2 Biomedical Data: Their Acquisition, Storage, and Use

Edward H. Shortliffe and G. Octo Barnett

After reading this chapter, you should know the answers to these questions:

- What are clinical data?
- How are clinical data used?
- What are the drawbacks of the traditional paper medical record?
- What is the potential role of the computer in data storage, retrieval, and interpretation?
- What distinguishes a database from a knowledge base?
- How are data collection and hypothesis generation intimately linked in clinical diagnosis?
- What are the meanings of the terms *prevalence*, *predictive value*, *sensitivity*, and *specificity*?
- How are the terms related?
- What are the alternatives for entry of data into a clinical database?

2.1 What Are Clinical Data?

From earliest times, the ideas of ill health and its treatment have been wedded to those of the observation and interpretation of data. Whether we consider the disease descriptions and guidelines for management in early Greek literature or the modern physician's use of complex laboratory and X-ray studies, it is clear that gathering data and interpreting their meaning are central to the health care process. With the move toward the use of genomic information in assessing individual patients (their risks, prognosis, and likely responses to therapy), the sheer amounts of data that may be used in patient care have become huge. A textbook on informatics will accordingly refer time and again to issues in data collection, storage, and use. This chapter lays the foundation for this recurring set of issues that is pertinent to all aspects of the use of information, knowledge, and computers in biomedicine, both in the clinical world and in applications related to, public health, biology and human genetics.

If data are central to all health care, it is because they are crucial to the process of decision making (as described in detail in Chaps. 3 and 4 and again in Chap. 22). In fact, simple reflection will reveal that all health care activities involve gathering, analyzing, or using data. Data provide the basis for categorizing the problems a patient may be having or for identifying subgroups within a population of patients. They also help a physician to decide what additional information

E.H. Shortliffe, MD, PhD (✉)
Departments of Biomedical Informatics
at Columbia University and Arizona State University,
Weill Cornell College of Medicine,
and the New York Academy of Medicine,
272 W 107th St #5B, New York 10025, NY, USA
e-mail: ted@shortliffe.net

G.O. Barnett, MD, FACP, FACMI
Laboratory of Computer Science
(Harvard Medical School and Massachusetts
General Hospital),
Boston 02114, MA, USA

E.H. Shortliffe, J.J. Cimino (eds.), *Biomedical Informatics*,
DOI 10.1007/978-1-4471-4474-8_2, © Springer-Verlag London 2014

is needed and what actions should be taken to gain a greater understanding of a patient's problem or most effectively to treat the problem that has been diagnosed.

It is overly simplistic to view data as the columns of numbers or the monitored waveforms that are a product of our increasingly technological health care environment. Although laboratory test results and other numeric data are often invaluable, a variety of more subtle types of data may be just as important to the delivery of optimal care: the awkward glance by a patient who seems to be avoiding a question during the medical interview, information about the details of a patient's symptoms or about his family or economic setting, or the subjective sense of disease severity that an experienced clinician will often have within a few moments of entering a patient's room. No clinician disputes the importance of such observations in decision making during patient assessment and management, yet the precise role of these data and the corresponding decision criteria are so poorly understood that it is difficult to record them in ways that convey their full meaning, even from one clinician to another. Despite these limitations, clinicians need to share descriptive information with others. When they cannot interact directly with one another, they often turn to the chart or electronic health record for communication purposes.

We consider a **clinical datum** to be any single observation of a patient—e.g., a temperature reading, a red blood cell count, a past history of rubella, or a blood pressure reading. As the blood pressure example shows, it is a matter of perspective whether a single observation is in fact more than one datum. A blood pressure of 120/80 might well be recorded as a single element in a setting where knowledge that a patient's blood pressure is normal is all that matters. If the difference between diastolic (while the heart cavities are beginning to fill) and systolic (while they are contracting) blood pressures is important for decision making or for analysis, however, the blood pressure reading is best viewed as two pieces of information (systolic pressure = 120 mmHg, diastolic pressure = 80 mmHg). Human beings can glance at a written blood pressure value and easily make the transition between its unitary view as a single data point and the decomposed information about systolic and diastolic pressures. Such dual views can be much more difficult for computers, however, unless they are specifically allowed for in the design of the method for data storage and analysis. The idea of a *data model* for computer-stored medical data accordingly becomes an important issue in the design of medical data systems.

If a clinical *datum* is a single observation about a patient, clinical *data* are multiple observations. Such data may involve several different observations made concurrently, the observation of the same patient parameter made at several points in time, or both. Thus, a single datum generally can be viewed as defined by five elements:

1. The *patient* in question
2. The *parameter* being observed (e.g., liver size, urine sugar value, history of rheumatic fever, heart size on chest X-ray film)
3. The *value* of the parameter in question (e.g., weight is 70 kg, temperature is 98.6 °F, profession is steel worker)
4. The *time* of the observation (e.g., 2:30 A.M. on 14FEB2013[1])
5. The method by which the observation was made (e.g., patient report, thermometer, urine dipstick, laboratory instrument).

Time can particularly complicate the assessment and computer-based management of data. In some settings, the date of the observation is adequate—e.g., in outpatient clinics or private offices where a patient generally is seen infrequently and the data collected need to be identified in time with no greater accuracy than a calendar date. In others, minute-to-minute variations may be important—e.g., the frequent blood sugar readings obtained for a patient in diabetic ketoacidosis (acid production due to poorly controlled blood sugar levels) or the continuous measurements of mean arterial blood pressure for a

[1] Note that it was the tendency to record such dates in computers as "14FEB12" that led to the end-of-century complexities that we called the *Year 2K problem*. It was shortsighted to think that it was adequate to encode the year of an event with only two digits.

patient in cardiogenic shock (dangerously low blood pressure due to failure of the heart muscle).

It often also is important to keep a record of the circumstances under which a data point was obtained. For example, was the blood pressure taken in the arm or leg? Was the patient lying or standing? Was the pressure obtained just after exercise? During sleep? What kind of recording device was used? Was the observer reliable? Such additional information, sometimes called contexts, methods, or modifiers, can be of crucial importance in the proper interpretation of data. Two patients with the same basic problem or symptom often have markedly different explanations for their problem, revealed by careful assessment of the modifiers of that problem.

A related issue is the uncertainty in the values of data. It is rare that an observation—even one by a skilled clinician—can be accepted with absolute certainty. Consider the following examples:

- An adult patient reports a childhood illness with fevers and a red rash in addition to joint swelling. Could he or she have had scarlet fever? The patient does not know what his or her pediatrician called the disease nor whether anyone thought that he or she had scarlet fever.
- A physician listens to the heart of an asthmatic child and thinks that she hears a heart murmur—but is not certain because of the patient's loud wheezing.
- A radiologist looking at a shadow on a chest X-ray film is not sure whether it represents overlapping blood vessels or a lung tumor.
- A confused patient is able to respond to simple questions about his or her illness, but under the circumstances the physician is uncertain how much of the history being reported is reliable.

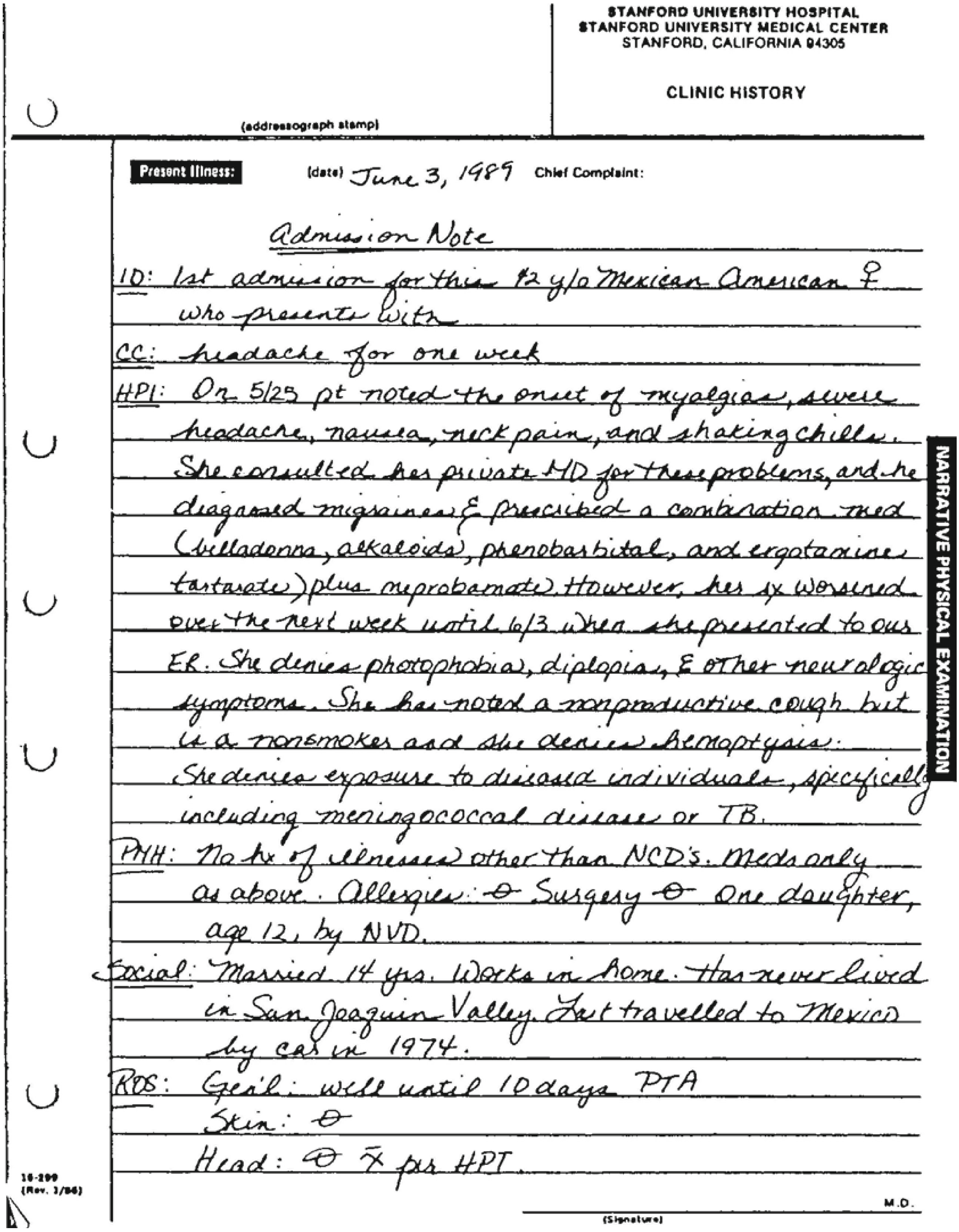
As described in Chaps. 3 and 4, there are a variety of possible responses to deal with incomplete data, the uncertainty in them, and in their interpretation. One technique is to collect additional data that will either confirm or eliminate the concern raised by the initial observation. This solution is not always appropriate, however, because the costs of data collection must be considered. The additional observation might be expensive, risky for the patient, or wasteful of time during which treatment could have been instituted. The idea of trade-offs in data collection thus becomes extremely important in guiding health care decision making.

2.1.1 What Are the Types of Clinical Data?

The examples in the previous section suggest that there is a broad range of data types in the practice of medicine and the allied health sciences. They range from narrative, textual data to numerical measurements, genetic information, recorded signals, drawings, and even photographs or other images.

Narrative data account for a large component of the information that is gathered in the care of patients. For example, the patient's description of his or her present illness, including responses to focused questions from the physician, generally is gathered verbally and is recorded as text in the medical record. The same is true of the patient's social and family history, the general review of systems that is part of most evaluations of new patients, and the clinician's report of physical examination findings. Such narrative data were traditionally handwritten by clinicians and then placed in the patient's medical record (Fig. 2.1). Increasingly, however, the narrative summaries are dictated and then transcribed by typists who produce printed summaries or electronic copies for inclusion in paper or electronic medical records. The electronic versions of such reports can easily be integrated into electronic health records (EHRs) and clinical data repositories so that clinicians can access important clinical information even when the paper record is not available.[2] Electronically stored transcriptions of dictated information often include not only patient histories and physical examinations but also other narrative descriptions such as reports of specialty consultations, surgical procedures,

[2] As is discussed in Chap. 12, health care organizations are increasingly relying on electronic health records to the exclusion of printed records.



