CHAPTER 13

USING INFORMATION SYSTEMS FOR PUBLIC HEALTH ADMINISTRATION

James Studnicki*
Donald J. Berndt
John W. Fisher

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
—T.S. Eliot, *Choruses from the Rock*

Chapter Overview

Public health organizations require well-designed information systems in order to make optimal use of the mounting supply of health-related data. Organizations rely on these systems to inform managerial decision making and improve operations in areas such as epidemiologic surveillance, health outcomes assessment, program and clinic administration, program evaluation and performance measurement, public health planning, and policy analysis. Key design considerations in developing information systems include service-based and population-based application objectives, units of analysis, data sources, data linkage methods, technology selection and integration strategies, and information privacy protections. A growing collection of models and resources now exists for developing effective information systems for public health organizations.

Information systems have emerged as an essential public health tool. Today, information systems provide real-time data to guide public health decisions. The rise in importance of health information systems (HISs) has three fundamental sources: (l) the expanding breadth of data available from multiple public and private sources, (2) advances in information technology (IT), and (3) the growing recognition of the power of information in public health decision making. Administrative data from public and private health service providers as well as insurers contain an electronic history of healthcare cost

*The authors wish to acknowledge the work of Stephen Parente, the author of the previous version of this chapter.

and use. Government surveys provide an unprecedented level of detailed information on health status, functional status, medical care use and expenditures, nutrition, sociodemographics, and health behaviors.

HISs support a wide variety of public health system objectives, including the following:

- Epidemiologic disease and risk factor surveillance
- Medical and public health outcomes assessment
- Facility and clinic administration (billing, inventory, clinical records, utilization review), cost-effectiveness, and productivity analysis
- Utilization analysis and demand estimation
- Program planning and evaluation
- Quality assurance and performance measurement
- Public health policy analysis
- Clinical research
- Health education and health information dissemination

IT has now advanced to the point that one year of the Medicare program's entire claims history—roughly 200 million observations—can be analyzed on a high-end personal computer (PC) workstation. Advances in IT are dramatically influencing public health organizations and their historical roles in collecting and disseminating data. Vital statistics and disease registries—critical functions of public health departments at both the local and national level—are being transformed by IT and its emphasis on evidence-based decision making. Yet, HIS resources remain difficult to develop and manage in addressing current public health challenges. Data sets are located in a balkanized array of separate computing platforms with little interconnectivity. For HISs to be effective, public health administrators must assess available data sources, design blueprints for extracting information and knowledge, and evaluate the benefits derived from these systems.

This chapter examines concepts, resources, and examples of HISs for public health organizations. Issues and implications for public health management are explored in the following five areas:

1. Contemporary concepts and applications of HISs in public health
2. Information systems architectures
3. Available databases
4. Operational models
5. Privacy and security

Contemporary Concepts and Applications

What is public health information? A more telling question may be what is *not* public health information, because the scope of data required to examine scientifically the multiple and overlapping health, social, and environmental factors that affect a population can be enormous. Traditionally, public health or epidemiologic data consist of vital statistics, disease registries, and other surveillance-based resources. However, these resources are often limited in scope because they only record natality, morbidity, mortality, and perhaps some measure of environmental and behavioral influences. Managing health

resources effectively at the population level requires a much broader scope of data resources to measure the effectiveness and cost of health interventions and policies.

An examination of public health applications of HISs is facilitated by an understanding of the two most common applications of these systems in practice. First, information systems are used to store and make available service data that reflect activities performed by public health organizations and other health-related entities. Second, information systems store and make available population-based data that are important for surveillance, program evaluation, policy making, and priority setting in public health. These two common applications are not separate but interact extensively.

For example, routinely collected service data by local public health agencies often include the results of blood lead screening of children under 5 years of age, immunization status, and encounter data recording the results of client visits for tuberculosis (TB) and sexually transmitted diseases (STDs). Other routinely collected service data include records of individual client encounters in the federal Special Supplemental Food Program for Women, Infants, and Children (WIC) and other early intervention programs. These service data are important for the effective management of individual care by public health and ambulatory care providers. Importantly, these data reflect individual transactions and can be used to monitor program performance and to describe a group of users at a particular facility—but they do not necessarily offer information about an entire community or population.

An important practical distinction exists between the service-based application of HISs and the population-based application, which offers information about defined communities and population groups of interest. To support this latter application, information systems must integrate data from major population-wide sources such as vital statistics registries and disease surveillance systems. In some cases, service data may also contribute to population-based information.

For example, the National Notifiable Diseases Surveillance System (NNDSS), formed more than a century ago, serves as a major source of population-wide data. This system captures information on disease incidence for approximately 50 diseases, which require accurate and timely information for effective prevention and control. The Centers for Disease Control and Prevention (CDC) receives reports of disease from the 50 states, two cities (New York City and the District of Columbia), and five territories.[1] This database is most useful to public health agencies because of its ability to analyze trends and conduct comparisons of disease incidence among communities.

Population-based information systems may also be constructed from service-level data. The immunization registries recently implemented by many state and local public health agencies provide an excellent example of this use. These registries record immunization status and vaccinations provided to all children residing within a defined geographic area so that this information is available not only to the initial provider, but also to other providers, health plans, and schools. Many of these registries incorporate birth certificate data for children born in the community, adding a population denominator. This is an example of an information system that provides service-level information that is helpful to individual providers and their patients, while

also providing population-level information that is helpful to public health organizations for surveillance, program evaluation, and policy making. A key qualification, however, is that a large proportion of the children in the community must be captured by such information systems in order for population-based information to have validity and reliability.

Drawing on the successes of immunization registries, a growing number of local public health organizations are developing computerized information systems for other purposes. For example, some local systems track the results of blood lead screenings performed at public health clinics, thereby producing important service information regarding the number of children screened, those with elevated blood lead levels, and those receiving follow-up treatment and lead abatement services. This information is based on service data, but if the systems can capture data on all children in a defined community, then valuable population-based information can become available.

The relatively recent availability of state-of-the art computing technology has enabled public health organizations to collect health data rapidly and extract meaningful information about community health status.[2] The major challenge is to integrate data sources and develop networks that make this information optimally available to public health organizations at all levels of government as well as to appropriate entities in the private sector. New service-oriented computing architectures are intended to build these types of networked information systems. The current impetus to have a surveillance capability supported by a national network of public health HISs is fueled by concerns about bioterrorism and emerging infectious diseases, resulting in sizeable investments by the CDC for constructing linked information systems. (See Chapter 23 for more information on the use of public health information in managing disasters.)

Major practical goals for the future development of HISs for public health organizations include the following:

- Integrate the multiple data sources available for public health purposes.
- Network information systems to make interaction and information flow between different entities feasible.
- Use health care delivery information systems to produce public health information regarding preventive services, preventable diseases, and quality of care.

Integration

Government public health agencies have historically designed computer-based information systems for single programs. For years, the same data were entered and maintained in many different, often incompatible, systems that supported different public health programs.[3] This duplicative and fragmented information infrastructure hindered the ability of public health managers to know what data existed and how to access them. For example, most local public health agencies maintain multiple programs for children, including lead toxicity prevention, immunization, WIC, and early intervention services. Meanwhile, the local departments of social services enroll families in Medicaid. Despite the fact that Medicaid and public health programs serve client populations that overlap substantially in most communities, the data-

bases used to manage these programs are entirely separate in most cases, reflecting the categorical mechanisms that support these programs. Information systems integration can offer opportunities for improved service delivery and enhanced population-based decision making and management.

