremely high or low blood pressure), prognosis associated with a diagnosis (e.g., any patient with an acute aortic dissection is sick), or other factors. Two clinicians when asked to describe a "sick" individual may legitimately focus on different facts. In contrast, a bank account balance (e.g., $1058.93) is relatively objective and is captured by the symbols. Thus, data-manipulating machines (IT) are much better suited to manipulating bank accounts than clinical descriptors.

In general, if the problem relates strictly to form (data) or is easily reduced to a form-based problem, then computers can easily be applied to solve the problem. Retrieving all abstracts in PubMed containing the string "*breast cancer*" is a question related to data and is easily reducible to a form-based data query. On the other hand, retrieving all documents that report a positive correlation between beta blockers (a class of medications) and weight gain is an information retrieval question that depends on the meaning of the query and the meaning of the text in

Table 2.2: Comparison of health and banking data

	Banking data	Health data
Concepts and descriptions	Precise *Example:* Account 123 balance = $15.98	General, subjective *Example:* sick patient
Actions	Usually (not always) reversible *Example:* Move money A B	Often not easily reversible *Example:* Give a medication Perform procedure
Context	Precise, constant *Example:* US $	Vague, variable *Example:* Normal lab values differ by lab
User autonomy	Well-defined and constrained *Example:* What I can do with my checking account = what you can do	Variable and dependent on circumstance *Example:* Clinical privileges depend on training, change over time, depend on circumstances
Users	Clerical staff	Varied, including highly trained professionals
Time sensitivity	Few true emergencies (seconds)	Many time sensitive tasks, highly variable time sensitivity depending on context
Workflow	Well-defined	Highly variable, implicit

the documents. The latter question is not easily reducible to form and is therefore much harder to automate.

Concepts definable with necessary and sufficient conditions are usually relatively easy to reduce to form, and thereby permit some limited automated processing of meaning. However, concepts without necessary and sufficient conditions (e.g., recognizing a sick patient, or defining pain) cannot be easily reduced to data and are much more difficult to capture computationally. Informatics is interesting (and difficult), in part, because many biomedical concepts defy definition via necessary and sufficient conditions.

Blois argued that, to compute upon a system, one must first determine the system's boundaries.[27] In other words, one must define all the relevant components and assume that everything else is irrelevant. However, this is very difficult to do for biological (or human) systems. If the goal is to model the circulatory system, can the renal system be excluded? The endocrine system that includes the adrenal glands (releases epinephrine that constricts blood vessels and raises blood pressure)? The nervous system? And so on. With a bank account, it is easy to draw boundaries around the real-world concepts that affect an accurate account balance. On the other hand, in biomedicine these boundaries are often impossible to precisely define, so our conceptual and computational models are rarely complete and often lead to inaccurate results, such as was seen with e-Patient Dave.

COMPLEXITY OF KNOWLEDGE MODELS

Modeling health care is difficult, but this has not stopped informaticians from trying. Notable modeling attempts include the HL7 Reference Information Model or RIM (see chapter on data standards). Work on the RIM started in 1997 and Release 1 was approved by the American National Standards Institute (ANSI) in 2003. The RIM is one of the major differences between the commonly adopted HL7 version 2.x that has been widely used for decades and version 3, which has not been as

widely adopted.[28] One of the problems is that the RIM is very complex (see Figure 2.9) and does not necessarily match all health care environments.

Biomedical informatics is also difficult because biomedical information can be imperfect in several different ways:

- Incomplete information: Information for which some data are missing, but potentially obtainable.
  - Example: What is the past medical history of an unconscious patient who arrives at ED?
- Uncertain information: Information for which it is not possible to objectively determine whether it is true or false. This can also be called epistemic uncertainty, because it arises from a lack of knowledge of some underlying fact. This type of imperfection is addressed by probability and statistics.
  - Example: how many female humans are in the US? Although there is a precise answer to this question at any given moment, we can only estimate the answer using statistics.
- Imprecise information: Information that is not as specific as it should be.
  - Example: Patient has pneumonia. This may be precise enough for some purposes but is not sufficiently precise to determine treatment. For example, antibiotics can treat bacterial pneumonia, but are of little use to a patient with viral pneumonia.
- Vague information: Information that includes elements (e.g., predicates or quantifiers) that permit boundary cases (tall woman, may have happened, large bruise, big wound, elderly man, sharp radiating pain, etc.). Unlike uncertain information, with vague information there is no underlying matter of fact. Even if the age of every female human in the US was known, one could not precisely answer the question of how many mature women were in the US at that time, because "mature" is a term that has boundary cases; there are women who are clearly mature, those who clearly are not, and a number in between for whom one cannot be sure that term applies.
- Inconsistent information: Information that contains two or more assertions that cannot simultaneously hold.
  - Example: Birthdate: 8/29/66 AND 9/17/66