STANFORD UNIVERSITY HOSPITAL
STANFORD UNIVERSITY MEDICAL CENTER
STANFORD, CALIFORNIA 94305

CLINIC HISTORY

(addressograph stamp)

Present Illness: (date) June 3, 1989 Chief Complaint:

Admission Note

ID: 1st admission for this 42 y/o Mexican American ♀ who presents with

CC: headache for one week

HPI: On 5/25 pt noted the onset of myalgias, severe headache, nausea, neck pain, and shaking chills. She consulted her private MD for these problems, and he diagnosed migraines c̄ prescribed a combination med (belladonna, alkaloids), phenobarbital, and ergotamine tartrate) plus meprobamate). However, her sx worsened over the next week until 6/3 when she presented to our ER. She denies photophobia, diplopia, c̄ other neurologic symptoms. She has noted a nonproductive cough but is a nonsmoker and she denies hemoptysis. She denies exposure to diseased individuals, specifically including meningococcal disease or TB.

PMH: No hx of illnesses other than NCD's. Meds only as above. Allergies: ⊖ Surgery ⊖ One daughter, age 12, by NVD.

Social: Married 14 yrs. Works in home. Has never lived in San Joaquin Valley. Last travelled to Mexico by car in 1974.

ROS: Gen'l: well until 10 days PTA
Skin: ⊖
Head: ⊖ x̄ per HPI.

NARRATIVE PHYSICAL EXAMINATION

16-299
(Rev. 3/86)

M.D.
(Signature)

Fig. 2.1 Much of the information gathered during a physician–patient encounter is written in the medical record

pathologic examinations of tissues, and hospitalization summaries when a patient is discharged.

Some narrative data are loosely coded with shorthand conventions known to health personnel, particularly data collected during the physical examination, in which recorded observations reflect the stereotypic examination process taught to all practitioners. It is common, for example, to find the notation "PERRLA" under the eye examination in a patient's medical record. This encoded form indicates that the patient's "Pupils are Equal (in size), Round, and Reactive to Light and Accommodation (the process of focusing on near objects)."

Note that there are significant problems associated with the use of such abbreviations. Many are not standard and can have different meanings depending on the context in which they are used. For example, "MI" can mean "mitral insufficiency" (leakage in one of the heart's valves) or "myocardial infarction" (the medical term for what is commonly called a heart attack). Many hospitals try to establish a set of "acceptable" abbreviations with meanings, but the enforcement of such standardization is often unsuccessful.

Complete phrases have become loose standards of communication among medical personnel. Examples include "mild dyspnea (shortness of breath) on exertion," "pain relieved by antacids or milk," and "failure to thrive." Such standardized expressions are attempts to use conventional text notation as a form of summarization for otherwise heterogeneous conditions that together characterize a simple concept about a patient.

Many data used in medicine take on discrete numeric values. These include such parameters as laboratory tests, vital signs (such as temperature and pulse rate), and certain measurements taken during the physical examination. When such numerical data are interpreted, however, the issue of precision becomes important. Can physicians distinguish reliably between a 9-cm and a 10-cm liver span when they examine a patient's abdomen? Does it make sense to report a serum sodium level to two-decimal-place accuracy? Is a 1-kg fluctuation in weight from 1 week to the next significant? Was the patient weighed on the same scale both times (i.e., could the different values reflect variation between measurement instruments rather than changes in the patient)?

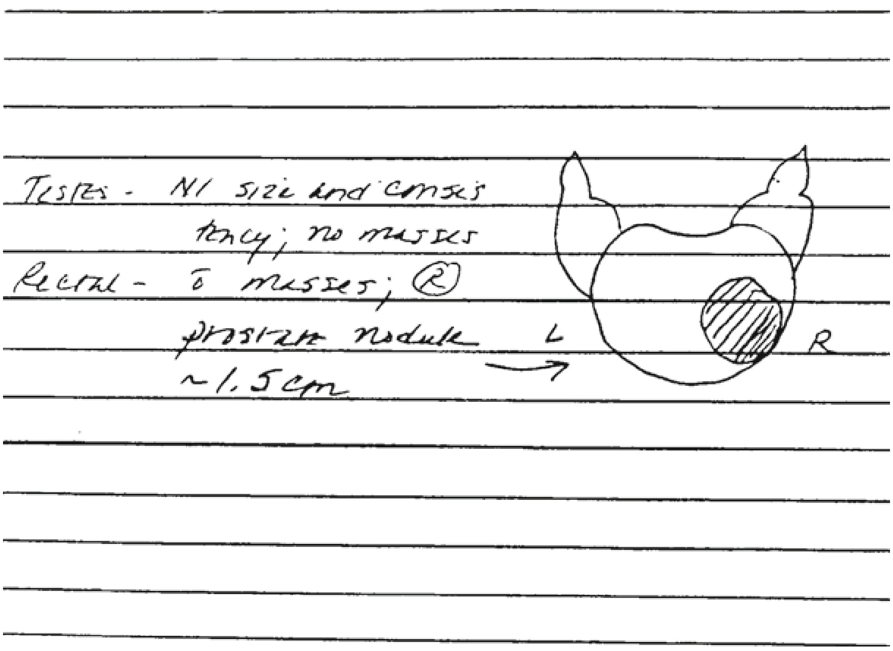
In some fields of medicine, analog data in the form of continuous signals are particularly important (see Chap. 19). Perhaps the best-known example is an electrocardiogram (ECG), a tracing of the electrical activity from a patient's heart. When such data are stored in medical records, a graphical tracing frequently is included, with a written interpretation of its meaning. There are clear challenges in determining how such data are best managed in computer-based storage systems.

Visual images—acquired from machines or sketched by the physician—are another important category of data. Radiologic images or photographs of skin lesions are obvious examples. It also is common for physicians to draw simple pictures to represent abnormalities that they have observed; such drawings may serve as a basis for comparison when they or another physician next see the patient. For example, a sketch is a concise way of conveying the location and size of a nodule in the prostate gland (Fig. 2.2).

As should be clear from these examples, the idea of data is inextricably bound to the idea of **data recording**. Physicians and other health care personnel are taught from the outset that it is crucial that they do not trust their memory when caring for patients. They must record their observations, as well as the actions they have taken and the rationales for those actions, for later communication to themselves and other people. A glance at a medical record will quickly reveal the wide variety of data-recording techniques that have evolved. The range goes from handwritten text to commonly understood shorthand notation to cryptic symbols that only specialists can understand; few physicians without specialized training know how to interpret the data-recording conventions of an ophthalmologist, for example (Fig. 2.3). The notations may be highly structured records with brief text or numerical information, hand-drawn sketches, machine-generated tracings of analog signals, or photographic images (of the patient or of radiologic or other studies). This range of data-recording conventions presents significant challenges to the person implementing electronic health record systems.

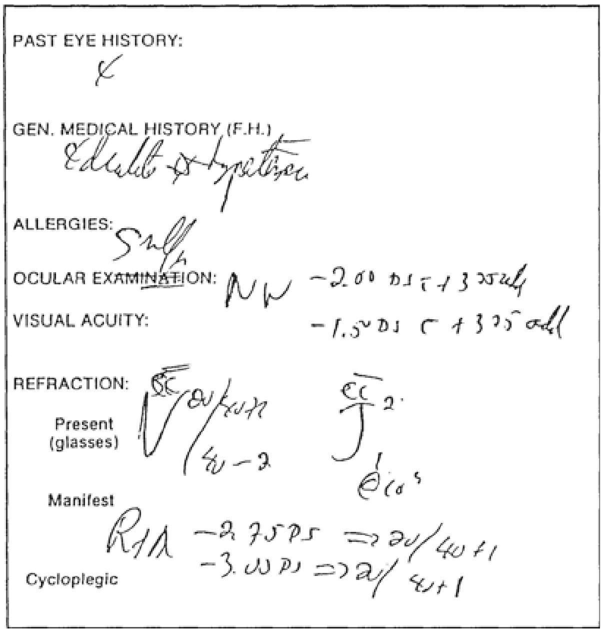
2.1.2 Who Collects the Data?

Health data on patients and populations are gathered by a variety of health professionals. Although conventional ideas of the **health care team** evoke images of coworkers treating ill patients, the team



Testes - Nl size and consistency; no masses
Rectal - ō masses; (R)
prostate nodule
~1.5cm

Fig. 2.2 A physician's hand-drawn sketch of a prostate nodule. A drawing may convey precise information more easily and compactly than a textual description



PAST EYE HISTORY: ø

GEN. MEDICAL HISTORY (F.H.) Diabetes & hypertension

ALLERGIES: Sulfa

OCULAR EXAMINATION: NL −2.00 DS c̄ +1.375 add

VISUAL ACUITY: −1.50 DS c̄ +1.375 add

REFRACTION: sc 20/400 20/40−2 cc 2 ...

Present (glasses)

Manifest

R/L −2.75 DS => 20/40+1
−3.00 DS => 20/40+1

Cycloplegic

Fig. 2.3 An ophthalmologist's report of an eye examination. Most physicians trained in other specialties would have difficulty deciphering the symbols that the ophthalmologist has used

has much broader responsibilities than treatment per se; data collection and recording are a central part of its task.

Physicians are key players in the process of data collection and interpretation. They converse with a patient to gather narrative descriptive data on the chief complaint, past illnesses, family and social information, and the system review. They examine the patient, collecting pertinent data and recording them during or at the end of the visit. In addition, they generally decide what additional data to collect by ordering laboratory or radiologic studies and by observing the patient's response to therapeutic interventions (yet another form of data that contributes to patient assessment).

In both outpatient and hospital settings, nurses play a central role in making observations and recording them for future reference. The data that they gather contribute to nursing care plans as well as to the assessment of patients by physicians and by other health care staff. Thus, nurses' training includes instruction in careful and accurate observation, history taking, and examination of the patient. Because nurses typically spend more time with patients than physicians do, especially in the hospital setting, nurses often build relationships with patients that uncover information and insights that contribute to proper diagnosis, to understanding of pertinent psychosocial issues, or to proper planning of therapy or discharge management (Fig. 2.4). The role of information systems in contributing to patient care tasks such as care planning by nurses is the subject of Chap. 15.

Various other health care workers contribute to the data-collection process. Office staff and admissions personnel gather demographic and financial information. Physical or respiratory

Fig. 2.4 Nurses often develop close relationships with patients. These relationships may allow the nurse to make observations that are missed by other staff. This ability is just one of the ways in which nurses play a key role in data collection and recording (Photograph courtesy of Janice Anne Rohn)

therapists record the results of their treatments and often make suggestions for further management. Laboratory personnel perform tests on biological samples, such as blood or urine, and record the results for later use by physicians and nurses. Radiology technicians perform X-ray examinations; radiologists interpret the resulting data and report their findings to the patients' physicians. Pharmacists may interview patients about their medications or about drug allergies and then monitor the patients' use of prescription drugs. As these examples suggest, many different individuals employed in health care settings gather, record, and make use of patient data in their work.

Finally, there are the technological devices that generate data—laboratory instruments, imaging machines, monitoring equipment in intensive care units, and measurement devices that take a single reading (such as thermometers, ECG machines, sphygmomanometers for taking blood pressure, and spirometers for testing lung function). Sometimes such a device produces a paper report suitable for inclusion in a traditional medical record. Sometimes the device indicates a result on a gauge or traces a result that must be read by an operator and then recorded in the patient's chart. Sometimes a trained specialist must interpret the output. Increasingly, however, the devices feed their results directly into computer equipment so that the data can be analyzed or formatted for electronic storage as well as reported on paper. With the advent of comprehensive EHRs (see Chap. 12), the printing of such data summaries may no longer be required as we move to "paperless" records whereby all access to information is through computer workstations, hand-held tablets, or even smart phones.