Linkage of data sets is often an effective method to obtain information across programs. For example, linkage of WIC records with Medicaid, birth and death, and hospital discharge files has enabled program analysts to document the effectiveness of the WIC program in reducing infant death and costly neonatal hospitalizations. Similarly, linking lead screening registries, Medicaid eligibility files, and managed care plan enrollment files can enable public health organizations to monitor compliance with lead screening by health plans. These linkages for special-purpose studies are often highly customized and assembled only for the duration of particular studies. The ongoing surveillance and community assessment activities that represent core public health functions require HISs to accumulate and integrate data for continuous use. (For more detailed discussion on community assessment, refer to Chapter 15.) This is a data warehousing problem. Data warehousing technologies are widely available and should become a key technology in the public health arena. Data warehouses organize data as cubes that can be "sliced and diced," providing a flexible environment to pursue analyses. Vital statistics, hospital discharge data, and disease registries can be integrated with demographic and economic data to populate public health data warehouses. For example, the Comprehensive Assessment for Tracking Community Health (CATCH) data warehouse integrates Florida data for use by health planners.[4] The warehouse has been used to generate more than three dozen assessment reports, along with many more targeted research projects.

Public health agencies are also beginning to innovate by using unconventional data sources such as market research databases. For example, electronic information compiled from grocery and drug store sales can be used as part of an HIS to identify the purchase of cigarettes concomitantly with products associated with pregnancy or infants, such as diapers. This information by ZIP code can help target or evaluate public health intervention programs, such as efforts to prevent tobacco use in the perinatal period.

Networks

Another major function of HISs in public health is to create linked networks of information that can strengthen public health operations by: (1) facilitating communication among public health practitioners throughout the United States, (2) enhancing the accessibility of information, and (3) allowing swift and secure exchange of public health data.[5] As a prominent example, the CDC initiated the Information Network for Public Health Officials (INPHO) in 1992. The CDC has been the major supporter of efforts to create networks that link public health information from localities and states with that of federal agencies. Information networks of this type are increasingly indispensable for disease surveillance activities, particularly in cases of local disease outbreaks that have the potential to spread regionally and nationally. In this way, HISs can help to create and sustain effective interorganizational relationships among public health organizations.

Utilization of Health Care Delivery Systems

Public health organizations can also benefit from timely access to health care services information from providers of personal healthcare services.[5] For example, immunization registries must acquire information on immunization status from multiple community providers who deliver vaccinations. In a growing number of communities, public health organizations are able to obtain relevant and timely information from the systems that are maintained by health care delivery organizations. Large delivery systems can offer information on the delivery and utilization of preventive services (including missed opportunities for prevention), the incidence of preventable diagnoses and comorbidities, and the quality of healthcare facilities and providers (such as rates of medical errors, mortality, and hospital infections).

These types of resources drive the contemporary development of HISs among public health organizations, and they reflect a basic change of thought regarding the delivery of medical and public health services subsequent to the 1993 federal health reform initiative. This initiative accentuated the need for informed decision making by consumers, providers, employers, and governments. For example, the Clinton administration reform plan relied solely on analyses of the 1987 National Medical Expenditure Survey (NMES) to draw conclusions about the future demand for and cost of health care in the United States. Between the time the NMES was fielded and 1993, the dominance of fee-for-service gave way to managed care as the primary health financing mechanism for the private and public insurance market. As a result, the 1987 NMES could not reliably estimate the impact of the administration's healthcare reform proposal without significant and possibly questionable assumptions. The limitations of the data increased the administration's interest in an annual survey that could provide better estimates of a rapidly changing market. In 1996, the Agency for Health Care Policy and Research fielded the Medical Expenditure Panel Survey (MEPS), providing a national annual survey instrument to track changes in health care use and cost as well as health status and insurance coverage. A similar demand for information came from employers, who wanted health plans to provide standardized information on the value of their products. The result was a cooperative effort between employers and health insurers to develop a common set of health plan performance measures known formally as the Health Plan Employers Information Data Set (HEDIS), developed by the National Committee for Quality Assurance. Some of the HEDIS measures were prevention oriented (e.g., immunization) and thus illustrated the principle of obtaining public health information from a health care delivery information system.[5] (See Chapter 11 for more information on data and Chapter 18 for more information on the evaluation of public health information.)

Building new databases for multiple purposes such as MEPS and HEDIS required a clear identification of HIS objectives as well as knowledge of the strengths and weaknesses of established data structures. This knowledge is essential in determining which structures can be recycled in building a new database, such as using existing health insurer records for HEDIS, and which structures need to be newly constructed, such as designing medical record abstraction protocols for obtaining disease and outcomes data for HEDIS. With appropriate design, medical encounter data (service data) can be used for sev-

eral population-based purposes, including community health assessment, surveillance, and evaluation.

Information Systems Architectures

A common misperception in developing HISs for public health applications is the expectation that such systems are analogous to their counterparts in IT-intensive industries such as banking or manufacturing. Health is a combination of many uncertain inputs. These inputs range from the unique biologic and behavioral characteristics of the individual patient or population under study, to health insurance characteristics and the accessibility of health resources, to the practice styles of physicians and other health professionals, as well as to thousands of possible diagnoses, comorbidities, risk factors, and interventions. In combination, these inputs generate millions of possible outcomes for a given health episode. Consequently, the HISs used to support public health applications and decision making may need to be more complex and costly than the systems supporting applications in other industries and professions.

In building an HIS, the field of health informatics constitutes a multidisciplinary core of expertise, including specialists from the following fields:

- Computer science
- Electrical engineering
- Medicine, nursing, and allied health management
- Finance and accounting operations research
- Economics
- Sociology
- Survey design
- Epidemiology
- Statistics

These disciplines work in combination to produce HISs to serve the public health system objectives described above. In designing and managing HISs, public health administrators require the ability to: (1) distinguish between data, information, and knowledge; (2) define units of analysis for the level of data aggregation; and (3) understand the health IT architecture of system(s) to be used.

Data vs. Information vs. Knowledge vs. Wisdom

There have been endless discussions on the differences between data, information, and knowledge. Although the boundaries seem somewhat blurred, the distinction can be helpful at a more abstract level. The current conception of the data, information, knowledge, wisdom (DIKW) hierarchy can be traced in part to T.S. Eliot's poetry that started the chapter. *Data* are raw facts and statistics that are collected as part of the normal functioning of a business, clinical encounter, or research experiment. *Information* is data that has been processed in a structured, intelligent way to obtain results that are directly useful to managers and analysts. This is often the case once data has been organized in a database management system. *Knowledge* is obtained by

using information to explain the context of a problem or situation. Finally, *wisdom* is knowledge tempered by experience.

In public health, data are obtained from a variety of sources ranging from patient history at a clinical visit to health insurance claims to bacteriology laboratory reports. To be valuable for generating information and knowledge, data must be readily accessible and reliable. Generally, electronic data in standardized formats are most efficient. However, data that are easy to obtain may not be the most accurate or precise. For example, electronic health insurance claims data can identify a specific immunization on a particular date but do not indicate the child's overall immunization status. For that information, medical records or reports from a computerized immunization registry are needed. Thus, the cost of obtaining accurate and precise knowledge concerning a child's immunization status may be outside the scope of existing data collection processes. Ethical questions also arise if immunization data had to be transferred from another source and parental consent had not been given. (See Chapter 5 for a more detailed discussion on ethics.)

Service-Oriented Computing

Healthcare planners and administrators are likely to interact with large-scale information systems both as end users and participants on implementation teams charged with the responsibility of deploying new technologies. Even direct providers of care can use an increasingly integrated set of information systems to capture patient-level data in electronic medical records, as well as more general knowledge for clinical decision support systems. This section explores the architectural considerations of large-scale information systems as computing power continues to dramatically improve with each new release and the growth of networking provides increasingly reliable interconnections.