As illustrated in the above examples, all these imperfections may be found in healthcare information. Humans can deal with these imperfections. For example, it can be decided that for clinical purposes, a difference in

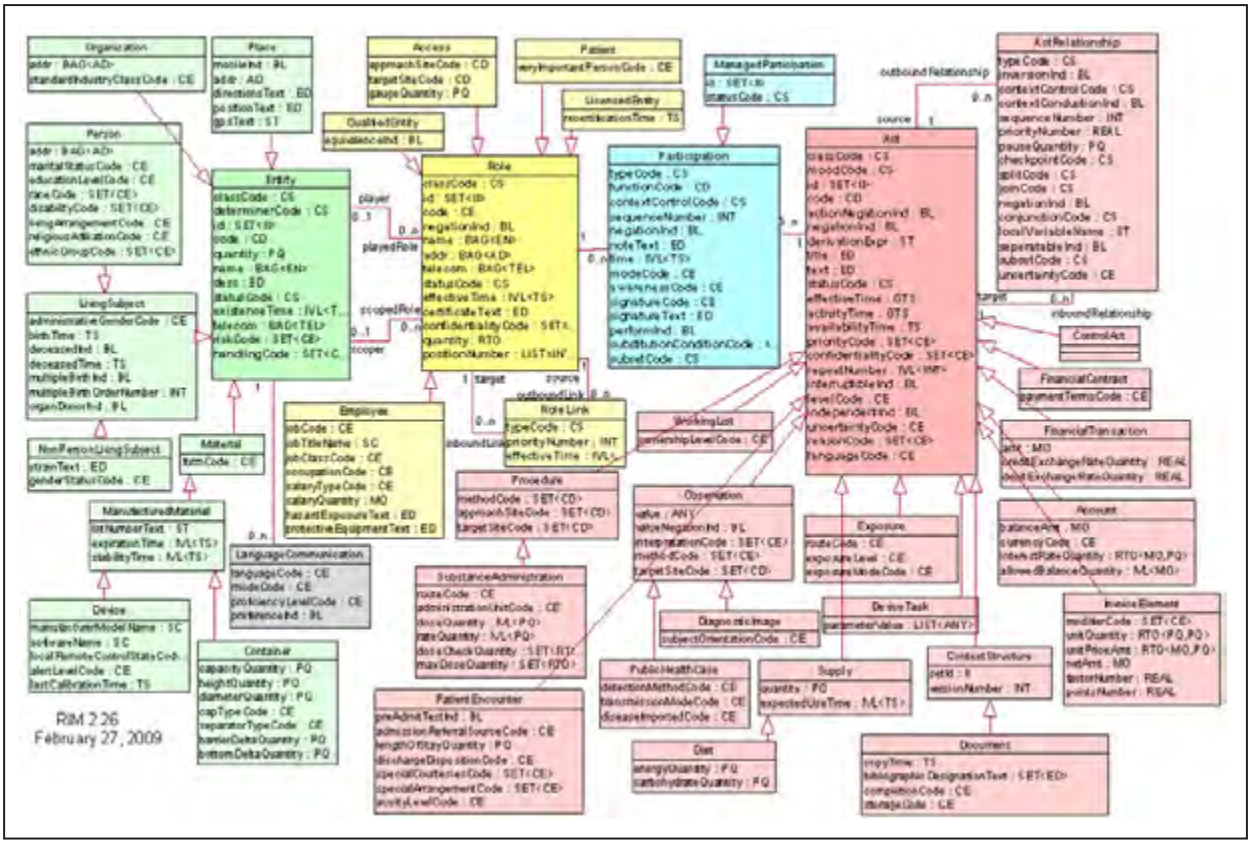

Figure 2.9: Overview of the HL7 version 3 RIM (Courtesy HL7 [29])

patient age of a little over two weeks (a vague statement), is insignificant for clinical purposes. Computers, on the other hand, must be explicitly programmed to make such "judgments." However, the number of possible variances and exceptions is effectively infinite. Thus, they cannot all be anticipated and addressed in advance. This is one reason why clinical decision support often gives advice that is, to a clinician, obviously inappropriate to the current patient situation.