2.2 Uses of Health Data

Health data are recorded for a variety of purposes. Clinical data may be needed to support the proper care of the patient from whom they were obtained, but they also may contribute to the good of society through the aggregation and analysis of data regarding populations of individuals (supporting clinical research or public health assessments; see Chaps. 16 and 26). Traditional data-recording techniques and a paper record may have worked reasonably well when care was given by a single physician over the life of a patient. However, given the increased complexity of modern health care, the broadly trained team of individuals who are involved in a patient's care, and the need for multiple providers to access a patient's data and to communicate effectively with one another through the chart, the paper record no longer adequately supports optimal care of individual patients. Another problem occurs because traditional paper-based data-recording techniques have made clinical research across populations of patients extremely cumbersome. Electronic record keeping offers major advantages in this regard, as we discuss in more detail later in this chapter and in Chaps. 12 and 16.

2.2.1 Create the Basis for the Historical Record

Any student of science learns the importance of collecting and recording data meticulously when carrying out an experiment. Just as a laboratory notebook provides a record of precisely what a scientist has done, the experimental data observed, and the rationale for intermediate decision points, medical records are intended to provide a detailed compilation of information about individual patients:

- What is the patient's history (development of a current illness; other diseases that coexist or have resolved; pertinent family, social, and demographic information)?
- What symptoms has the patient reported? When did they begin, what has seemed to aggravate them, and what has provided relief?
- What physical signs have been noted on examination?
- How have signs and symptoms changed over time?
- What laboratory results have been, or are now, available?
- What radiologic and other special studies have been performed?
- What medications are being taken and are there any allergies?
- What other interventions have been undertaken?
- What is the reasoning behind the management decisions?

Each new patient problem and its management can be viewed as a therapeutic experiment, inherently confounded by uncertainty, with the goal of answering three questions when the experiment is over:

1. What was the nature of the disease or symptom?
2. What was the treatment decision?
3. What was the outcome of that treatment?

As is true for all experiments, one purpose is to learn from experience through careful observation and recording of data. The lessons learned in a given encounter may be highly individualized (e.g., the physician may learn how a specific patient tends to respond to pain or how family interactions tend to affect the patient's response to disease). On the other hand, the value of some experiments may be derived only by pooling of data from many patients who have similar problems and through the analysis of the results of various treatment options to determine efficacy.

Although laboratory research has contributed dramatically to our knowledge of human disease and treatment, especially over the last half century, it is careful observation and recording by skilled health care personnel that has always been of fundamental importance in the generation of new knowledge about patient care. We learn from the aggregation of information from large numbers of patients; thus, the historical record for individual patients is of inestimable importance to clinical research.

2.2.2 Support Communication Among Providers

A central function of structured data collection and recording in health care settings is to assist personnel in providing coordinated care to a patient over time. Most patients who have significant medical conditions are seen over months or years on several occasions for one or more problems that require ongoing evaluation and treatment. Given the increasing numbers of elderly patients in many cultures and health care settings, the care given to a patient is less oriented to diagnosis and treatment of a single disease episode and increasingly focused on management of one or more chronic disorders—possibly over many years.

It was once common for patients to receive essentially all their care from a single provider: the family doctor who tended both children and adults, often seeing the patient over many or all the years of that person's life. We tend to picture such physicians as having especially close relationships with their patients—knowing the family and sharing in many of the patient's life events, especially in smaller communities. Such doctors nonetheless kept records of all encounters so that they could refer to data about past illnesses and treatments as a guide to evaluating future care issues.

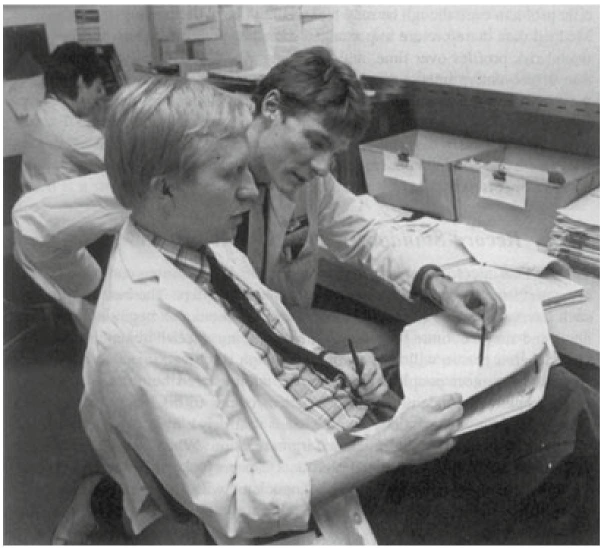


Fig. 2.5 One role of the medical record: a communication mechanism among health professionals who work together to plan patient care (Photograph courtesy of Janice Anne Rohn)

In the world of modern medicine, the emergence of subspecialization and the increasing provision of care by teams of health professionals have placed new emphasis on the central role of the medical record. Shared access to a paper chart (Fig. 2.5) is now increasingly being replaced by clinicians accessing electronic records, sometimes conferring as they look at the same computer screen (Fig. 2.6). Now the record not only contains observations by a physician for reference on the next visit but also serves as a communication mechanism among physicians and other medical personnel, such as physical or respiratory therapists, nursing staff, radiology technicians, social workers, or discharge planners. In many outpatient settings, patients receive care over time from a variety of physicians—colleagues covering for the primary physician, or specialists to whom the patient has been referred, or a managed care organization's case manager. It is not uncommon to hear complaints from patients who remember the days when it was

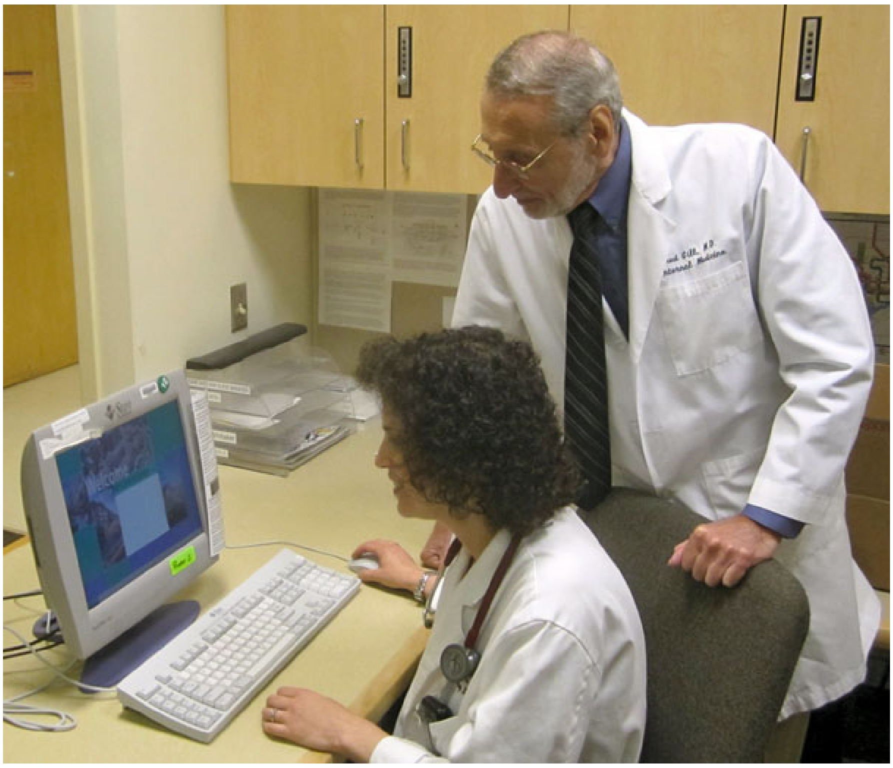


Fig. 2.6 Today similar communication sessions occur around a computer screen rather than a paper chart (see Fig. 2.5) (Photograph courtesy of James J. Cimino)

possible to receive essentially all their care from a single physician whom they had come to trust and who knew them well. Physicians are sensitive to this issue and therefore recognize the importance of the medical record in ensuring quality and **continuity of care** through adequate recording of the details and logic of past interventions and ongoing treatment plans. This idea is of particular importance in a health care system, such as the one in the United States, in which chronic diseases rather than care for trauma or acute infections increasingly dominate the basis for interactions between patients and their doctors.

2.2.3 Anticipate Future Health Problems

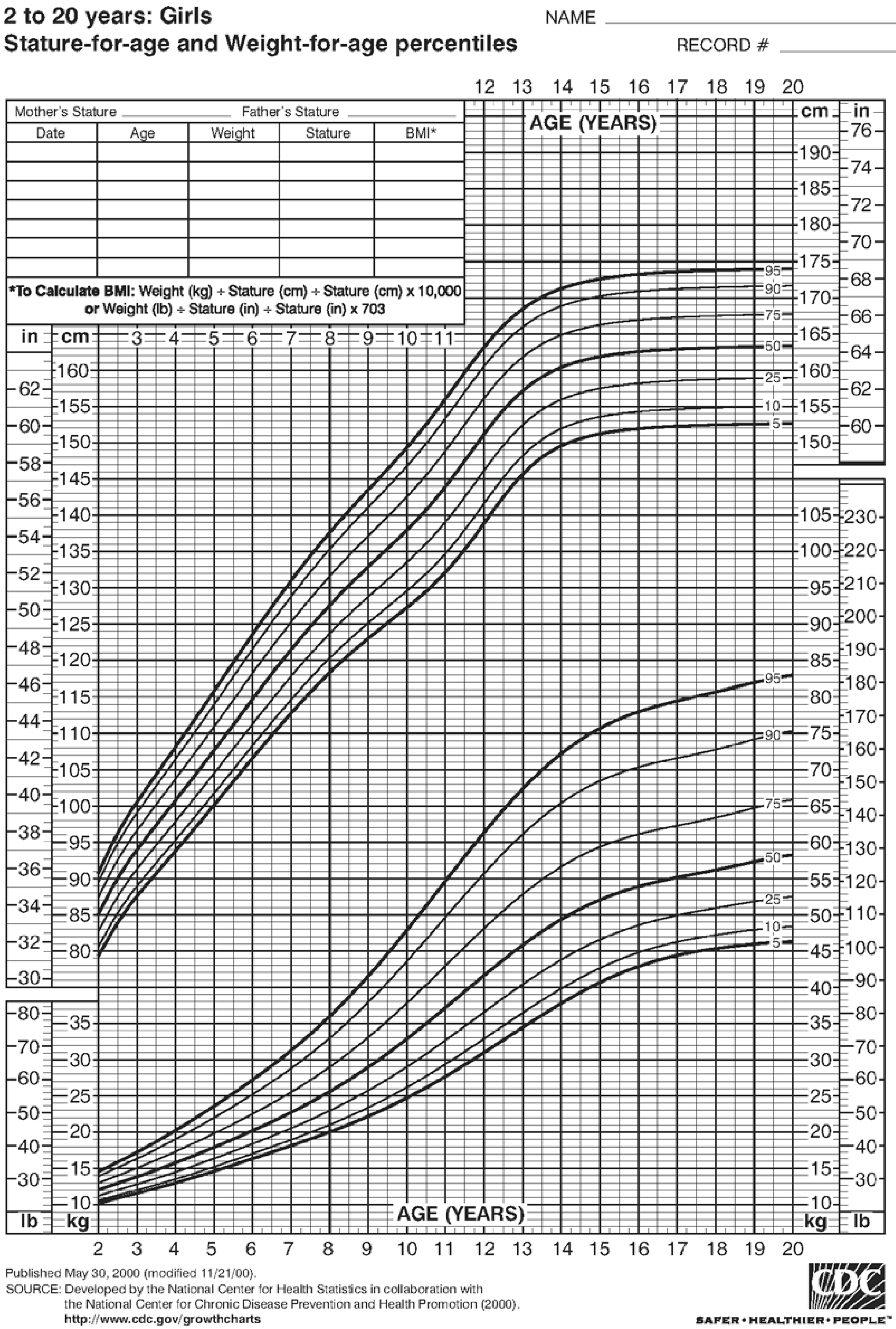
Providing high-quality health care involves more than responding to patients' acute or chronic health problems. It also requires educating patients about the ways in which their environment and lifestyles can contribute to, or reduce the risk of, future development of disease. Similarly, data gathered routinely in the ongoing care of a patient may suggest that he or she is at high risk of developing a specific problem even though he or she may feel well and be without symptoms at present. Clinical data therefore are important in screening for risk factors, following patients' risk profiles over time, and providing a basis for specific patient education or preventive interventions, such as diet, medication, or exercise. Perhaps the most common examples of such ongoing risk assessment in our society are routine monitoring for excess weight, high blood pressure, and elevated serum cholesterol levels. In these cases, abnormal data may be predictive of later symptomatic disease; optimal care requires early intervention before the complications have an opportunity to develop fully.