Among the most interesting and promising trends in information systems architecture is the growing body of technologies and standards for service-oriented architectures. In the past, software engineering involved building monolithic systems from customized single-use components. Although the components might be well built and offer sophisticated capabilities, the number of dependencies between components is a source of failure in such highly customized complex systems. Newer programming language extensions and object-oriented approaches emphasize the encapsulation of component details and more explicit programming interfaces to better manage growing software complexity. These evolving software development tools support servicelike approaches based on reusable components. More current service-oriented architectures continue this trend, providing standards for defining and using Web-based services.

Service-oriented computing (SOC) is not new, but maturing standards support the approach and make implementing complex service-based systems practical. As many traditional software engineers observed, it is possible to adhere to the principles of encapsulation and other good programming practices in any language, but explicit support makes the task much easier. Service-oriented computing standards such as the Extensible Markup Language (XML), Simple Object Access Protocol (SOAP), and Web Service Description Language (WDSL), govern the structure, transmission, and description of services. Registry standards such as the Universal Description, Discovery, and Integration

(UDDI) protocol provide a method for publishing and finding services as components of complex systems. These standards allow developers to implement high-quality, tightly focused computing services that can be published and serve as components in large information systems.[6]

Among the most important goals of well-designed information system architectures are scalability to meet growing demands, flexibility to meet changing demands, and reliability or fault tolerance so systems are continually available to meet all demands. Being an agile organization places a premium on the flexibility to change business processes and supporting information systems by adding or modifying components. This is particularly challenging given the number of independent or quasi-independent providers that coordinate to deliver health services. A key advantage of service-oriented computing is the loose coupling between components. No detailed knowledge of the internal operation of a service is required and all coordination is managed through standardized protocols. Using this approach, services can be reused, rearranged, and new services can be added, as systems are adapted in the pursuit of new opportunities.

Computer Networking

Computer networking and communication technologies provide the glue that binds the different components or services that form complex, distributed information systems. Typically, networking tasks are separated by the general nature of the connection and the underlying technologies used to handle the communications (see Figure 13-1). The major technology classes are wide area networking (WAN), local area networking (LAN), and storage area networking (SAN), although this latter category receives much less attention in the popular press.

Wide area networking is the term applied to the task of interconnecting large numbers of geographically dispersed computers. This is typically accomplished in piecemeal fashion by interconnecting smaller networks to create more global connectivity—the Internet being the quintessential example. Internetworking relies on standard protocols or rules of engagement that allow data to be routed through cooperating networks. Although there have been many competing proposals, the current standard that governs the Internet is the Transmission Control Protocol/Internet Protocol (TCP/IP). The Internet Protocol provides for routing or message delivery through cooperating networks, with the most basic service being best-effort delivery of a message (with no guarantees). The Transmission Control Protocol provides a reliable end-to-end delivery service that costs a bit more (computationally speaking). So the services are much like the dilemma that faces any physical mail user, cheap delivery with no guarantees and such premium services as return receipt and package tracking. These wide area networking protocols provide the foundation for the emerging service-oriented computing approaches discussed above.

A local area network (LAN) spans an office, the floor of building, or similarly restricted geographic area. There can be many computers in this small area that are typically interconnected by shared media. That is, devices compete for access to a wire or other means of communication, but the sporadic nature of traffic means overall performance is reasonable. Not every computer needs access to the network at the same time. Ethernet is the dominant

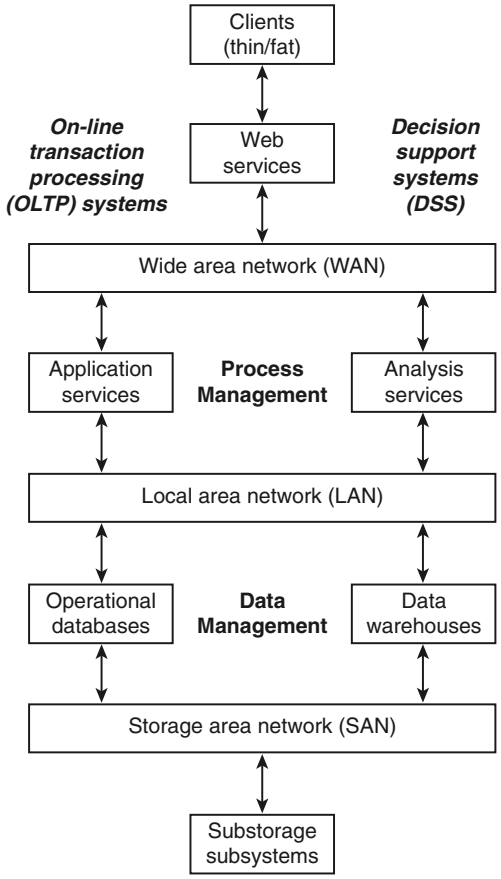

FIGURE 13-1 Information Systems Architecture

technology in this arena because of its very low costs, wide availability, and continually improving speeds. In a clinical environment, local area networks are likely to interconnect departmental devices, including handheld wireless devices that are becoming increasingly important.

Storage area networking (SAN) technologies provide an infrastructure for the lower levels of tiered architectures. Database systems, file transfers, document handling, and image storage and retrieval all require the transmission of large amounts of data. Storage area networks provide a high-performance alternative for such demanding tasks. SAN technologies allow storage to be centralized and flexibly reallocated through network addressing.

Tiered Architectures

Figure 13-1 presents an architectural framework that illustrates some of the major logical functions that are embodied in complex information systems, from the interfaces with the ultimate end users to the storage subsystems that protect very large collections of data. The arrangement of subsystems in separate tiers or layers provides more flexibility as whole systems need to be reconfigured to meet new demands. The ability to meet larger demands is also

enhanced by allowing bottlenecks to be isolated and removed by improving computer performance or having several computers focus on a single task. Finally, tiered architectures separate tasks and thereby reduce the dependencies that often cause system failures. These tiers are viewed as logical functions that might physically reside on a single server in the case of a small system or on many machines in complex distributed systems.

Storage Management

Anchoring tiered architectures are storage subsystems, the actual data repositories. This is shown as a separate tier as there have been many advances that allow storage to be more centrally managed and shared across different information systems. The total cost of ownership for storage is dominated by management costs, with a current rule of thumb estimating that for every dollar spent buying storage, 10 dollars is spent managing that storage. Therefore, effectively allocating and managing storage has become an important architectural goal. The high-performance storage area networking infrastructure just discussed has given system designers both the freedom to centralize storage for management and the flexibility to deploy and dynamically reallocate storage as demands change. Storage subsystems can be as simple as redundant arrays of independent (or inexpensive) disks (RAID) that combine two or more disks for fault tolerance and performance gains. In most configurations, redundant data spread across the disk drives allows the storage system to recover from disk failures, as a hot-swappable spare can be rebuilt from surviving drives. This technology has been widely adopted and is now available on even very low-cost servers. The use of networking also allows the storage to be physically separated from servers running databases or other applications. Network attached storage (NAS) uses existing network capabilities to make storage accessible to many computers as a network resource. For demanding applications, dedicated storage area networks can be used to attach advanced subsystems that offer large storage sizes, dynamic performance tuning, and integrated backup capabilities. These enterprise-class storage systems seem well suited to healthcare applications that must be highly reliable and meet stringent security demands. These devices also support backup and disaster recovery strategies that are critical in the healthcare industry.

Data Management

The database management systems market has seen considerable consolidation with a few major vendors holding substantial market shares, along with some interesting open source alternatives. Most products are based on relational database technologies, with extensions that make most systems object-relational database management systems. Relational database systems allow users to query large collections of data without knowledge of the detailed storage structure. A high-level query language, such as the structured query language (SQL), is used to express the desired result. The detailed execution plan for actually retrieving the data is automatically constructed by a query optimizer in the database engine. Users of relational database systems need not be overly concerned with the physical storage characteristics and are largely isolated from changes as specific databases evolve.