In addition, definitions in health care and biomedicine often change over time. Consider the definition of a gene.[30]

Designing systems that adapt to changes in definition that, in turn, can affect other definitions is difficult. Our computers and programming languages process discrete symbols according to precise formal rules or mathematical expressions. They do not make sense of a highly ambiguous, noisy world or do meaning-based processing. With this background, one can now consider health IT and its various successes and failures in the real world.

WHY HEALTH IT SOMETIMES FAILS

> *"To improve the quality of our health care while lowering its cost, we will make the immediate investments necessary to ensure that within five years all of America's medical records are computerized. This will cut waste, eliminate red tape, and reduce the need to repeat expensive medical tests… it will save lives by reducing the deadly but preventable medical errors that pervade our health care system." —Barack Obama (Speech on the Economy, George Mason University, January 8, 2009)*

Widespread dissatisfaction with health care in America and rapid advancement in information technology has focused attention on Health IT (HIT) as a possible solution. The need for HIT is one of the few topics upon which Democrats and Republicans agree. Both former President Bush and President Obama set 2014 as the goal date for computerizing medical records. To many, HIT seems like an obvious solution to our health care woes. The government's HIT website says that HIT adoption will: improve health care quality, prevent medical errors, reduce health care costs, increase administrative efficiencies, decrease paperwork and expand access to affordable care.[9] However, there is increasing evidence that HIT adoption does not guarantee these benefits. Unmitigated enthusiasm is dangerous for HIT adoption. Similar enthusiasm repeatedly threatens the field of artificial intelligence, resulting in cycles of excitement and disappointment (in artificial intelligence, these cycles are sometimes called "AI winters").

Effects of HIT

HIT is an "*easy sell*" to an American public increasingly dissatisfied with our health care system. Indeed, there is evidence that HIT can improve health care quality, prevent medical errors, and increase efficiency.[31-32] Thus, there is reason for optimism. With the American Recovery and Reinvestment Act (ARRA) of 2009, the US government made a multi-billion-dollar investment in HIT.[33] Similar investments have been made by the governments of Australia,[34] Belgium,[35] Canada,[36] Denmark,[37] and the United Kingdom.[38]

However, many and perhaps even most HIT projects fail.[39] There is also evidence that HIT can worsen health care quality to the point of increasing mortality, increasing errors and decreasing efficiency.[40-42] In November 2011, the Institute of Medicine issued a report entitled "*Health IT and Patient Safety: Building Safer Systems for Better Care*" that concluded: "*...some products have begun being associated with increased safety risks for patients.*"[43] There is even a term, "e-iatrogenesis," that refers to the unintended deleterious consequences of HIT.[44] Notably, systems that increase mortality at one institution , do not seem to have the same effect at another institution;[40,45] even though the clinical setting (pediatric intensive care) was similar. Thus, one cannot simply conclude that the system itself is wholly responsible. It is not just the system being implemented, but how it is implemented and in what context that affects the clinical outcomes.

We've Been Here Before: AI Winters

During the 1950s, we were faced with a different problem: the Cold War. Similarly, the government saw IT as a promising (at least partial) solution. If researchers could develop automated translation, we could monitor Russian communications and scientific reports in "real time." There was a great deal of optimism and "*…many predictions of fully automatic systems operating within a few years.*"[46]

Although there were promising applications of poor-quality automated translation, the optimistic predictions of the 1950s were not realized. The fundamental problem of context and meaning remains unsolved. This made disambiguation difficult resulting in amusing failures. Humorous examples include: "*the spirit is willing but the flesh is weak*" translated English → Russian → English

resulted in the phrase "*the vodka is good but the meat is rotten*."

In 1966, the influential Automatic Language Processing Advisory Committee (ALPAC) concluded that "*there is no immediate or predictable prospect of useful machine translation*."[47] As a result, research funding was stopped and there was little automated translation research in the United States from 1967 until a revival in 1976-1989.[46]

Similarly, there is currently tremendous interest in HIT. Although there is good evidence that HIT can be useful, some will certainly be disappointed. A recent report by the National Research Council (the same body that published the ALPAC report) concluded that "… *current efforts aimed at the nationwide deployment of health care IT will not be sufficient to achieve the vision of 21st century health care, and may even set back the cause if these efforts continue wholly without change from their present course*."[48] Thus, there is reason for concern that HIT (and perhaps even informatics, in general) may be headed for a bust. Such an "HIT winter" would be unfortunate, since there are real benefits of pursuing research and implementation of HIT.