2.2.4 Record Standard Preventive Measures

The medical record also serves as a source of data on interventions that have been performed to prevent common or serious disorders. Sometimes the interventions involve counseling or educational programs (for example, regarding smoking cessation, measures for stopping drug abuse, safe sex practices, and dietary changes to lower cholesterol). Other important preventive interventions include immunizations: the vaccinations that begin in early childhood and continue throughout life, including special treatments administered when a person will be at particularly high risk (e.g., injections of gamma globulin to protect people from hepatitis, administered before travel to areas where hepatitis is endemic). When a patient comes to his local hospital emergency room with a laceration, the physicians routinely check for an indication of when he most recently had a tetanus immunization. When easily accessible in the record (or from the patient), such data can prevent unnecessary treatments (in this case, a repeat injection) that may be associated with risk or significant cost.

2.2.5 Identify Deviations from Expected Trends

Data often are useful in medical care only when viewed as part of a continuum over time. An example is the routine monitoring of children for normal growth and development by pediatricians (Fig. 2.7). Single data points regarding height and weight may have limited use by themselves; it is the trend in such data points observed over months or years that may provide the first clue to a medical problem. It is accordingly common for such parameters to be recorded on special charts or forms that make the trends easy to discern at a glance. Women who want to get pregnant often keep similar records of body temperature. By measuring temperature daily and recording the values on special charts, women can identify the slight increase in temperature that accompanies ovulation and thus may discern the days of maximum fertility. Many physicians will ask a patient to keep such graphical records so that they can later discuss the data with the patient and include the record in the medical chart for ongoing reference. Such graphs are increasingly created and displayed for viewing by clinicians as a feature of a patient's medical record.


2 to 20 years: Girls
Stature-for-age and Weight-for-age percentiles

Published May 30, 2000 (modified 11/21/00).
SOURCE: Developed by the National Center for Health Statistics in collaboration with
the National Center for Chronic Disease Prevention and Health Promotion (2000).
http://www.cdc.gov/growthcharts

Fig. 2.7 A pediatric growth chart. Single data points would not be useful; it is the changes in values over time that indicate whether development is progressing normally (Source: National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000). http://www.cdc.gov/growthcharts)

2.2.6 Provide a Legal Record

Another use of health data, once they are charted and analyzed, is as the foundation for a legal record to which the courts can refer if necessary. The medical record is a legal document; the responsible individual must sign most of the clinical information that is recorded. In addition, the chart generally should describe and justify both the presumed diagnosis for a patient and the choice of management.

We emphasized earlier the importance of recording data; in fact, data do not exist in a generally useful form unless they are recorded. The legal system stresses this point as well. Providers' unsubstantiated memories of what they observed or why they took some action are of little value in the courtroom. The medical record is the foundation for determining whether proper care was delivered. Thus, a well-maintained record is a source of protection for both patients and their physicians.

2.2.7 Support Clinical Research

Although experience caring for individual patients provides physicians with special skills and enhanced judgment over time, it is only by formally analyzing data collected from large numbers of patients that researchers can develop and validate new clinical knowledge of general applicability. Thus, another use of clinical data is to support research through the aggregation and statistical or other analysis of observations gathered from populations of patients (see Chap. 1).

A **randomized clinical trial** (**RCT**) (see also Chaps. 11 and 26) is a common method by which specific clinical questions are addressed experimentally. RCTs typically involve the random assignment of matched groups of patients to alternate treatments when there is uncertainty about how best to manage the patients' problem. The variables that might affect a patient's course (e.g., age, gender, weight, coexisting medical problems) are measured and recorded. As the study progresses, data are gathered meticulously to provide a record of how each patient fared under treatment and precisely how the treatment was administered. By pooling such data, sometimes after years of experimentation (depending on the time course of the disease under consideration), researchers may be able to demonstrate a statistical difference among the study groups depending on precise characteristics present when patients entered the study or on the details of how patients were managed. Such results then help investigators to define the standard of care for future patients with the same or similar problems.

Medical knowledge also can be derived from the analysis of large patient data sets even when the patients were not specifically enrolled in an RCT, often referred to as **retrospective studies**. Much of the research in the field of epidemiology involves analysis of population-based data of this type. Our knowledge of the risks associated with cigarette smoking, for example, is based on irrefutable statistics derived from large populations of individuals with and without lung cancer, other pulmonary problems, and heart disease.

2.3 Weaknesses of the Traditional Medical Record System

The preceding description of medical data and their uses emphasizes the positive aspects of information storage and retrieval in the record. All medical personnel, however, quickly learn that use of the traditional paper record is complicated by a bevy of logistical and practical realities that greatly limit the record's effectiveness for its intended uses.

2.3.1 Pragmatic and Logistical Issues

Recall, first, that data cannot effectively serve the delivery of health care unless they are recorded. Their optimal use depends on positive responses to the following questions:

- Can I find the data I need when I need them?
- Can I find the medical record in which they are recorded?

- Can I find the data within the record?
- Can I find what I need quickly?
- Can I read and interpret the data once I find them?
- Can I update the data reliably with new observations in a form consistent with the requirements for future access by me or other people?

All too frequently, the traditional paper record creates situations in which people answer such questions in the negative. For example:

- The patient's paper chart may be unavailable when the health care professional needs it. It may be in use by someone else at another location; it may have been misplaced despite the record-tracking system of the hospital, clinic, or office (Fig. 2.8); or it may have been taken by someone unintentionally and is now buried on a desk.
- Once the chart is in hand, it might still be difficult to find the information required. The data may have been known previously but never recorded due to an oversight by a physician or other health professional. Poor organization in the chart may lead the user to spend an inordinate time searching for the data, especially in the massive paper charts of patients who have long and complicated histories.
- Once the health care professional has located the data, he or she may find them difficult to read. It is not uncommon to hear one physician asking another as they peer together into a chart: "What is that word?" "Is that a two or a five?" "Whose signature is that?" Illegible and sloppy entries can be a major obstruction to effective use of the chart (Fig. 2.9).
- When a chart is unavailable, the health care professional still must provide patient care. Thus, providers make do without past data, basing their decisions instead on what the patient can tell them and on what their examination reveals. They then write a note for inclusion in the chart—when the chart is located. In a large institution with thousands of medical records, it is not surprising that such loose notes often fail to make it to the patient's chart or are filed out of sequence so that the actual chronology of management is disrupted in the record.
- When patients who have chronic or frequent diseases are seen over months or years, their records grow so large that the charts must be broken up into multiple volumes. When a hospital clinic or emergency room orders the patient's chart, only the most recent volume typically is provided. Old but pertinent data

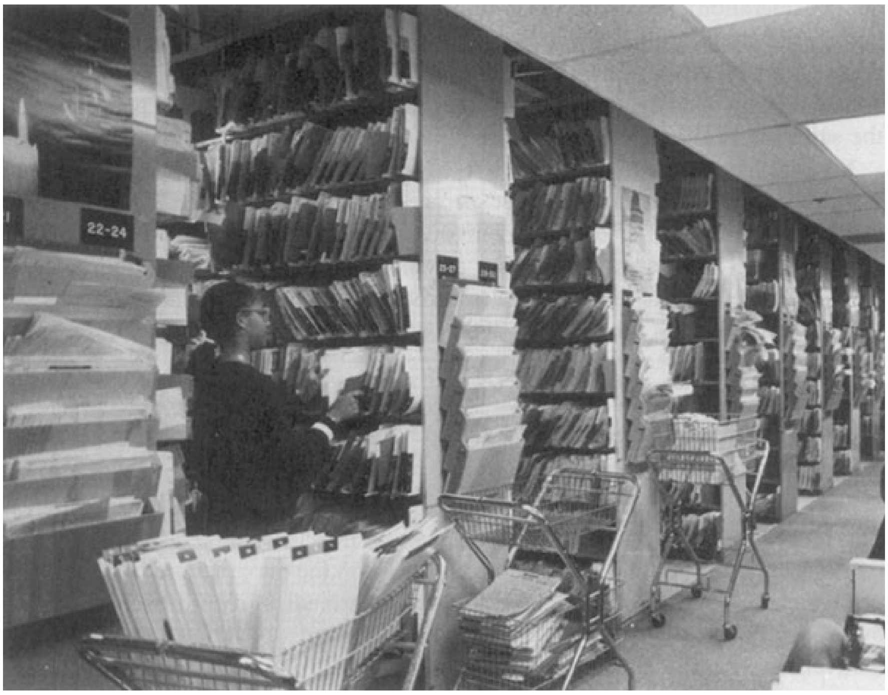
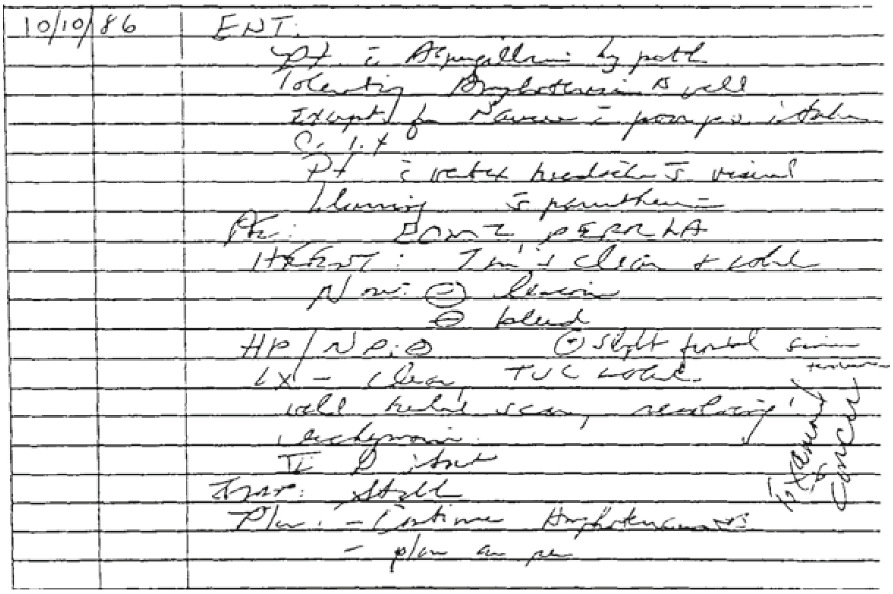


Fig. 2.8 A typical storage room for medical records. It is not surprising that charts sometimes were mislaid, and similarly clear why such paper repositories are being replaced as EHRs increasingly become the standard (Photograph courtesy of Janice Anne Rohn)



10/10/86 ENT:
Pt. c̄ Aspergillosis by path
Tolerating Amphotericin B well
Except for nausea c̄ poor po intake
C: 1.4
Pt c̄ retro headache s̄ visual
blurring s̄ paresthesia
PE: EOMI PERRLA
HEENT: TM's clear & [illegible]
Nose: ⊕ lesion
⊖ bleed
HP/NP: ⊖ ⊕ slight frontal sinus tenderness
Lx – clear TVC mobile
will follow scan, [illegible]
[illegible]
[illegible]
Imp: Stable
Plan: – Continue Amphotericin
– plan as per
[illegible]

Fig. 2.9 Written entries are standard in paper records, yet handwritten notes may be illegible. Notes that cannot be interpreted by other people due to illegibility may cause delays in treatment or inappropriate care—an issue that is largely eliminated when EHRs are used