An important distinction is made between operational systems that support the day-to-day operations of an organization and decision support systems that provide analytic services. Data warehouses are very large collections of data accumulated over time that allow decision makers to better understand trends and conduct "what-if" analyses. The typical operations in a data warehouse are retrievals or "reads" that summarize large subsets of data. The predictable nature of these retrievals has led to the development of many online analytic processing (OLAP) tools, which allow users to easily navigate and visualize the data, creating customized reports for specific decisions. Data warehouses are populated through fairly complex extraction, transformation, and loading (ETL) processes that collect data from the operational systems and ensure data quality.

Process Management

The next tier focuses on business processes or workflow management, often on dedicated application servers. There are many tools available for designing and implementing workflows that allow systems to handle many contingencies. A workflow is typically a mixture of human and computer activities coordinated through a process model. There are many examples of workflows or processes in the healthcare industry. For instance, the admission process to a hospital can be formally modeled and embedded in an information system. The individual tasks that makeup a workflow may or may not be dependent on previous tasks, so some can be pursued in parallel, while others must await the completion of prior tasks. Each step may also produce data that is passed to the data management tier. In the case of hospital admissions, it is clearly an operational system that captures newly created data as a patient arrives at the hospital, provides insurance coverage details, and is examined by staff. As an example of service-oriented computing, imagine a Web service provided by health plans that allows the hospital to obtain up-to-the-minute details of insurance coverage without bothering the patient with paperwork. Other excellent examples of healthcare workflows are care guidelines that can be embedded in information systems. As patients undergo specific regimens, departures from accepted care guidelines could be more carefully monitored.

The Client Side

The upper tier of an information system is concerned with the presentation of any results to the ultimate end users of an information system and should thus reflect the needs of the end user. The software complexity and computing demands reside on the end-user's or client's computer. At one extreme is a computer that uses nothing but a Web browser to interact with a large information system through a Web portal (a very thin client). Though the services might be quite sophisticated and require substantial computing resources, the burden is on the collection of servers that are used to build the core information system, not on the end user's computer. An alternative, somewhat "heavier" client is a computer with a traditional statistical package installed locally that might be used to analyze a large data set from a network accessible database. The data are shipped to the end user's computer where all the analyses are conducted.

Of course, the end users (or their support staff) are responsible for correctly configuring the statistical package, upgrading to newer versions, and ensuring the computer is powerful enough to handle the demanding tasks. Therefore, many corporate computing policies attempt to control and minimize the burden of complex software installed locally on client computers.

In many cases, interaction with the client uses the ubiquitous Internet protocols to deliver content through a Web browser or portal environment. If the system is designed as a true Web service, the protocols discussed above can be used to provide a public interface. Therefore, the presentation tier typically includes a Web server listening for connections from client computers over a wide area network.

Putting the Pieces Together

As examples of information system architectures consider the Florida healthcare data warehouse cited earlier along with the many operational systems that serve as primary data collection points. In Florida, the Agency for Healthcare Administration (AHCA) requires that all acute care hospitals report standardized data after patients are discharged. This data is originally collected using hospital information systems, complex commercial systems that are typically tiered architectures with storage subsystems that provide a high degree of reliability for electronic medical record applications. Data is extracted from hospital systems on a quarterly basis and reported to the state, where various data quality procedures are used to verify the data. The hospital discharge data and many other data sets, including vital statistics and specific disease registries, are loaded into the data warehouse and integrated for decision support activities. The data warehouse itself is a tiered system, accessible to health planners through a Web portal. A middle tier provides analysis services based on data cubes that can be filtered and aggregated as needed, allowing analysts to select an appropriate unit of analysis. The data cubes are constructed from a base tier that includes a large relational data warehouse, where all the data quality and integration procedures are implemented. These types of systems will become standard public health tools, integrating conventional and unconventional data for evidence-based public health.

Sources of Data for Information Systems

The heart of any public health information system is the data that it contains. Understanding the fundamental characteristics of databases is essential for effectively structuring and employing database technologies. Public health information databases share many characteristics with business enterprise data warehouses. That is, the systems are generally "subject oriented, integrated, nonvolatile, time variant collection[s] of data in support of management's decisions."[7]

Dimensional model data warehouses are constructed of two main components—*fact* tables and *dimension* tables. Facts are most often numeric, continuously valued, and additive measures of interest. For instance, a hospital discharge record includes such fields as length of stay and charge data that can be summed or averaged across various population groupings. Most

commonly, however, the discharge counts are aggregated to compute event rates for specified demographic population segments.

Dimensions, on the other hand, are textual and discrete, providing a rich query environment for investigating associations between the dimensions and outcomes. Common hospital discharge dimensions include race, age, gender, physician, diagnosis, procedure code, payer, and so on. To facilitate analysis, many of these dimensions are refined to create *hierarchies* for aggregation. For instance, instead of comparing the hospitalizations for every different age by individual years, records are "rolled up" or aggregated by age bands, which are simply predefined age groupings. Similarly, a geographic hierarchy (ZIP code, community, county, state, region, or nation) identifies multiple levels of aggregation for comparing hospitalization rates.

Data Characteristics

Grain

Data granularity refers to the level of aggregation, or distance from individual events, of the fact tables. The finest grain data is the individual transactions themselves, such as birth or death certificates or individual hospital discharge records. As data is aggregated or summarized, there is a commensurate loss of information. For instance, death records are commonly rolled up geographically (to the county or state level), temporally (quarterly or annual rates), by gender (separate rates for males and females), race (rates for black, white, Asian, etc.), and causes (combining various ICD-9 or -10 codes). Virtually all reports are aggregated data. Although these aggregates are useful for comparing rates, they can not be later disaggregated without access to the underlying transaction records.

In general, the finer the grain of the fact tables, the greater the flexibility in aggregating and analyzing it. On the other hand, finer granularity also increases the number of records that must be maintained and accessed, and increases the complexity of queries that must be formed to create reports. Moreover, as the grain of the data gets smaller, so do the cell sizes, particularly for uncommon events. As cell size declines below approximately 30 events, statistical significance becomes a serious concern. When dealing with multiple data sets, it is important to ensure that the granularity definitions match and have not changed over time. For instance, annual data may reflect an average value over a calendar year, a single beginning, mid-, or end-year value, or even a fiscal year average.

Determining the granularity in a data warehouse is one of the critical design decisions, providing a lower bound on subsequent analyses. Therefore, designers typically err on the side of more rather than less detail. Data warehouses provide a flexible query environment by allowing users to "roll up" or summarize data, enabling a decision maker to choose a *unit of analysis*. The unit of analysis determines a level at which data is aggregated and analyzed in order to generate information. The four most common units of analysis in public health are the following:

1. Person/patient
2. Vendor/supplier
3. Program
4. Region/population

For example, an average diabetic patient is associated with 200 billing records in a year. The information can be bundled at the person/patient level by building a set of counts for the frequency of a certain service that is vital to recommended diabetic care, such as the number of hemoglobin A1c or diabetic retinopathy screening tests. Once completed, a database with one observation per diabetic patient is created to compare patient-level variation in quality of care. This type of bundled procedure can be completed for a variety of different levels of analysis.