The Problem: Health Information Technology is Really Health Data Technology

The fundamental problem is that existing technology stores, manipulates and transmits data (symbols), not information (data + meaning). Thus, the utility of HIT is limited by the extent to which data approximates meaning, or more precisely to the ability of HIT to act "as if" it understands the meaning of the data. Unfortunately, in health care, data do not fully represent the meaning. In other words, there is a large gap between data and information. Since the difference between data and information is meaning (semantics), this gap is referred to as the "*semantic gap*."

Social and Administrative Barriers to HIT Adoption. Manipulating data and not information has many consequences for HIT. Note that there is no shortage of computers in hospitals. Hospitals managed their financial data electronically long before they computerized clinical activities. Just like any other organization, many hospitals have functioning e-mail systems and maintain a Web presence. Many clinicians used personal digital assistants,[49] some even communicate with patients using e-mail.

The social and administrative barriers to HIT adoption have been discussed by multiple authors in countless papers. Such barriers include a mismatch between costs and benefits, cultural resistance to change, lack of an appropriately trained workforce to implement HIT and multiple others.[50] To some, clinicians' resistance to computerization appears irrational. However, caution seems increasingly reasonable given the mixed evidence regarding the benefits of poorly-implemented HIT.

FUTURE TRENDS

Significant research problems must be addressed before HIT becomes more attractive to clinicians. Many of these are outlined in a National Research Council report.[48] First, there is a mismatch between what HIT can represent (data) and concepts relevant to health care (data + meaning). This is a very difficult and fundamental challenge that includes multiple long-standing challenges in artificial intelligence (e.g., how computers can be "taught" context or common sense) that have proven very difficult to solve. It seems that until one has true information processing, rather than data processing, technology, the benefits of HIT will be limited.

Second, HIT must augment human cognition and abilities. Friedman expressed this elegantly as the "*fundamental theorem of informatics*": *human + computer > human* (humans working with computers should perform better than a human alone).[51] The theorem argues that there must be a clear and demonstrable benefit from HIT. Despite the problems with current HIT, there are clearly situations where HIT can be beneficial. In some ways, human cognition and computer technology are very complementary. For example, monitoring (e.g., waveforms) is much easier for computers than for humans. In contrast, reasoning by analogy across domains is natural for humans but difficult for computers.

How Progress Will Be Made

Researchers are exploring multiple promising paradigm-shifting ideas. Examples of approaches that address some of the fundamental problems described in this chapter can be provided.

One approach is to recognize the complementary strengths of humans and computers. Humans are good at constructing and processing meaning. In contrast, computers are much better at processing data. Users can leverage this understanding to design systems that harness the data-processing power of computers to present (display) data in ways that make it easier for humans to grasp and manipulate meaning. For example, a *word cloud visualization* shows the term frequency in text.[52] The size of the font is proportional to the frequency of the term.

Returning to HIT, one can apply these same principles. For example, Figure 2.10 shows an example of an

EHR that integrates clinical decision support. This is not novel, but this example illustrates what could be done by combining multiple types of information on the same screen with an understanding of the user's task.

Defining scenarios when HIT is beneficial with all relevant parameters and demonstrating that using HIT is *reliably* beneficial in these scenarios remains a research challenge. In its present form, HIT will not transform healthcare in the same way that IT has transformed other industries. This is due in part to the large semantic gap between health data and health information (concepts). In addition, many problems with healthcare require non-technological solutions, such as changes in healthcare policy and financing.