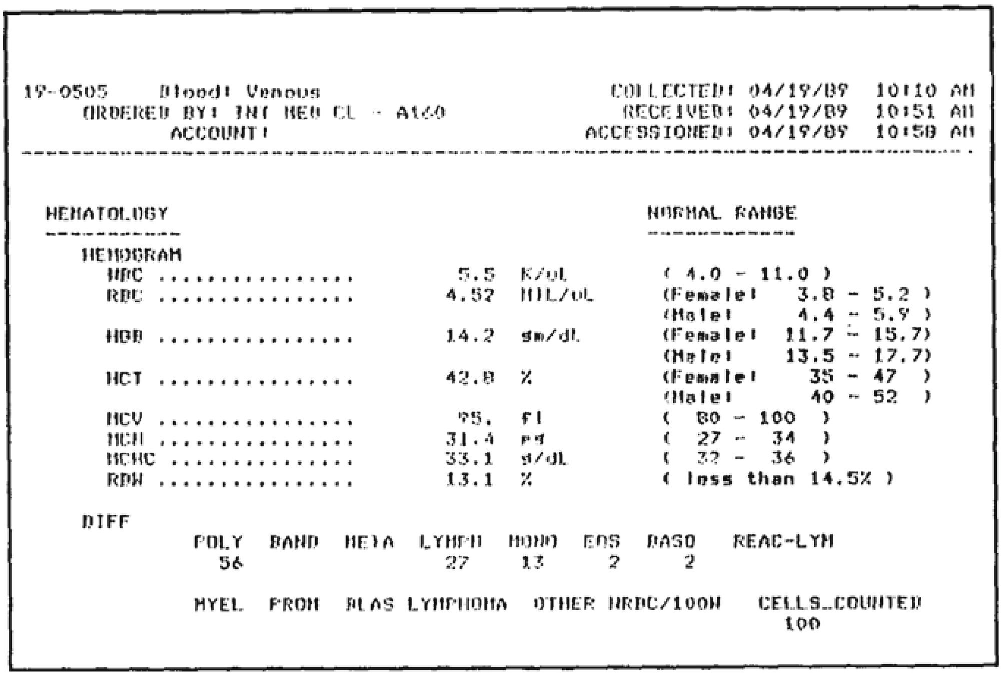
may be in early volumes that are stored offsite or are otherwise unavailable. Alternatively, an early volume may be mistaken for the most recent volume, misleading its users and resulting in documents being inserted out of sequence.

As described in Chap. 12, electronic health record systems offer solutions to all these practical problems in the use of the paper record. It is for this reason that more and more hospitals, health systems, and individual practitioners are implementing EHRs–further encouraged in the US by Federal incentive programs that help to cover the costs of EHR acquisition and maintenance (see Chaps. 1 and 27).

2.3.2 Redundancy and Inefficiency

To be able to find data quickly in the chart, health professionals have developed a variety of techniques that provide redundant recording to match alternate modes of access. For example, the result of a radiologic study typically is entered on a standard radiology reporting form, which is filed in the portion of the chart labeled "X-ray." For complicated procedures, the same data often are summarized in brief notes by radiologists in the narrative part of the chart, which they enter at the time of studies because they know that the formal report will not make it back to the chart for 1 or 2 days. In addition, the study results often are mentioned in notes written by the patient's admitting and consulting physicians and by the nursing staff. Although there may be good reasons for recording such information multiple times in different ways and in different locations within the chart, the combined bulk of these notes accelerates the physical growth of the document and, accordingly, complicates the chart's logistical management. Furthermore, it becomes increasingly difficult to locate specific patient data as the chart succumbs to "obesity". The predictable result is that someone writes yet another redundant entry, summarizing information that it took hours to track down.

A similar inefficiency occurs because of a tension between opposing goals in the design of reporting forms used by many laboratories. Most health personnel prefer a consistent, familiar paper form, often with color-coding, because it helps them to find information more quickly (Fig. 2.10). For example, a physician may know that a urinalysis report form is printed on yellow paper and records the bacteria count halfway down the middle column of the form. This knowledge allows the physician to work backward quickly in the laboratory section of the chart to find the most recent urinalysis sheet and to check at a glance the bacterial count. The problem is that such forms typically store only sparse information. It is clearly suboptimal if a rapidly growing physical chart is filled with sheets of paper that report only a single data element.



```
19-0505     Blood: Venous                        COLLECTED: 04/19/89  10:10 AM
    ORDERED BY: INT MED CL - A160                 RECEIVED: 04/19/89  10:51 AM
           ACCOUNT:                              ACCESSIONED: 04/19/89  10:58 AM
------------------------------------------------------------------------------------


  HEMATOLOGY                                     NORMAL RANGE
  ----------                                     ------------
    HEMOGRAM
      WBC ................        5.5   K/uL     ( 4.0 - 11.0 )
      RBC ................       4.52   MIL/uL   (Female:   3.8 - 5.2 )
                                                 (Male:     4.4 - 5.9 )
      HGB ................       14.2   gm/dL    (Female:  11.7 - 15.7)
                                                 (Male:    13.5 - 17.7)
      HCT ................       42.8   %        (Female:    35 - 47  )
                                                 (Male:      40 - 52  )
      MCV ................        95.   fl       (  80 - 100  )
      MCH ................       31.4   pg       (  27 -  34  )
      MCHC ...............       33.1   g/dL     (  32 -  36  )
      RDW ................       13.1   %        ( less than 14.5% )

    DIFF
             POLY  BAND  META  LYMPH  MONO  EOS  BASO   REAC-LYM
              56                27     13    2     2

             MYEL  PROM  BLAS LYMPHOMA  OTHER NRBC/100W   CELLS_COUNTED
                                                              100
```

Fig. 2.10 Laboratory reporting forms record medical data in a consistent, familiar format

2.3.3 Influence on Clinical Research

Anyone who has participated in a clinical research project based on chart review can attest to the tediousness of flipping through myriad medical records. For all the reasons described in Chap. 1, it is arduous to sit with stacks of patients' charts, extracting data and formatting them for structured statistical analysis, and the process is vulnerable to transcription errors. Observers often wonder how much medical knowledge is sitting untapped in paper medical records because there is no easy way to analyze experience across large populations of patients without first extracting pertinent data from those charts.

Suppose, for example, that physicians on a medical consultation service notice that patients receiving a certain common oral medication for diabetes (call it drug X) seem to be more likely to have significant postoperative hypotension (low blood pressure) than do surgical patients receiving other medications for diabetes. The doctors have based this hypothesis—that drug X influences postoperative blood pressure—on only a few recent observations, however, so they decide to look into existing hospital records to see whether this correlation has occurred with sufficient frequency to warrant a formal investigation. One efficient way to follow up on their theory from existing medical data would be to examine the hospital records of all patients who have diabetes and also have been admitted for surgery. The task would then be to examine those records (difficult and arduous with paper charts as will be discussed shortly, but subject to automated analysis in the case of EHRs) and to note for all patients (1) whether they were taking drug X when admitted and (2) whether they had postoperative hypotension. If the statistics showed that patients receiving drug X were more likely to have low blood pressure after surgery than were similar diabetic patients receiving alternate treatments, a controlled trial (prospective observation and data gathering) might well be appropriate.

Note the distinction between **retrospective chart review** to investigate a question that was not a subject of study at the time the data were collected and **prospective studies** in which the clinical hypothesis is known in advance and the **research protocol** is designed specifically to collect future data that are relevant to the question under consideration (see also Chaps. 11 and 26). Subjects are assigned **randomly** to different study groups to help prevent researchers—who are bound to be biased, having developed the hypothesis—from unintentionally skewing the results by assigning a specific class of patients all to one group. For the same reason, to the extent possible, the studies are **double blind**; i.e., neither the researchers nor the subjects know which treatment is being administered. Such blinding is of course impractical when it is obvious to patients or physicians what therapy is being given (such as surgical procedures versus drug therapy). Prospective, randomized, double-blind studies are considered the best method for determining optimal management of disease, but it is often impractical to carry out such studies, and then methods such as retrospective chart review are used.

Returning to our example, consider the problems in paper chart review that the researchers would encounter in addressing the postoperative hypotension question retrospectively. First, they would have to identify the charts of interest: the subset of medical records dealing with surgical patients who are also diabetic. In a hospital record room filled with thousands of charts, the task of chart selection can be overwhelming. Medical records departments generally do keep indexes of diagnostic and procedure codes cross-referenced to specific patients (see Sect. 2.5.1). Thus, it might be possible to use such an index to find all charts in which the discharge diagnoses included diabetes and the procedure codes included major surgical procedures. The researcher might compile a list of patient identification numbers and have the individual charts pulled from the file room for review.

The researchers' next task is to examine each chart serially to find out what treatment each patient was receiving for diabetes at the time of the surgery and to determine whether the patient had postoperative hypotension. Finding such information may be extremely time-consuming. Where should the researcher look for it? The admission drug orders might show what the patient received for diabetes control, but it would also be wise to check the medication sheets to see whether the therapy was also administered (as well as ordered) and the admission history to see whether a routine treatment for diabetes, taken right up until the patient entered the hospital, was not administered during the inpatient stay. Information about hypotensive episodes might be similarly difficult to locate. The researchers might start with nursing notes from the recovery room or with the anesthesiologist's datasheets from the operating room, but the patient might not have been hypotensive until after leaving the recovery room and returning to the ward. So the nursing notes from the ward need to be checked too, as well as vital signs sheets, physicians' progress notes, and the discharge summary.

It should be clear from this example that retrospective paper chart review is a laborious and tedious process and that people performing it are prone to make transcription errors and to overlook key data. One of the great appeals of EHRs (Chap. 12) is their ability to facilitate the chart review process. They obviate the need to retrieve hard copy charts; instead, researchers can use computer-based data retrieval and analysis techniques to do most of the work (finding relevant patients, locating pertinent data, and formatting the information for statistical analyses). Researchers can use similar techniques to harness computer assistance with data management in prospective clinical trials (Chap. 26).

2.3.4 The Passive Nature of Paper Records

The traditional manual system has another limitation that would have been meaningless until the emergence of the computer age. A manual archival system is inherently passive; the charts sit waiting for something to be done with them. They are insensitive to the characteristics of the data recorded within their pages, such as legibility, accuracy, or implications for patient management. They cannot take an active role in responding appropriately to those implications.

Increasingly, EHR systems have changed our perspective on what health professionals can expect from the medical chart. Automated record systems introduce new opportunities for dynamic responses to the data that are recorded in them. As described in many of the chapters to follow, computational techniques for data storage, retrieval, and analysis make it feasible to develop record systems that (1) monitor their contents and generate warnings or advice for providers based on single observations or on logical combinations of data; (2) provide automated quality control, including the flagging of potentially erroneous data; or (3) provide feedback on patient-specific or population-based deviations from desirable standards.

2.4 New Kinds of Data and the Resulting Challenges

The revolution in human genetics that emerged with the **Human Genome Project** in the 1990s is already having a profound effect on the diagnosis, prognosis, and treatment of disease (Palotie et al. 2013). The vast amounts of data that are generated in biomedical research (see Chaps. 24 and 25), and that can be pooled from patient datasets to support clinical research (Chap. 26) and public health (Chap. 16), have created new challenges as well as opportunities. Researchers are finding that the amount of data that they must manage and assess has become so large that they often find that they lack either the capabilities or expertise to handle the analytics that are required. This problem, sometimes dubbed the "big data" problem, has gathered the attention of government funding agencies as well (Mervis 2012; NSF-NIH Interagency Initiative 2012). Some suggest that the genetic material itself will become our next-generation method for storing large amounts of data (Church et al. 2012). Data analytics, and the management of large amounts of genomic/proteomic or clinical/public-health data, have accordingly become major research topics and key opportunities for new methodology development by biomedical informatics scientists (Ohno-Machado 2012).