Generally, the unit of analysis should correlate to the population of interest for a management decision. For example, if the principal focus of an evaluation is patient compliance with a disease management program, then the correct unit of analysis is the patient. If, however, the goal of the analysis is to compare how different disease management programs perform to improve the health of the population, the appropriate unit of analysis is the program. Vendors or suppliers in health care can be physicians, hospitals, group practices, public health clinics, health insurers, or any other health-related organization. Regions can be defined in a variety of ways, including state, interstate regions (e.g., the Midwest), metropolitan statistical area, county, and ZIP code. Populations can be defined by residence within a given geopolitical subdivision, by sociodemographic characteristics such as age and ethnicity, or by health conditions such as diagnoses or behavioral risk factors. Other health-related units of analysis are defined by diseases, medical procedures, or other health interventions.

Scope

Scope is a measure of breadth of coverage across any of the dimensions. For instance, geographic grain describes the unit of coverage, such as, census tract level data, and geographic scope is the coverage of all tracts in a county or state. Temporal grain reflects the finest unit of time, and temporal scope reflects the total units available (e.g., monthly or annual data covering the last 5 years).

Source Type

Public health data can be gathered from a variety of source types, with a correspondingly broad spectrum of reliability. Vital statistics and hospital discharges (and their derived aggregations) are generally among the most accurate, as their sources are individual events recorded by objective individuals and subject to postcollection cleansing efforts by official agencies. Moreover, the formatting and coding systems are often standardized across the states to facilitate ease of reporting to the CDC, which compiles state data into national reports. Similarly, state registries most commonly collect information through questionnaires completed in real-time by third parties at the point of service and are generally validated for accuracy and completeness before entering the database.

In contrast, data sets such as the Behavioral Risk Factor Surveillance System (BRFSS) are collected using survey instruments that are completed by individuals with varying levels of commitment to completeness and accuracy. Moreover, surveys are, by their very nature, partial samplings of the total populace and results must be projected to include the whole population. This necessarily introduces sampling error and the potential for selection bias (respondents may selectively opt out of embarrassing questions).

Finally, some data, such as demographic changes between census years or estimates of per capita income, are simply estimated based on observations of proxy events, such as school registrations, vehicle registration changes, and voter rolls. Different agencies use different methods for arriving at their estimates with necessarily different results. It is important that estimate sources not be mixed in the database.

Regardless of the source of the data, however, it should be checked for completeness and consistency before being added to the data repository. A clear policy for handling missing or inconsistent data elements must be thought out and enforced, unit definitions understood and reconciled, and records containing clearly erroneous data flagged or removed.

Metadata

Central to the maintainability of the data repository is keeping a clear record of its contents. This record is called *metadata*, or data about the data. Elements such as sources of the data and the agencies responsible for its collection, definitions of each of the fields, the date the data set was last updated, and the number of records each update contains can be invaluable for performing data quality checks, as well as providing needed context for end users. Making such metadata available to end users should be an integral part of any information system, but it is particularly important for health information systems, where timeliness and context are critical for proper interpretation.

Integration

Although individual databases can be used to investigate simple count or rate questions, tapping the real power of information involves integrating the information from multiple datasets. This is accomplished by linking the tables through *key fields*. In database design, a *primary key* (PK) is a value that can be used to identify a particular row in a table. A *foreign key* (FK) is a field or group of fields in a database record that point to a key field or group of fields forming a key of another database record in some (usually different) table. Usually a foreign key in one table refers to the primary key of another table. For instance, a vital statistics data set can be queried to determine the number of deaths due to cervical cancer in a given county over a given period of time. To calculate a rate, however, requires linking this count with the demographic table for the same county and same period. Fields used to link the numerator (death count) to the denominator (population) necessarily include the county, the time period, and the gender (since only females are susceptible to cervical cancer).

Common Data Problems

Not all data is created equal, and the prudent investigator choose's sources carefully. Fuzzy data element definitions, inconsistent collection and screening processes, changing variable definitions or scales, and intentional hiding of data values are simply a few of the threats to data quality that must be considered and addressed before bringing new data sources into the data warehouse.

Race is a particularly problematic dimension because it is generally self-reported and poorly understood. Many individuals responding to surveys or

questionnaires confuse race with ethnicity or nationality and may classify themselves as multiracial or other if their nationality is not listed as a racial option. This confusion was magnified by the significant expansion of racial and ethnic categories offered in the 2000 census. From the single selection from the relatively simple four racial categories offered in the 1990 census, respondents in 2000 could select from an expanded list of over 30 options. Moreover, since this same smorgasbord of racial options is generally not duplicated in most event data collection instruments (e.g., hospital discharge records or vital statistics forms), reconciling event data with demographics requires careful conformation of racial definitions. That is, the demographic value categories must be conformed to the definitions used in the event records. The CDC has created a bridging methodology for reconciling the different race categorization schemes (see the full description at http://wonder.cdc.gov/wonder/help/bridged-race.html); however, this issue promises to grow significantly with time as more racial, national, and ethnic groups assert their distinctiveness.

A related, more general, data concern is the tendency of data collection agencies to change the definition of data elements or the circumstances of collection. For example, ZIP code boundaries change frequently, with 5–10% of the codes changing each year. Besides the obvious problem of aligning the numerator (event) values with the denominator (population estimates) for rate calculations, the creation and deletion of ZIP codes each year presents challenges when trying to trend data over time.

More serious are changes to definitions of the data itself. ICD-9/10 changes most often involve additions or deletions of codes, rather than changes in definitions themselves; however, aggregations based on these codes often do change. Communicable disease reports, for instance, may simply report hepatitis C incidence one year, and then split the data out to report acute, chronic, and congenital incidences the next year. Unless the change is detected and the new subcategories aggregated, the data warehouse values will be in error.

A more subtle problem with public-use data sets may arise from privacy concerns. Many government agencies responsible for collecting and distributing event level data will mask one or more of the fields that may be used to identify individuals. Masking simply replaces the actual value of the masked field(s) with one or more placeholder values for some predetermined percentage of the records. The most commonly masked fields are ZIP codes, age, gender, and race. For instance, in the case of California hospital discharge records, the public use data set masks the gender of approximately 18% of the records, 26% of the race values, 30% of the ethnicity values, and over 46% of the ages. Although the masking process should not affect comparison of rates between geographical entities within the state (since all entity rates should be reduced similarly), any rates calculated for comparison with other states or national statistics must account for the artificial diminution of the numerator values. Unmasked data may be available, but is generally provided only for specific, defined research projects and requires formal oversight by an approved institutional review board (IRB).

Common Databases Available for Public Health

A solid understanding of health IT and data structure is required for the optimal design of public HISs. Fortunately for public health managers, there are

rich data resources available at federal and state levels. This section provides an overview of the most common databases available to health managers and researchers in developing HISs (more details are available in Chapter 11). Most of the databases described here are federal or state specific in their focus. Although a federal focus may be too broad for local and regional health policy issues, federal surveys can still provide two significant benefits. First, national databases provide field-tested survey instruments or data abstraction tools that can be applied to a more focused information system. Second, federal surveys can provide a comparison database for information systems that also use state and local data sources in order to gauge the effectiveness of local initiatives.

It is worth noting the distinction between health statistics databases and health reports. Web sites that serve as data sources, such as the US Census (http://www.census.gov) and some state departments of health allow users to execute relatively broad queries that return fine grain data across the full scope of one or more dimensions. Report sites, on the other hand, provide either preformatted reports, often in fixed formats such as Adobe PDF files, or point queries that return aggregated data for a limited scope. The census site, for instance, allows end users to generate very comprehensive queries covering a large number of available indicator statistics, grouped by the full spectrum of gender, race, age, and geographical groupings. The data is downloadable as spreadsheet or comma separated value (CSV) files that can be directly imported into database programs for end-user manipulation.