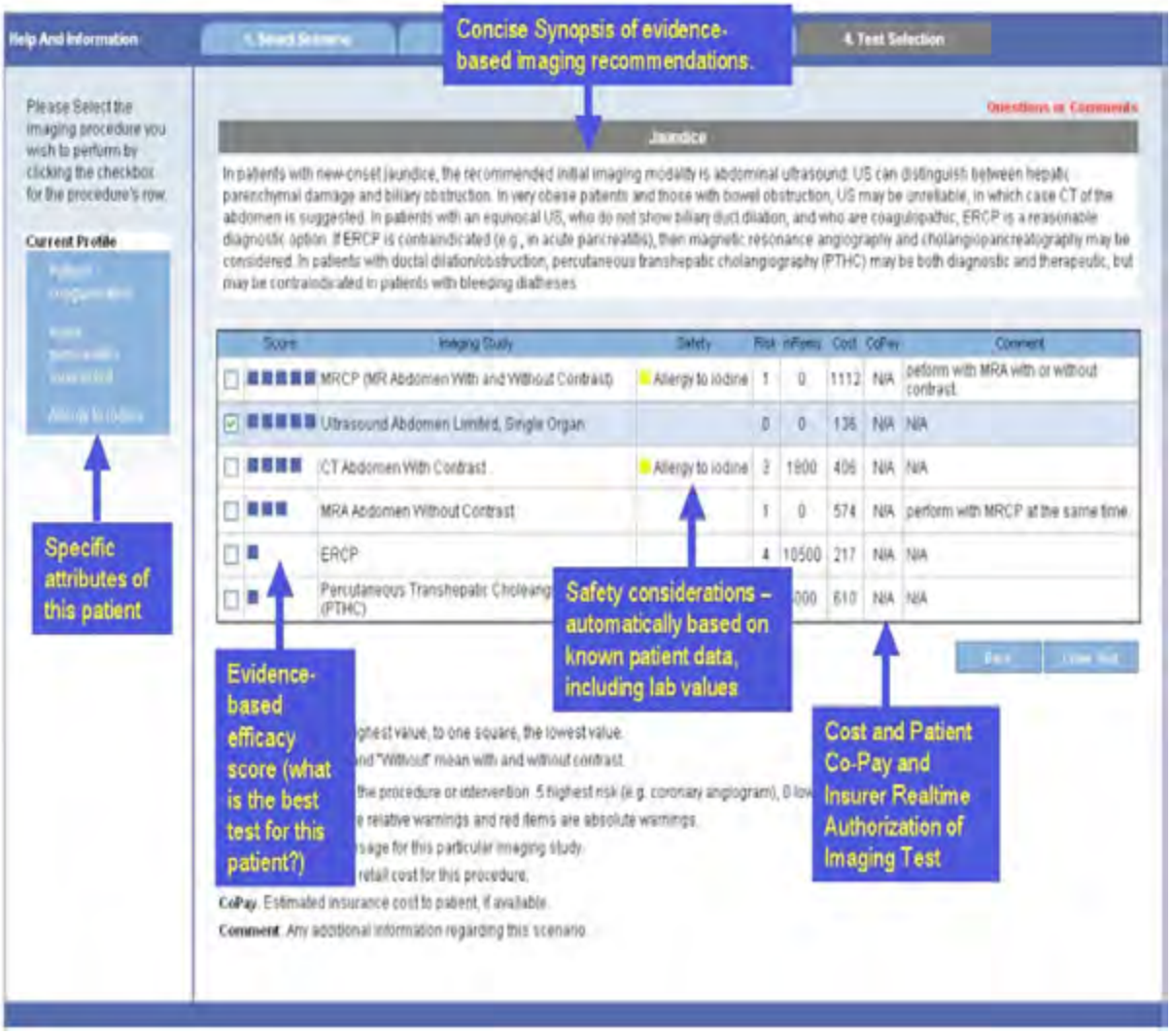

Figure 2.10: EHR screen (from John Halamka) showing integration of decision support into the EHR [53]

KEY POINTS

- Data are observations reflecting differences in the world (e.g., "C34.9") while information is meaningful data or facts from which conclusions can be drawn and knowledge is information that is justifiably believed to be true
- Data are largely the domain of information technology (IT) professionals and computer scientists; information and knowledge are the domains of informatics and informaticians
- Vocabularies help convert data into information
- The transformation of data into information and knowledge is a core concern of informaticians
- When the real world, the conceptual model and the computational model match, we get useful answers from the computer
- Concepts relevant to health are relatively poorly defined compared to e.g. banking concepts
- There is a large "semantic gap" between health data and health information

CONCLUSION

Problems in healthcare are information and knowledge intensive. Current technology is centered on processing data. This mismatch, or semantic gap, between the problems healthcare IT tries to address and the available technology explains the difficulties that informaticians face every day. It also explains the differences between Informatics and Computer Science. Informatics must advance our information and knowledge-processing capabilities to continue improving healthcare through technology.