The issues that arise are practical as well as scientifically interesting. For example, developers of EHRs have begun to grapple with questions regarding how they might be store an individual's personal genome with the electronic health record. New standards will be required, and tactical questions need answering regarding, for example, whether to store an entire genome or only those components (e.g., genetic markers) that are already understood (Masys et al. 2012). In cancer, for example, where mutations in cell lines can occur, an individual may actually have many genomes represented among his or her cells. These issues will undoubtedly influence the evolution of data systems and EHRs, as well as the growth of **personalized medicine**, in the years ahead.

2.5 The Structure of Clinical Data

Scientific disciplines generally develop a precise terminology or notation that is standardized and accepted by all workers in the field. Consider, for example, the universal language of chemistry embodied in chemical formulae, the precise definitions and mathematical equations used by physicists, the predicate calculus used by logicians, or the conventions for describing circuits used by electrical engineers. Medicine is remarkable for its failure to develop a widely accepted standardized vocabulary and **nomenclature**, and many observers believe that a true scientific basis for the field will be impossible until this problem is addressed (see Chap. 7). Other people argue that common references to the "art of medicine" reflect an important distinction between medicine and the "hard" sciences; these people question whether it is possible to introduce too much standardization into a field that prides itself in humanism.

The debate has been accentuated by the introduction of computers for data management, because such machines tend to demand conformity to data standards and definitions. Otherwise, issues of data retrieval and analysis are confounded by discrepancies between the meanings intended by the observers or recorders and those intended by the individuals retrieving information or doing data analysis. What is an "upper respiratory infection"? Does it include infections of the trachea or of the main stem bronchi? How large does the heart have to be before we can refer to "cardiomegaly"? How should we deal with the plethora of disease names based on eponyms (e.g., Alzheimer's disease, Hodgkin's disease) that are not descriptive of the illness and may not be familiar to all practitioners? What do we mean by an "acute abdomen"? Are the boundaries of the abdomen well agreed on? What are the time constraints that correspond to "acuteness" of abdominal pain? Is an "ache" a pain? What about "occasional" cramping?

Imprecision and the lack of a standardized vocabulary are particularly problematic when we wish to aggregate data recorded by multiple health professionals or to analyze trends over time. Without a controlled, predefined vocabulary, data interpretation is inherently complicated, and the automatic summarization of data may be impossible. For example, one physician might note that a patient has "shortness of breath." Later, another physician might note that she has "dyspnea." Unless these terms are designated as synonyms, an automated program will fail to indicate that the patient had the same problem on both occasions.

Regardless of arguments regarding the "artistic" elements in medicine, the need for health personnel to communicate effectively is clear both in acute care settings and when patients are seen over long periods. Both high-quality care and scientific progress depend on some standardization in terminology. Otherwise, differences in intended meaning or in defining criteria will lead to miscommunication, improper interpretation, and potentially negative consequences for the patients involved.

Given the lack of formal definitions for many medical terms, it is remarkable that medical workers communicate as well as they do. Only occasionally is the care for a patient clearly compromised by miscommunication. If EHRs are to become dynamic and responsive manipulators of patient data, however, their encoded logic must be able to presume a specific meaning for the terms and data elements entered by the observers. This point is discussed in greater detail in Chap. 7, which deals in part with the multiple efforts to develop health care-computing standards, including a shared, controlled terminology for biomedicine.

2.5.1 Coding Systems

We are used to seeing figures regarding the growing incidences of certain types of tumors, deaths from influenza during the winter months, and similar health statistics that we tend to take for granted. How are such data accumulated? Their role in health planning and health care financing is clear, but if their accumulation required chart review through the process described earlier in this chapter, we would know much less about the health status of the populations in various communities (see Chap. 16).

Because of the needs to know about health trends for populations and to recognize epidemics in their early stages, there are various health-reporting requirements for hospitals (as well as other public organizations) and practitioners. For example, cases of gonorrhea, syphilis, and tuberculosis generally must be reported to local public-health organizations, which code the data to allow trend analyses over time. The Centers for Disease Control and Prevention in Atlanta (CDC) then pool regional data and report national as well as local trends in disease incidence, bacterial-resistance patterns, etc.

Another kind of reporting involves the coding of all discharge diagnoses for hospitalized patients, plus coding of certain procedures (e.g., type of surgery) that were performed during the hospital stay. Such codes are reported to state and federal health-planning and analysis agencies and also are used internally at the institution for case-mix analysis (determining the relative frequencies of various disorders in the hospitalized population and the average length of stay for each disease category) and for research. For such data to be useful, the codes must be well defined as well as uniformly applied and accepted.

The World health Organization publishes adiagnostic coding scheme called the International Classification of Disease (ICD). The 10th revision of this standard, ICD10,[3] is currently in use in much of the world, although in the United States a derivative of the previous version, the *International Classification of Diseases, 9th*

[3] http://www.icd10data.com/ (Accessed 12/2/2012).

Edition – Clinical Modifications (*ICD9-CM*), is still transitioning to the new version (see Chap. 7). ICD9-CM is used by all nonmilitary hospitals in the United States for discharge coding, and must be reported on the bills submitted to most insurance companies (Fig. 2.11). Pathologists have developed another widely used diagnostic coding scheme; originally known as Systematized Nomenclature of Pathology (SNOP), it was expanded to the Systematized Nomenclature of Medicine (SNOMED) (Côté and Rothwell 1993; American College of Pathologists 1982) and then merged with the Read Clinical Terms from the Great Britain to become

J45 Asthma
Includes: allergic (predominantly) asthma, allergic bronchitis NOS, allergic rhinitis with asthma, atopic asthma, extrinsic allergic asthma, hay fever with asthma, idiosyncratic asthma, intrinsic nonallergic asthma, nonallergic asthma

Use additional code to identify: exposure to environmental tobacco smoke (Z77.22), exposure to tobacco smoke in the perinatal period (P96.81), history of tobacco use (Z87.891), occupational exposure to environmental tobacco smoke (Z57.31), tobacco dependence (F17.-), tobacco use (Z72.0)

Excludes: detergent asthma (J69.8), eosinophilic asthma (J82), lung diseases due to external agents (J60-J70), miner's asthma (J60), wheezing NOS (R06.2), wood asthma (J67.8), asthma with chronic obstructive pulmonary disease (J44.9), chronic asthmatic (obstructive) bronchitis (J44.9), chronic obstructive asthma (J44.9)

```
J45.2  Mild intermittent asthma
   J45.20  Mild intermittent asthma, uncomplicated
            Mild intermittent asthma NOS
   J45.21  Mild intermittent asthma with (acute) exacerbation
   J45.22  Mild intermittent asthma with status asthmaticus
J45.3  Mild persistent asthma
   J45.30  Mild persistent asthma, uncomplicated
           Mild persistent asthma NOS
   J45.31  Mild persistent asthma with (acute) exacerbation
   J45.32  Mild persistent asthma with status asthmaticus
J45.4  Moderate persistent asthma
   J45.40  Moderate persistent asthma, uncomplicated
           Moderate persistent asthma NOS
   J45.41  Moderate persistent asthma with (acute) exacerbation
   J45.42  Moderate persistent asthma with status asthmaticus
J45.5  Severe persistent asthma
   J45.50  Severe persistent asthma, uncomplicated
           Severe persistent asthma NOS
   J45.51  Severe persistent asthma with (acute) exacerbation
   J45.52  Severe persistent asthma with status asthmaticus
J45.9  Other and unspecified asthma
   J45.90  Unspecified asthma
           Asthmatic bronchitis NOS
           Childhood asthma NOS
           Late onset asthma
      J45.901  Unspecified asthma with (acute) exacerbation
      J45.902  Unspecified asthma with status asthmaticus
      J45.909  Unspecified asthma, uncomplicated
               Asthma NOS
   J45.99  Other asthma
   J45.990  Exercise induced bronchospasm
   J45.991  Cough variant asthma
   J45.998  Other asthma
```

Fig. 2.11 The subset of disease categories for asthma taken from ICD-10-CM, the new diagnosis coding system that is being developed as a replacement for ICD-9-CM, Volumes 1 and 2 (Source: Centers for Medicare and Medicaid Services, US Department of Health and Human Services, http://www.cms.gov/Medicare/Coding/ICD10/2013-ICD-10-CM-and-GEMs.html, accessed September 11, 2013)

SNOMED-CT (Stearns et al. 2001). In recent years, support for SNOMED-CT, has been assumed by the International Health Terminology Standards Development Organization, based in Copenhagen.[4] Another coding scheme, developed by the American Medical Association, is the Current Procedural Terminology (CPT) (Finkel 1977). It is similarly widely used in producing bills for services rendered to patients. More details on such schemes are provided in Chap. 7. What warrants emphasis here, however, is the motivation for the codes' development: health care personnel need standardized terms that can support pooling of data for analysis and can provide criteria for determining charges for individual patients.

The historical roots of a coding system reveal themselves as limitations or idiosyncrasies when the system is applied in more general clinical settings. For example, ICD9-CM was derived from a classification scheme developed for epidemiologic reporting. Consequently, it has more than 500 separate codes for describing tuberculosis infections. SNOMED versions have long permitted coding of pathologic findings in exquisite detail but only in later years began to introduce codes for expressing the dimensions of a patient's functional status. In a particular clinical setting, none of the common coding schemes is likely to be completely satisfactory. In some cases, the granularity of the code will be too coarse; on the one hand, a hematologist (person who studies blood diseases) may want to distinguish among a variety of hemoglobinopathies (disorders of the structure and function of hemoglobin) lumped under a single code in ICD8-CM. On the other hand, another practitioner may prefer to aggregate many individual codes—e.g., those for active tuberculosis—into a single category to simplify the coding and retrieval of data.

Such schemes cannot be effective unless health care providers accept them. There is an inherent tension between the need for a coding system that is general enough to cover many different patients and the need for precise and unique terms that accurately apply to a specific patient and do not unduly constrain physicians' attempts to describe what they observe. Yet if physicians view the EHR as a blank sheet of paper on which any unstructured information can be written, the data they record will be unsuitable for dynamic processing, clinical research, and health planning. The challenge is to learn how to meet all these needs. Researchers at many institutions have worked for over two decades to develop a unified medical language system (UMLS), a common structure that ties together the various vocabularies that have been created. At the same time, the developers of specific terminologies are continually working to refine and expand their independent coding schemes (Humphreys et al. 1998) (see Chap. 7).

2.5.2 The Data-to-Knowledge Spectrum

A central focus in bio medical informatics is the information base that constitutes the "substance of medicine." Workers in the field have tried to clarify the distinctions among three terms frequently used to describe the content of computer-based systems: data, information, and knowledge (Blum 1986b; Bernstam et al. 2010). These terms are often used interchangeably. In this volume, we shall refer to a **datum** as a single observational point that characterizes a relationship.[5] It generally can be regarded as the value of a specific parameter for a particular object (e.g., a patient) at a given point in time. The term **information** refers to analyzed data that have been suitably curated and organized so that they have meaning. Data do not constitute information until they have been organized in some way, e.g., for analysis or display. **Knowledge**, then, is derived through the formal or informal analysis (or interpretation) of information that was in turn derived from data. Thus, knowledge includes the results of formal studies and also common sense facts, assumptions, heuristics (strategic rules of thumb), and models—any of which may reflect the expe-

[4] http://www.ihtsdo.org/ (Accessed 12/2/2012).

[5] Note that *data* is a plural term, although it is often erroneously used in speech and writing as though it were a collective (singular) noun.

rience or biases of people who interpret the primary data and the resulting information.

The observation that patient Brown has a blood pressure of 180/110 is a *datum*, as is the report that the patient has had a myocardial infarction (heart attack). When researchers pool such data, creating information, subsequent analysis may determine that patients with high blood pressure are more likely to have heart attacks than are patients with normal or low blood pressure. This analysis of organized data (information) has produced a piece of knowledge about the world. A physician's belief that prescribing dietary restriction of salt is unlikely to be effective in controlling high blood pressure in patients of low economic standing (because the latter are less likely to be able to afford special low-salt foods) is an additional personal piece of *knowledge*—a **heuristic** that guides physicians in their decision making. Note that the appropriate interpretation of these definitions depends on the context. Knowledge at one level of abstraction may be considered data at higher levels. A blood pressure of 180/110 mmHg is a raw piece of data; the statement that the patient has hypertension is an interpretation of several such data and thus represents a higher level of information. As input to a diagnostic decision aid, however, the presence or absence of hypertension may be requested, in which case the presence of hypertension is treated as a data item.