At the other end of the spectrum is the Florida Department of Children and Families Youth Substance Abuse Survey reports (http://www.dcf.state.fl.us/mentalhealth/publications/fysas/countyreports04.shtml) that present aggregated county data in individual PDF files, without benefit of race, gender, or grade-level breakdowns. This requires manual conversion of the data tables into spreadsheet format and even then precludes any end user reaggregation of the data or creating different dimensional views.

Most sites fall between these two extremes. For example, the Florida Community Health Assessment Resource Tool Set (CHARTS) at http://www.floridacharts.com/charts/chart.aspx allows users to return statewide incidence counts and rates for hundreds of diseases and injuries, grouped by county, ZIP code, gender, race, or age bands and formatted in spreadsheets for easy importation into a database. A comprehensive view of the health status of communities across the state can be generated very quickly using this system, although the data is grouped by only one dimension at a time, preventing end-user crossing of the demographic variables.

Government Survey Data

The federal government collects a broad array of data that may be used by public HISs. The US Department of Health and Human Services (HHS) has the largest health data collection responsibility. However, other federal government departments such as defense, labor, and commerce also collect critical health data.

The phrase “national probability sample” describes a survey instrument that has been deliberately designed to reflect the US national population’s sociodemographic variation in age, gender, race, income, and education. If a state-level analysis was attempted, the survey could produce misleading esti-

mates if survey respondents were over- or underweighted to reflect their proportional representation within the nation.

Another important concept is the panel survey. In this design, a panel or cohort of survey participants is followed during several rounds of the questionnaire. For example, some surveys such as the MEPS and the Medicare Current Beneficiary Survey (MCBS) follow participants for at least 2 years to track health status and cost. Panel surveys are valuable to assess long-term impacts in health care, such as a lack of health insurance or follow-up from a massive heart attack.

Most federal surveys are collected on an annual basis and are generally available as public use files 1 to 2 years following the completion of the data collection period. These data are available for a small fee to cover the cost of producing the databases. A list of nearly all of the government surveys used for health is available on the Internet at http://www.cdc.gov/nchs/.

Several examples of government-sponsored survey data are provided in the following paragraphs.

Current Population Survey (CPS)

This survey is completed monthly by the Census Bureau for the Department of Labor and updated annually. It contains basic information in healthcare use and can be queried online at http://www.census.gov/hhes/www/cpstc/cps_table_creator.html. It is often available before any other federal survey with health data. The sample consists of approximately 52,000 housing units and the persons in them. The survey's primary goals are to provide estimates of employment, unemployment, and other socioeconomic characteristics of the general labor force, of the population as a whole, and of various subgroups of the population.

National Health Interview Survey (NHIS)

This survey, collected by the National Center for Health Statistics (NCHS) within the HHS, is a national probability sample of the health status of the population. A two-part questionnaire is used with a sample size of approximately 49,000 households yielding 127,000 persons. The NHIS has had continuous data collection since 1957 for national estimates through household interviews by US Census Bureau interviewers. The NHIS provides the sampling frame for other NCHS surveys and is linked to the National Death Index (discussed later). Both a core survey of demographic and general health information and a supplement focusing on different populations are deployed.

National Health and Nutrition Examination Survey (NHANES)

The NHANES is sponsored jointly by the CDC and the NCHS as part of the HHS. The primary goal of the NHANES is to estimate the national prevalence of selected diseases and risk factors. Target diseases and areas of special interest include (but are not limited to) cardiovascular disease, chronic obstructive pulmonary disease, diabetes, kidney disease, gallbladder disease, osteoporosis, arthritis, infectious diseases, substance abuse, tobacco use, child health, mental health, environmental health, and occupation health. Public use files from the NHANES are currently available.

NCHS Medical Care Use Surveys

The NCHS has several annual surveys of healthcare services designed to profile the use of services regardless of public or private payer. The surveys are specific to inpatient, ambulatory care, home care, and other types of services. These are excellent surveys for national comparisons of changes in the use of ambulatory and inpatient care. However, they are not able to generalize to any area smaller than a multistate sample (e.g., the northeast United States).

Administrative/Claims Data

The use of administrative data in public HISs has dramatically increased as the cost to work with the data has been reduced and the quality of data, relative to its past quality, has improved significantly. Administrative data are defined as the data elements that are generated as part of a healthcare organization's operations. For example, health insurers generate claims data to record the services that are reimbursed by the insurer. There are three significant advantages to using administrative data. First, the data cover a large breadth of services ranging from inpatient services to prescription drug use and immunizations. Second, administrative data are an inexpensive source of data when contrasted to other forms of health service data such as medical records. The third advantage is the timeliness of availability when compared with government surveys and other data sources. The most commonly used administrative databases are described to illustrate the range of data available for use in information systems.

Medicare National Claims History File (NCHF)

The NCHF is more of a database architecture than a single file. Generally, it includes two file types. One file type is an annual 5% sample of the roughly 40 million Medicare beneficiary population. This file is sold as a public use file by the Centers for Medicare and Medicaid Services.

The second file type includes specialized data extracts of the NCHF across the Medicare population. An example of this type of file is any patient who received either a coronary artery bypass graft procedure (CABG) or angioplasty in 1990 and their claims for the next 5 years. Within these data, one can track health outcomes, such as repeat hospitalizations for cardiac conditions as well as mortality. Reimbursed services included in the claims file are inpatient, outpatient, hospice, medical equipment, provider services (e.g., physician), home health care, and skilled nursing care. The key identifying variables for the NCHF data extracts are inpatient diagnosis-related groups (DRGs), physician procedure codes, and diagnosis codes. Unlike survey data, the NCHF can be used to develop state-, county-, and possibly even ZIP code-level analysis, depending on the prevalence of the condition or treatment under investigation. See http://www.cms.hhs.gov/data/default.asp for a listing of Medicare data available.

State Hospital Discharge Records

Over half of all states maintain annual hospital discharge summary records. These data are valuable for examining changes in inpatient service use and

cost. For example, changes in the use of CABGs and angioplasty over several years can be assessed by different age, gender, and health insurance payer categories. The principal advantage of using hospital discharge records for health policy purposes is that they contain data on all payers, whereas the NCHF only provides data on Medicare. Hospital discharge data can be obtained directly from a state's government or from the AHRQ Healthcare Cost and Utilization Profile (HCUP) standardized databases. HCUP databases include the following:

- The Nationwide Inpatient Sample (NIS) contain inpatient data from a national sample of over 1000 hospitals.
- The State Inpatient Databases (SID) contain the universe of inpatient discharge abstracts, including over 100 structured clinical and nonclinical data elements, from 36 participating states. The advantage of the SID is that it allows an analyst to obtain data from 16 of the 36 states in a standardized format from AHRQ; the other 20 states provide the data directly in roughly similar formats.
- The Kids' Inpatient Database (KID) is a nationwide sample of pediatric inpatient discharges, drawn from the SID database and is the only all-payer inpatient care database for children in the United States.

For more information on the HCUP databases, refer to http://www.ahrq.gov/data/hcup/.

State Medicaid Claims Data

Most states maintain claims data for reimbursements from their Medicaid programs. The states with more advanced Medicaid systems include (but are not limited to) California, Maryland, Pennsylvania, and Wisconsin. As with the Medicare claims data, Medicaid claims include data on inpatient, outpatient, physician, pharmacy, and skilled nursing services. Also available are provider data and Medicaid-eligible beneficiary data. It is vital to secure the eligibility file to properly account for truncated beneficiary enrollment periods. For example, one Medicaid recipient may have been enrolled for one month, whereas another may have been enrolled for one year. If both recipients received an equal number of physician services during a calendar year, the absence of applying the denominator of enrolled months leads to faulty conclusions on service use. The quality of these data varies widely. For example, managed care capitation contracts may not require the collection of encounter information. Therefore, an analyst seeking to complete a multistate Medicaid study is faced with the task of understanding each state's claims data idiosyncrasies.