REFERENCES

1. Stead WW, Lin HS, editors. Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. National Academies Press; Washington, D.C.: 2009. P. 2
2. E-patient Dave http://epatientdave.com/ (Accessed September 5, 2017)
3. Electronic Health Records Raise Doubt, Boston Globe, April 13, 2009. (Accessed September 5, 2017)
4. Bernstam, E.V., J.W. Smith, and T.R. Johnson. What is biomedical informatics? Journal of biomedical informatics, 2010. 43(1): p. 104-105
5. Floridi, L. Semantic conceptions of information. 2005 October 5, 2005; http://plato.stanford.edu/entries/information-semantic/ (Accessed September 5, 2017)
6. Bernstam, E.V., et al., Synergies and distinctions between computational disciplines in biomedical research: perspective from the Clinical and Translational Science Award programs. Academic Medicine : Journal of the Association of American Medical Colleges, 2009. 84(7): 964-70.
7. Manning CD, Raghavan P, Schutze H. Introduction to information retrieval. Cambridge University Press Cambridge 2008.
8. Andre Esteva A, Kuprel B, Novoa RA, Ko K, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542: 115–118. doi:10.1038/nature21056
9. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics. 2017 May 6:bbx044.
10. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. 2013. Cornell University Library. https://arxiv.org/abs/1312.6199 (Accessed August 30, 2017)
11. Chapman, G.B., & Elstein, A.S. (2000). Cognitive processes and biases in medical decision-making. In G. B. Chapman & F. S. Sonnenberg, (Eds.) Decision-making in health care: Theory, psychology, and applications (pp. 183-210). Cambridge: Cambridge University Press.
12. HealthIT.gov https://www.healthit.gov/buzz-blog/meaningful-use/interoperability-health-information-exchange-setting-record-straight/ (Accessed August 30, 2017)
13. Embi, P.J. and P.R. Payne, Clinical research informatics: challenges, opportunities and definition for an emerging domain. Journal of the American Medical Informatics Association: JAMIA, 2009. 16(3): p. 316-27.
14. Payne, P.R., P.J. Embi, and M.G. Kahn, Clinical research informatics: The maturing of a translational biomedical informatics sub-discipline. Journal of biomedical informatics, 2011.