A **database** is a collection of individual observations without any summarizing analysis. An EHR system is thus primarily viewed as a database—the place where patient data are stored. When properly collated and pooled with other data, these elements in the EHR provide *information* about the patient. A **knowledge base**, on the other hand, is a collection of facts, heuristics, and models that can be used for problem solving and analysis of organized data (information). If the knowledge base provides sufficient structure, including semantic links among knowledge items, the computer itself may be able to apply that knowledge as an aid to case-based problem solving. Many decision-support systems have been called knowledge-based systems, reflecting this distinction between knowledge bases and databases (see Chap. 22).

2.6 Strategies of Clinical Data Selection and Use

It is illusory to conceive of a "complete clinical data set." All medical databases, and medical records, are necessarily incomplete because they reflect the selective collection and recording of data by the health care personnel responsible for the patient. There can be marked interpersonal differences in both style and problem solving that account for variations in the way practitioners collect and record data for the same patient under the same circumstances. Such variations do not necessarily reflect good practices, however, and much of medical education is directed at helping physicians and other health professionals to learn what observations to make, how to make them (generally an issue of technique), how to interpret them, and how to decide whether they warrant formal recording.

An example of this phenomenon is the difference between the first medical history, physical examination, and summarizing report developed by a medical student and the similar process undertaken by a seasoned clinician examining the same patient. Medical students tend to work from comprehensive mental outlines of questions to ask, physical tests to perform, and additional data to collect. Because they have not developed skills of selectivity, the process of taking a medical history and performing a physical examination may take more than 1 h, after which students develop extensive reports of what they observed and how they have interpreted their observations. It clearly would be impractical, inefficient, and inappropriate for physicians in practice to spend this amount of time assessing every new patient. Thus, part of the challenge for the neophyte is to learn how to ask only the questions that are necessary, to perform only the examination components that are required, and to record only those data that will be pertinent in justifying the ongoing diagnostic approach and in guiding the future management of the patient.

What do we mean by **selectivity** in data collection and recording? It is precisely this process that often is viewed as a central part of the "art of medicine," an element that accounts for individual styles and the sometimes marked distinctions among clinicians. As is discussed with numerous clinical examples in Chaps. 3 and 4, the idea of selectivity implies an ongoing decision-making process that guides data collection and interpretation. Attempts to understand how expert clinicians internalize this process, and to formalize the ideas so that they can better be taught and explained, are central in biomedical informatics research. Improved guidelines for such decision making, derived from research activities in biomedical informatics, not only are enhancing the teaching and practice of medicine (Shortliffe 2010) but also are providing insights that suggest methods for developing computer-based decision-support tools.

2.6.1 The Hypothetico-Deductive Approach

Studies of clinical decision makers have shown that strategies for data collection and interpretation may be imbedded in an iterative process known as the **hypothetico-deductive approach** (Elstein et al. 1978; Kassirer and Gorry 1978). As medical students learn this process, their data collection becomes more focused and efficient, and their medical records become more compact. The central idea is one of sequential, staged data collection, followed by data interpretation and the generation of hypotheses, leading to hypothesis-directed selection of the next most appropriate data to be collected. As data are collected at each stage, they are added to the growing database of observations and are used to reformulate or refine the active hypotheses. This process is iterated until one hypothesis reaches a threshold level of certainty (e.g., it is proved to be true, or at least the uncertainty is reduced to a satisfactory level). At that point, a management, disposition, or therapeutic decision can be made.

The diagram in Fig. 2.12 clarifies this process. As is shown, data collection begins when the patient presents to the physician with some issue (a symptom or disease, or perhaps the need for routine care). The physician generally responds with a few questions that allow one to focus rapidly on the nature of the problem. In the written report, the data collected with these initial questions typically are recorded as the patient identification, chief complaint, and initial portion of the history of the present illness. Studies have shown that an experienced physician will have an initial set of hypotheses (theories) in mind after hearing the patient's response to the first six or seven questions (Elstein et al. 1978). These hypotheses then serve as the basis for selecting additional questions. As shown in Fig. 2.12, answers to these additional questions allow the physician to refine hypotheses about the source of the patient's problem. Physicians refer to the set of active hypotheses as the **differential diagnosis** for a patient; the differential diagnosis comprises the set of possible diagnoses among which the physician must distinguish to determine how best to administer treatment.

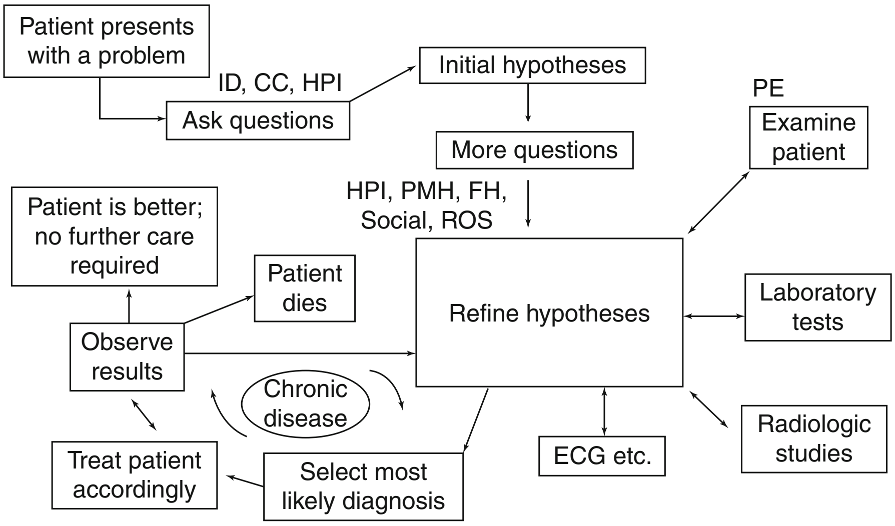


Fig. 2.12 A schematic view of the hypothetico-deductive approach. The process of medical data collection and treatment is intimately tied to an ongoing process of hypothesis generation and refinement. See text for full discussion. *ID* patient identification, *CC* chief complaint, *HPI* history of present illness, *PMH* past medical history, *FH* family history, *Social* social history, *ROS* review of systems, *PE* physical examination

Note that the question selection process is inherently heuristic; e.g., it is personalized and efficient, but it is not guaranteed to collect every piece of information that might be pertinent. Human beings use heuristics all the time in their decision making because it often is impractical or impossible to use an exhaustive problem-solving approach. A common example of heuristic problem solving is the playing of a complex game such as chess. Because it would require an enormous amount of time to define all the possible moves and countermoves that could ensue from a given board position, expert chess players develop personal heuristics for assessing the game at any point and then selecting a strategy for how best to proceed. Differences among such heuristics account in part for variations in observed expertise.

Physicians have developed safety measures, however, to help them to avoid missing important issues that they might not discover when collecting data in a hypothesis-directed fashion when taking the history of a patient's present illness (Pauker et al. 1976). These measures tend to be focused in four general categories of questions that follow the collection of information about the chief complaint: past medical history, family history, social history, and a brief **review of systems** in which the physician asks some general questions about the state of health of each of the major organ systems in the body. Occasionally, the physician discovers entirely new problems or finds important information that modifies the hypothesis list or modulates the treatment options available (e.g., if the patient reports a serious past drug reaction or allergy).

When physicians have finished asking questions, the refined hypothesis list (which may already be narrowed to a single diagnosis) then serves as the basis for a focused physical examination. By this time, physicians may well have expectations of what they will find on examination or may have specific tests in mind that will help them to distinguish among still active hypotheses about diseases based on the questions that they have asked. Once again, as in the question-asking process, focused hypothesis-directed examination is augmented with general tests that occasionally turn up new abnormalities and generate hypotheses that the physician did not expect on the basis of the medical history alone. In addition, unexplained findings on examination may raise issues that require additional history taking. Thus, the asking of questions generally is partially integrated with the examination process.

When physicians have completed the physical examination, their refined hypothesis list may be narrowed sufficiently for them to undertake specific treatment. Additional data gathering may still be necessary, however. Such testing is once again guided by the current hypotheses. The options available include laboratory tests (of blood, urine, other body fluids, or biopsy specimens), radiologic studies (X-ray examinations, nuclear-imaging scans, computed tomography (CT) studies, magnetic resonance scans, sonograms, or any of a number of other imaging modalities), and other specialized tests (electrocardiograms (ECGs), electroencephalograms, nerve conduction studies, and many others), as well as returning to the patient to ask further questions or perform additional physical examination. As the results of such studies become available, physicians constantly revise and refine their hypothesis list.

Ultimately, physicians are sufficiently certain about the source of a patient's problem to be able to develop a specific management plan. Treatments are administered, and the patient is observed. Note data collected to measure response to treatment may themselves be used to synthesize information that affects the hypotheses about a patient's illness. If patients do not respond to treatment, it may mean that their disease is resistant to that therapy and that their physicians should try an alternate approach, or it may mean that the initial diagnosis was incorrect and that physicians should consider alternate explanations for the patient's problem.

The patient may remain in a cycle of treatment and observation for a long time, as shown in Fig. 2.12. This long cycle reflects the nature of chronic-disease management—an aspect of medical care that is accounting for an increasing proportion of the health care community's work (and an increasing proportion of health care cost).

Alternatively, the patient may recover and no longer need therapy, or he or she may die. Although the process outlined in Fig. 2.12 is oversimplified in many regards, it is generally applicable to the process of data collection, diagnosis, and treatment in most areas of medicine.

Note that the hypothesis-directed process of data collection, diagnosis, and treatment is inherently knowledge-based. It is dependent not only on a significant fact base that permits proper interpretation of data and selection of appropriate follow-up questions and tests but also on the effective use of heuristic techniques that characterize individual expertise.

Another important issue, addressed in Chap. 3, is the need for physicians to balance financial costs and health risks of data collection against the perceived benefits to be gained when those data become available. It costs nothing but time to examine the patient at the bedside or to ask an additional question, but if the data being considered require, for example, X-ray exposure, coronary angiography, or a CT scan of the head (all of which have associated risks and costs), then it may be preferable to proceed with treatment in the absence of full information. Differences in the assessment of cost-benefit trade-offs in data collection, and variations among individuals in their willingness to make decisions under uncertainty, often account for differences of opinion among collaborating physicians.

2.6.2 The Relationship Between Data and Hypotheses

We wrote rather glibly in Sect. 2.6.1 about the "generation of hypotheses from data"; now we need to ask: What precisely is the nature of that process? As is discussed in Chap. 4, researchers with a psychological orientation have spent much time trying to understand how expert problem solvers evoke hypotheses (Pauker et al. 1976; Elstein et al. 1978; Pople 1982) and the traditional probabilistic decision sciences have much to say about that process as well. We provide only a brief introduction to these ideas here; they are discussed in greater detail in Chaps. 3 and 4.

When an observation evokes a hypothesis (e.g., when a clinical finding makes a specific diagnosis come to mind), the observation presumably has some close association with the hypothesis. What might be the characteristics of that association? Perhaps the finding is almost always observed when the hypothesis turns out to be true. Is that enough to explain hypothesis generation? A simple example will show that such a simple relationship is not enough to explain the evocation process. Consider the hypothesis that a patient is pregnant and the observation that the patient is female. Clearly, all pregnant patients are female. When a new patient is observed to be female, however, the possibility that the patient is pregnant is not immediately evoked. Thus, female gender is a highly sensitive indicator of pregnancy (there is a 100 % certainty that a pregnant patient is female), but it is not a good predictor of pregnancy (most females are not pregnant). The idea of **sensitivity**—the likelihood that a given datum will be observed in a patient with a given disease or condition—is an important one, but it will not alone account for the process of hypothesis generation in medical diagnosis.