National and State Vital Statistics (Births, Deaths)

The NCHS makes available for purchase the complete event-level records of all births and deaths in the United States. To prevent disclosure of individuals and institutions, NCHS excludes (a) geographic identities of counties, cities, and metropolitan areas with less than 100,000 population, and (b) exact day of birth and death, although data with these fields populated may be requested for specific research projects. These data are also generally

available directly from the appropriate state health department for a nominal copying fee.

HIS Applications in Public Health Administration

There are several operating public HISs of note. These initiatives range in scope from federal to local sponsorship. Some provide a general database for a full range of public health issues, while others are designed for specific disease tracking or program evaluation.

The CDC's INPHO

The INPHO system was developed as a framework for public health information and practice based on a state-of-the-art telecommunications network.[5] The INPHO is part of a strategy to strengthen public health infrastructure. The three concepts of the INPHO are linkage, information access, and data exchange. First, the CDC works with state and local area health agencies to build local and wide area networks. Second, the CDC has expanded "virtual networks" through the use of CDC WONDER. This is a software system that provides access to data in the CDC's public health databases. Third, the CDC has encouraged each state to connect with the Internet to have access to information.

Georgia (discussed in more detail later in the chapter) pioneered the program in early 1993. By 1997, 14 more states made the INPHO vision integral to their public health information strategies: California, Florida, Illinois, Indiana, Kansas, Michigan, Missouri, New Jersey, New York, North Carolina, Oregon, Rhode Island, Washington, and West Virginia. A second round of INPHO projects was funded through a cooperative agreement program, with awards made in the spring of 1998. The program promotes the integration of information systems, with special emphasis on immunization registries. The cooperative agreements were funded as either implementation projects (Florida, Georgia, Missouri, and New York) or demonstration projects (Iowa, Maryland, Montana, Nevada, and Texas). More information on the initiative is available at http://www.phppo.cdc.gov/PHTN.

CDC WONDER

CDC WONDER was designed by the CDC to put critical information into the hands of public health managers quickly and easily. Originally a PC-based system, it is now available from any computer with an Internet connection, solving the problem of dedicating workstations to a specific database. As such, it is one of the few truly national public health data resources available with real-time access to anyone in the world. With CDC WONDER, one can do the following:

1. Search for and retrieve *Morbidity and Mortality Weekly Review* articles and prevention guidelines published by the CDC.
2. Query dozens of numeric data sets on the CDC's mainframe and other computers via fill-in-the blank request screens. Public use data sets about mortality, cancer incidence, hospital discharges, AIDS, behavioral

risk factors, diabetes, and many other topics are available for query, and the requested data can be readily summarized and analyzed.
3. Locate the name and e-mail addresses of the CDC staff and registered CDC WONDER users.
4. Post notices, general announcements, data files, or software programs of interest to public health professionals in an electronic forum for use by CDC staff and other CDC WONDER users.

For more information on CDC WONDER, refer to http://wonder.cdc.gov.

State Public HISs

States have multiple public HISs mirroring the complicated array of categorical programs with different funding sources. Commonly maintained information systems include computerized immunization registries, lead toxicity tracking, early intervention databases for children with a disability, congenital disease registries, in addition to vital statistics data, Medicaid utilization, and disease reports. The need for integrated information systems and the support of the INPHO project has spurred models in a number of states. The next sections describe efforts in Missouri, Georgia, Illinois, and New York.

Missouri

The Missouri Department of Health had a problem with 67 information systems that ran on different platforms and could not communicate with one another.[3] To solve this problem, the Missouri Health Strategic Architectures and Information System (MOHSAIC) was developed. An integrated client service record was an important component of this initiative. From the client's perspective, it was irrelevant if the services were labeled WIC, prenatal care, diabetes, Maternal and Child Health Services block grant, or local funding. Considerable effort and staff resources were committed to develop this system. Also, integrated systems magnify concerns about confidentiality. Benefits include increased capability for community health assessment, coordination of services, outreach, and linkages to primary care delivered by larger networks.[3]

Georgia

Georgia was the first site of the CDC INPHO initiative. Georgia was able to develop quickly as a demonstration site through a unique consortium of state agencies with academic health partners and IT partners. For example, members of the consortium included the Medical College of Georgia as well as the Georgia Center for Advanced Telecommunications Technology and the Emory University School of Public Health. The program also had initial funding from the Robert Wood Johnson Foundation.[2] The infrastructure includes 81 clinics and 59 county health departments.

The Georgia INPHO system includes local and wide area computer networks, office automation and e-mail, a public health calendar, an executive HIS, and electronic notification of public health emergencies. Before the project began, the state public health office operated 13 small unlinked local area

networks. With the INPHO project, hardware and software were consolidated into one integrated network system.

Illinois

Cornerstone is a management information system developed in Illinois to integrate maternal and child health services. The design expands on the existing WIC program PC-based computer system.[8] This system is an example of a state information system integrating several related programs as compared with wide integration pursued by Missouri, Georgia, and New York State. Risk assessment and demographic information are captured once and used for multiple programs. Exchange of information, risk assessment, assurance of follow-up, and referral are assisted by this information system.

New York

New York State is implementing an ambitious and far-reaching plan for the integration of public health information. Development of this information system was assisted by funding from the CDC. The New York State Department of Health has developed an enterprise-wide infrastructure for electronic health commerce. This effort has three major components.

- A public Web site (http://www.health.state.ny.us/) of health information serving as an Internet portal with an average of 850,000 hits per week and provider of data to consumers, researchers, and providers on health issues and data
- The Health Information Network (HIN), a public and private health data interchange of information
- The Health Provider Network (HPN) targeted at private data information interchange between state and healthcare providers including clinical laboratories, managed care plans, pharmacies, hospitals, and continuing care facilities

The New York health e-commerce initiative is using the Internet and Web page interface to connect users and databases in a secure environment. For the HIN, the Web-based interface functions as a closed intranet where Web encryption of secure socket layers is established (though transparent to the user) to protect the security and confidentiality of data.[9]

A large effort has been undertaken to ensure information security on the HIN because of the confidential nature of data transactions between state and local public health departments. Organizational and individual security agreements are required for HIN access.[10] Very narrow access is provided for highly confidential items such as case reports for notifiable disease. Particular restrictions and security arrangements are in place for HIV reporting. More broadly defined access exists for statistical data queries.

Future Public Health Information Systems

The broad range of public HIS applications developed over the past 10 years demonstrates how managers are seeking to improve the scope and quality of their data systems. HIS experts consistently state that the future lies in build-

ing an infrastructure that is both easy to use and able to demonstrate value for its investment.[11,12] To build such an infrastructure requires data standards as well as translators for different standards to help bridge the transition from the current system.[13]

One of the most promising developments is the use of the Internet as the platform to collect data, turn data into information, and monitor the health of the population. The development of Internet-based software that is not dependent on operating systems or statistical computing software represents one less barrier to building an infrastructure. (The Internet's ease of software deployment through the use of a simple connection and a Web browser will lead to faster dissemination of standard data translation tools.) Even more powerful is the transition of the medical profession from an arcane paper-based data collection world toward e-commerce for business-to-business applications where new standards can be applied from the beginning of data collection and management activity, not retrofitted. Public concerns about the privacy of health-related information in this new environment are motivating new policies for information use that, it is hoped, will build the public's trust in emerging health information applications while preserving the ability of public health organizations to use health data for essential surveillance, research, and management activities.