15. van der Lei, J., Use and abuse of computer-stored medical records. Methods of information in medicine, 1991. 30(2): p. 79-80.
16. Zerhouni, E.A., Translational research: moving discovery to practice. Clin Pharmacol Ther, 2007. 81(1): p. 126-8.
17. Murphy, S.N., et al., Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association: JAMIA, 2010. 17(2): p. 124-30.
18. Informatics for Integrating Biology & the Bedside (I2b2) www.i2b2.org (Accessed September 5, 2017)
19. Szeto, H.C., et al., Accuracy of computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. The American journal of managed care, 2002. 8(1): p. 37-43.
20. Aronson, A.R. and F.M. Lang, An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association: JAMIA, 2010. 17(3): p. 229-36.
21. Savova, G.K., et al., Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association: JAMIA, 2010. 17(5): p. 507-13
22. Chen, E.S., et al., Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. Journal of the American Medical Informatics Association: JAMIA, 2008. 15(1): p. 87-98.
23. Chapman, W.W., et al., Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association: JAMIA, 2011. 18(5): p. 540-3.
24. Stanfill, M.H., et al., A systematic literature review of automated clinical coding and classification systems. Journal of the American Medical Informatics Association: JAMIA, 2010. 17(6): p. 646-51.
25. Meystre S, Haug PJ. Evaluation of medical problem extraction from electronic clinical documents using MetaMap Transfer (MMTx). Studies in health technology and informatics. 2005 Jan;116:823-8.
26. Friedman, C., et al., Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association: JAMIA, 2004. 11(5): p. 392-402.
27. Blois, M.S., Information and medicine: the nature of medical descriptions1984, Berkeley: University of California Press.
28. Smith, B. and W. Ceusters, HL7 RIM: an incoherent standard. Studies in health technology and informatics, 2006. 124: p. 133-8.
29. HL7 Version 3 RIM http://www.hl7.org/documentcenter/public_temp_B69AB426-1C23-BA17-0CA55CBFEF56C9A3/calendarofevents/himss/2011/HL7%20Reference%20Information%20Model.pdf (Accessed September 5, 2017)
30. Hopkin, K., The Evolving Definition of a Gene. BioScience, 2009. 59(11): p. 928-31.
31. Chaudhry, B., et al., Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med, 2006. 144(10): p. 742-52.
32. Bates, D.W., et al., Reducing the frequency of errors in medicine using information technology. J Am Med Inform Assoc, 2001. 8(4): p. 299-308.
33. American Recovery and Reinvestment Act (ARRA) of 2009. ARRA Economic Stimulus Package. Hitech Answers. https://www.hitechanswers.net/about/about-arra/ (Accessed September 5, 2017)
34. HealthConnect Evaluation Department of Health and Ageing (DoHA) 24 August 2009 http://www.health.gov.au/internet/main/publishing.nsf/Content/FAFD8FE999704592CA257BF00020A8CF/$File/HealthConnect.pdf (Accessed September 5, 2017)
35. France, F.R., eHealth in Belgium, a new "secure" federal network: role of patients, health professions and social security services. International Journal of Medical Informatics, 2011. 80(2): p. e12-6.
36. EHRS Blueprint: An interoperable EHR framework. Version 2. March 2006. [cited 2011 December 11]; Available from: https://www.infoway-inforoute.ca/en/component/edocman/resources/technical-documents/391-ehrs-blueprint-v2-full (Accessed September 5, 2017))
37. Protti, D. and I. Johansen, Widespread adoption of information technology in primary care physician offices in Denmark: a case study. Issue brief, 2010. 80: p. 1-14.
38. House of Commons Public Accounts Committee. The National Programme for IT in the NHS: Progress since 2006. Second Report of Session 2008-09. [cited 2011 December 11]; Available from: http://www.publications.parliament.uk/pa/cm200809/cmselect/cmpubacc/153/153.pdf (Accessed September 5, 2017)
39. Littlejohns, P., J.C. Wyatt, and L. Garvican. Evaluating computerized health information systems: hard lessons still to be learnt. BMJ, 2003. 326(7394): p. 860-3.
40. Han, Y.Y., et al., Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics, 2005. 116(6): p. 1506-12.

41. Levenson, N.G. and C.S. Turner, An Investigation of the Therac-25 Accidents. IEEE Computer, 1993(July): p. 18-41.
42. Koppel, R., et al., Role of computerized physician order entry systems in facilitating medication errors. JAMA, 2005. 293(10): p. 1197-203.
43. Services, C.o.P.S.a.H.I.T.B.o.H.C., Health IT and patient safety: building safer systems for better care, 2011, Institute of Medicine of the National Academies: Washington DC.
44. Weiner, J.P., et al., "e-Iatrogenesis": the most critical unintended consequence of CPOE and other HIT. J Am Med Inform Assoc, 2007. 14(3): p. 387-8; discussion 389.
45. Del Beccaro, M.A., et al., Computerized provider order entry implementation: no association with increased mortality rates in an intensive care unit. Pediatrics, 2006. 118(1): p. 290-5.
46. Hutchins, J., Machine translation: history, in Encyclopedia of language & linguistics, second edition, K. Brown, Editor 2006, Elsevier: Oxford. p. 375-83.
47. ALPAC, Language and machines: computers in translation and linguistics. Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council., 1966, National Academy of Sciences, National Research Council.: Washington, DC.
48. Computational technology for effective health care: immediate steps and strategic directions, W.W. Stead and H.S. Lin, Editors. 2009, Committee on Engaging the Computer Science Research Community in Health Care Informatics, Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, National Research Council of the National Academies: Washington, DC.
49. McLeod, T.G., J.O. Ebbert, and J.F. Lymp, Survey assessment of personal digital assistant use among trainees and attending physicians. J Am Med Inform Assoc, 2003. 10(6): p. 605-7.
50. Hersh, W., Health care information technology: progress and barriers. JAMA, 2004. 292(18): p. 2273-4.
51. Friedman, C.P., A "fundamental theorem" of biomedical informatics. J Am Med Inform Assoc, 2009. 16(2): p. 169-70.
52. Visualizations: JAMIA Content. May13, 2009. http://wordcloud.cs.arizona.edu/ (Accessed September 5, 2017)
53. My Life as a CMIO. http://geekdoctor.blogspot.com/2007/11/data-information-knowledge-and-wisdom.html (Accessed September 5, 2017)