Perhaps the clinical manifestation seldom occurs unless the hypothesis turns out to be true; is that enough to explain hypothesis generation? This idea seems to be a little closer to the mark. Suppose a given datum is never seen unless a patient has a specific disease. For example, a Pap smear (a smear of cells swabbed from the cervix, at the opening to the uterus, treated with Papanicolaou's stain, and then examined under the microscope) with grossly abnormal cells (called class IV findings) is never seen unless the woman has cancer of the cervix or uterus. Such tests are called **pathognomonic**. Not only do they evoke a specific diagnosis but they also immediately prove it to be true. Unfortunately, there are few pathognomonic tests in medicine and they are often of relatively low sensitivity (that is, although having a particular test result makes the diagnosis, few patients with the condition actually have that finding).

More commonly, a feature is seen in one disease or disease category more frequently than it is in others, but the association is not absolute.

For example, there are few disease entities other than infections that elevate a patient's white blood cell count. Certainly it is true, for example, that leukemia can raise the white blood cell count, as can the use of the drug prednisone, but most patients who do not have infections will have normal white blood cell counts. An elevated white count therefore does not prove that a patient has an infection, but it does tend to evoke or support the hypothesis that an infection is present. The word used to describe this relationship is **specificity**. An observation is highly specific for a disease if it is generally not seen in patients who do not have that disease. A pathognomonic observation is 100% specific for a given disease. When an observation is highly specific for a disease, it tends to evoke that disease during the diagnostic or data-gathering process.

By now, you may have realized that there is a substantial difference between a physician viewing test results that evoke a disease hypothesis and that physician being willing to act on the disease hypothesis. Yet even experienced physicians sometimes fail to recognize that, although they have made an observation that is highly specific for a given disease, it may still be more likely that the patient has other diseases (and does not have the suspected one) unless (1) the finding is pathognomonic or (2) the suspected disease is considerably more common than are the other diseases that can cause the observed abnormality. This mistake is one of the most common errors of intuition that has been identified in the medical decision-making process. To explain the basis for this confusion in more detail, we must introduce two additional terms: prevalence and predictive value.

The **prevalence** of a disease is simply the percentage of a population of interest that has the disease at any given time. A particular disease may have a prevalence of only 5 % in the general population (1 person in 20 will have the disease) but have a higher prevalence in a specially selected subpopulation. For example, black-lung disease has a low prevalence in the general population but has a much higher prevalence among coal miners, who develop black lung from inhaling coal dust. The task of diagnosis therefore involves updating the probability that a patient has a disease from the **baseline rate** (the prevalence in the population from which the patient was selected) to a post-test probability that reflects the test results. For example, the probability that any given person in the United States has lung cancer is low (i.e., the prevalence of the disease is low), but the chance increases if his or her chest X-ray examination shows a possible tumor. If the patient were a member of the population composed of cigarette smokers in the United States, however, the prevalence of lung cancer would be higher. In this case, the identical chest X-ray report would result in an even higher updated probability of lung cancer than it would had the patient been selected from the population of all people in the United States.

The **predictive value** (**PV**) of a test is simply the post-test (updated) probability that a disease is present based on the results of a test. If an observation supports the presence of a disease, the PV will be greater than the prevalence (also called the pretest risk). If the observation tends to argue against the presence of a disease, the PV will be lower than the prevalence. For any test and disease, then, there is one PV if the test result is positive and another PV if the test result is negative. These values are typically abbreviated PV+ (the PV of a positive test) and PV− (the PV of a negative test).

The process of hypothesis generation in medical diagnosis thus involves both the evocation of hypotheses and the assignment of a likelihood (probability) to the presence of a specific disease or disease category. The PV of a positive test depends on the test's sensitivity and specificity, as well as the prevalence of the disease. The formula that describes the relationship precisely is:

$$\text{PV+} = \frac{(\text{sensitivity})(\text{prevalence})}{(\text{sensitivity})(\text{prevalence}) + (1 - \text{specificity})(1 - \text{prevalence})}$$

There is a similar formula for defining PV− in terms of sensitivity, specificity, and prevalence. Both formulae can be derived from simple probability theory. Note that positive tests with high sensitivity and specificity may still lead to a low post-test probability of the disease (PV+) if the prevalence of that disease is low. You should substitute values in the PV + formula to convince yourself that this assertion is true. It is this relationship that tends to be poorly understood by practitioners and that often is viewed as counterintuitive (which shows that your intuition can misguide you!). Note also (by substitution into the formula) that test sensitivity and disease prevalence can be ignored only when a test is pathognomonic (i.e., when its specificity is 100 %, which mandates that PV+ be 100 %). The PV+ formula is one of many forms of **Bayes**' **theorem**, a rule for combining probabilistic data that is generally attributed to the work of Reverend Thomas Bayes in the 1700s. Bayes' theorem is discussed in greater detail in Chap. 3.

2.6.3 Methods for Selecting Questions and Comparing Tests

We have described the process of hypothesis-directed sequential data collection and have asked how an observation might evoke or refine the physician's hypotheses about what abnormalities account for the patient's illness. The complementary question is: Given a set of current hypotheses, how does the physician decide what additional data should be collected? This question also has been analyzed at length (Elstein et al. 1978; Pople 1982) and is pertinent for computer programs that gather data efficiently to assist clinicians with diagnosis or with therapeutic decision making (see Chap. 22). Because understanding issues of test selection and data interpretation is crucial to understanding medical data and their uses, we devote Chap. 3 to these and related issues of medical decision making. In Sect. 3.6, for example, we discuss the use of decision-analytic techniques in deciding whether to treat a patient on the basis of available information or to perform additional diagnostic tests.

2.7 The Computer and Collection of Medical Data

Although this chapter has not directly discussed computer systems, the potential role of the computer in medical data storage, retrieval, and interpretation should be clear. Much of the rest of this book deals with specific applications in which the computer's primary role is data management. One question is pertinent to all such applications: How do you get the data into the computer in the first place?

The need for data entry by physicians has posed a problem for medical-computing systems since the earliest days of the field. Awkward or nonintuitive interactions at computing devices—particularly ones requiring keyboard typing or confusing movement through multiple display screens by the physician—have probably done more to inhibit the clinical use of computers than have any other factor. Doctors, and many other health care staff, sometimes simply refuse to use computers because of the awkward interfaces that are imposed.

A variety of approaches have been used to try to finesse this problem. One is to design systems such that clerical staff can do essentially all the data entry and much of the data retrieval as well. Many clinical research systems (see Chap. 26) have taken this approach. Physicians may be asked to fill out structured paper datasheets, or such sheets may be filled out by data abstractors who review patient charts, but the actual entry of data into the database is done by paid transcriptionists.

In some applications, it is possible for data to be entered automatically into the computer by the device that measures or collects them. For example, monitors in intensive care or coronary care units, pulmonary function or ECG machines, and measurement equipment in the clinical chemistry laboratory can interface directly with a computer in which a database is stored. Certain data can be entered directly by patients; there are systems, for example, that take the patient's history by presenting on a computer screen or tablet multiple-choice questions that follow a branching logic. The patient's responses to the questions are used to generate electronic or hard copy reports for physicians and also may be stored directly in a computer database for subsequent use in other settings.

When physicians or other health personnel do use the machine themselves, specialized devices often allow rapid and intuitive operator–machine interaction. Most of these devices use a variant of the "point-and-select" approach—e.g., touch-sensitive computer screens, mouse-pointing devices, and increasingly the clinician's finger on a mobile tablet or smart phone (see Chap. 5). When conventional computer workstations are used, specialized keypads can be helpful. Designers frequently permit logical selection of items from menus displayed on the screen so that the user does not need to learn a set of specialized commands to enter or review data. There were clear improvements when handheld tablets using pen-based or finger-based mechanisms for data entry were introduced. With ubiquitous wireless data services, such devices are allowing clinicians to maintain normal mobility (in and out of examining rooms or inpatient rooms) while accessing and entering data that are pertinent to a patient's care.

These issues arise in essentially all application areas, and, because they can be crucial to the successful implementation and use of a system, they warrant particular attention in system design. As more physicians are becoming familiar with computers at home, they will find the use of computers in their practice less of a hindrance. We encourage you to consider human–computer interaction, and the cognitive issues that arise in dealing with computer systems (see Chap. 4), as you learn about the application areas and the specific systems described in later chapters.

Suggested Readings

Bernstam, E. V., Smith, J. W., & Johnson, T. R. (2010). What is biomedical informatics? *Journal of Biomedical Informatics, 43*(1), 104–110. The authors discuss the transformation of data into information and knowledge, delineating the ways in which this focus lies at the heart of the field of biomedical informatics.

Klasnja, P., & Pratt, W. (2012). Healthcare in the pocket: mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics, 45*(1), 184–198. This review article describes the multiple ways in which both patients and providers are being empowered through the introduction of affordable mobile technologies that manage data and apply knowledge to generate advice.

Ohno-Machado, L. (2012). Big science, big data, and a big role for biomedical informatics. *Journal of the American Medical Informatics Association, 19*, e1. This editorial introduces a special online issue of the Journal of the American Medical Informatics Association in which the rapidly evolving world of biomedical and clinical "big data" challenges are the focus. Papers deal with both translational bioinformatics, in which genomic and proteomic data dominate, and clinical research informatics, in which large clinical and public health datasets are prominent.

Patel, V. L., Arocha, J. F., & Kaufman, D. R. (1994). Diagnostic reasoning and medical expertise. *Psychology of Learning and Motivation, 31*, 187–252. This paper illustrates the role of theory-driven psychological research and cognitive evaluation as they relate to medical decision making and the interpretation of clinical data. See also Chap. 4.

Shah, N. H. (2012). Translational bioinformatics embraces big data. *Yearbook of Medical Informatics, 7*(1), 130–134. This article reviews the latest trends and major developments in translational bioinformatics, arguing that the field is ready to revolutionize human health and health care using large-scale measurements on individuals. It discusses data-centric approaches that compute on massive amounts of data ("Big Data") to discover patterns and to make clinically relevant predictions, arguing that research that bridges the latest multimodal measurement technologies with large amounts of electronic health care data is where new medical breakthroughs will occur. See also Chap. 25.

Questions for Discussion

1. You check your pulse and discover that your heart rate is 100 beats per minute. Is this rate normal or abnormal? What additional information would you use in making this judgment? How does the context in which data are collected influence the interpretation of those data?
2. Given the imprecision of many medical terms, why do you think that serious instances of miscommunication among health care professionals are not more common? Why is greater standardization of terminology necessary if computers rather than humans are to manipulate patient data?
3. Based on the discussion of coding schemes for representing clinical information, discuss three challenges you foresee in attempting to construct a standardized terminology to be used in hospitals, physicians' offices, and research institutions.
4. How would medical practice change if nonphysicians were to collect all medical data?
5. Consider what you know about the typical daily schedule of a busy clinician. What are the advantages of wireless devices, connected to the Internet, as tools for such clinicians? Can you think of disadvantages as well? Be sure to consider the safety and protection of information as well as workflow and clinical needs.
6. To decide whether a patient has a significant urinary tract infection, physicians commonly use a calculation of the number of bacterial organisms in a milliliter of the patient's urine. Physicians generally assume that a patient has a urinary tract infection if there are at least 10,000 bacteria per milliliter. Although laboratories can provide such quantification with reasonable accuracy, it is obviously unrealistic for the physician explicitly to count large numbers of bacteria by examining a milliliter of urine under the microscope. As a result, one article offers the following guideline to physicians: "When interpreting … microscopy of … stained centrifuged urine, a threshold of one organism per field yields a 95 % sensitivity and five organisms per field a 95 % specificity for bacteriuria [bacteria in the urine] at a level of at least 10,000 organisms per ml." (Senior Medical Review 1987, p. 4)
   (a) Describe an experiment that would have allowed the researchers to determine the sensitivity and specificity of the microscopy.
   (b) How would you expect specificity to change as the number of bacteria per microscopic field increases from one to five?
   (c) How would you expect sensitivity to change as the number of bacteria per microscopic field increases from one to five?
   (d) Why does it take more organisms per microscopic field to obtain a specificity of 95 % than it does to achieve a sensitivity of 95 %?