Amid the opportunities for developing HISs, substantial barriers remain, but these barriers are becoming less technical and more political. Public health managers seeking to develop and use the IT infrastructure must be prepared to demonstrate its value to society constantly.

Privacy Issues

The public's concern for the privacy of personal health information has become a major policy issue. Unfortunately, this concern is not easily addressed. At the heart of this issue is the paradox that health data must be identifiable if they are to be valuable for public health interventions. Complicating the issue is that even an encrypted personal identifier still yields a personal identifier. HISs must remain responsive to these evolving data privacy and confidentiality issues.

The public's desire for health data privacy appears to exceed its desire for public health and biomedical research. In a 1993 Lou Harris survey on the public's attitudes on health data privacy, 64% of the sample responded that they did not want medical records data used for biomedical research unless the researchers obtained the patient's consent. When asked if they favor the creation of a "national medical privacy board" to hold meetings, issue regulations, and enforce standards for protecting medical information privacy, 86% responded favorably.[14]

Two recent developments advanced the privacy debate. The first development was the passage of the Health Insurance Portability and Accountability Act (HIPAA) in 1996, which created a timetable for the adoption of national medical privacy legislation by the year 2000. The combination of HIPAA and privacy laws was adopted to ensure health coverage after leaving employment, while also creating the first national policy to prosecute those persons who breach the medical privacy of an individual. The penalties can range from fines to prison. The compliance date for the privacy rule was April 2003.

Protected health information (PHI) is individually identifiable health data that is transmitted or maintained in electronic media and related to the physical or mental health of an individual, the healthcare services provided to an individual, or the payment for those services provided to the individual. For covered entities using or disclosing PHI, the Privacy Rule establishes a range of health information privacy requirements and standards, including procedures for notification of individuals, internal policies and procedures, employee training, and technical and physical data security safeguards.

Public health practice and research uses protected health information to perform many of its required functions, including public health surveillance, outbreak investigation, program operations, terrorism preparedness, and others. Public health authorities have a long history of protecting the confidentiality of individually identifiable health information, and were given significant latitude in the Privacy Rule, which expressly permits PHI to be shared for specified public health purposes. Covered entities may disclose PHI to a public health agency legally authorized to collect information for the purpose of preventing or controlling disease, injury or disability, without separate authorization. It should be noted, however, that in addition to using PHI from covered entities, a public health agency may itself be a covered entity, providing services and producing covered electronic transactions.[15]

Of particular interest for both research and population level assessment are the use of deidentified information and limited use data sets. Deidentified data (stripped of individual identifiers rendering it "impossible" to associate a record with any individual) require no individual privacy protection and are not covered by the Privacy Rule. Deidentification can be accomplished by using accepted analytical techniques to conclude that the subject of the information cannot be identified or by removing 18 specific identifier fields (the "safe harbor" method) to render identification infeasible. Limited data sets may contain some of the 18 identifiers, as long as other safeguards are provided to prevent subject identification.

Ultimately, data are provided to public health managers and researchers as an act of trust. If one individual or organization violates that trust, the public's confidence may erode immediately. The Harris poll results show consistently that health data confidentiality and security issues are an important public concern.[12]

• • •

In developing and using HISs, public health administrators and researchers must demonstrate that the public's trust is deserved. To do so, contemporary HISs must ensure that society receives an optimal return on its public investments in data resources—a return that ultimately must be realized through more effective public health interventions and improved community health status.

Chapter Review

1. Public health organizations rely on information systems to support a number of key operations, including:
   - Epidemiologic disease and risk factor surveillance
   - Medical and public health outcomes assessment

- Program and clinic administration (billing, inventory, client tracking, clinical records, utilization)
- Cost-effectiveness and productivity analysis
- Utilization analysis and demand estimation
- Program planning and evaluation
- Quality assurance and performance measurement
- Public health policy analysis
- Clinical research
- Health education and health information dissemination

2. Two of the most common types of applications for information systems in public health organizations are:
   - Service-based applications that track encounter-level information on the users and providers of specific services. These applications are useful for program administration and management of services for individual clients.
   - Population-based applications that track information on defined populations of interest. These applications are useful for surveillance, program evaluation, planning, and policy development.
3. To build the public health information systems of the future, data must be extracted from many operational systems, integrated for analysis, and disseminated using multiple technologies such as Web portals. Service-oriented computing is an emerging information systems architecture that supports the construction of networked systems that are flexible, scalable, and reliable. There are many evolving standards to support the development of loosely coupled components that can be assembled into complex systems.
4. Data warehouses are "subject-oriented, integrated, nonvolatile, time-variant collection[s] of data in support of management's decisions." Data warehousing technologies have matured and found wide application in many industries. In the healthcare domain, these technologies offer a powerful method of integrating data for community health assessment, surveillance, clinical decision support, and outcomes review.
5. A variety of data sources can be integrated within a public HIS. Common data sources include survey data, administrative claims data, program administration data, regional and geographic data, registry data, and private industry data. Common data structures include transactional data, cross-sectional data, time series and panel data, and relational databases.
6. Well-designed public HISs must provide strong protections for the privacy and confidentiality of information derived from person-specific health-related data. These protections must cover data acquisition, storage, and linkage and retrieval activities, as well as analytic and reporting activities. Information systems must be responsive to the privacy provisions of recent federal and state legislation.

References

1. Koo D, Wetterhall SF. History and current status of the National Notifiable Diseases Surveillance System. *J Public Health Manage Pract.* 1996;2(4):4–10.

2. Chapman KA, Moulton AD. The Georgia Information Network for Public Health Officials (INPHO): a demonstration of the CDC INPHO concept. *J Public Health Manage Pract.* 1995;1(2):39–43.
3. Land GH, Stokes C, Hoffman N, Peterson R, Weiler MJ. Developing an integrated public health information system for Missouri. *J Public Health Manage Pract.* 1995;1(1):48–56.
4. Berndt D, Hevner A, Studnicki J. The CATCH data warehouse: support for community health care decision making. *Decision Support Syst.* 2003; 35:367–384.
5. Corrigan JM, Nielsen DM. Toward the development of uniform reporting standards for managed care organizations: the Health Plan Employer Data and Information Set. *J Joint Commission Qual Improv.* 1993;19(12): 566–575.
6. Huhns M, Singh M. Service-oriented computing: key concepts and principles. *IEEE Internet Comput.* 2005;9(1):75–81.
7. Inmon W. *Building the Data Warehouse.* 3rd ed. New York, NY: John Wiley and Sons; 2002.
8. Nelson JR. Cornerstone: Illinois' approach to service integration. *J Public Health Manage Pract.* 1996;2(1):71–74.
9. Gotham I. Personal communication, NYS Department of Health, August, 1999.
10. Baker EL, Friede A, Moulton AD, Ross DA. CDC's Information Network for Public Health Officials (INPHO): a framework for integrated public health information and practice. *J Public Health Manage Pract.* 1995;1(1):43–47.
11. Baker EL Jr, Ross D. Information and surveillance systems and community health: building the public health information infrastructure. *J Public Health Manage Pract.* 1996;2(4):58–60.
12. Milio N. Beyond informatics: an electronic community infrastructure for public health. *J Public Health Manage Pract.* 1999;1(4):84–94.
13. Lumpkin J, Atkinson D, Biery R, Cundiff D, McGlothlin M, Novick LF. The development of integrated public health information systems: a statement by the Joint Council of Governmental Public Health Agencies. *J Public Health Manage Pract.* 1995;1(4):55–59.
14. Lou Harris and Associates. *Health Information Privacy Survey, 1993.* New York, NY: Harris; 1993.
15. Centers for Disease Control and Prevention, Epidemiology Program Office. HIPPA Privacy Rule and public health. *MMWR.* 2003;52:1–12.