The Johns Hopkins University Applied Physics Laboratory Center of Excellence in Public Health Informatics

Joseph S. Lombardo

For the past 10 years, APL has been supporting the public health disease surveillance mission by developing and implementing automated applications to acquire, process, and present data using modern information technology. The effort grew from a modest independent research and development activity to APL's receiving an award as one of the National Centers for Disease Control and Prevention's Centers of Excellence in Public Health Informatics. This issue presents some of the recent work being performed within the Center. This article provides an introduction to disease surveillance and the work within the JHU/APL Center of Excellence in Public Health Informatics.

INTRODUCTION

Disease has at times contributed to the decline of empires and civilizations. For decades, one of the more critical security issues countries faced was the health of their population and those who defend that population. Smallpox has accounted for tens of millions of deaths and has been linked to the decline of Aztec and Inca empires in Mexico and the North American Indians.[1,2]

In 1918–1919, Spanish influenza infected up to one-third of the world's population.[3] The number of dead was estimated to be between 20 and 40 million, with the exact numbers unknown because of inadequate reporting. Occurring near the end of World War I, the disease incapacitated military training facilities preparing troops to enter the war.

The development of biological weapons as a source of generating disease outbreaks accelerated prior to and during World War II. The Japanese attacked numerous Chinese cities with biological weapons, infected food, contaminated water supplies, and performed human experimentation resulting in the loss of hundreds of thousands of Chinese lives.[4] In 1941, the United States, Canada, and the United Kingdom initiated a biological weapons program that resulted in the development of an anthrax weapon that could have cost thousands if not millions of German lives. After World War II, the Soviet Union accelerated the development, manufacture, and stockpiling of biological weapons. At the same time, the United States focused its resources on the nuclear threat with the belief that modern medicine could cure infectious diseases and the Biological Weapons Convention would end the development of engineered super-bugs. The rapid spread of HIV and severe acute respiratory syndrome (SARS), the appearance of West Nile virus in the United States, the appearance of antibiotic-resistant disease strains, the ease with which anthrax was distributed through the U.S. Postal Service, and the fear of a new highly virulent form of avian influenza have provided a reawakening to the need for early detection and control of major health risks.

The public health infrastructure within the United States consists of local, state, and federal components. The burden of surveillance, containment, and reporting is at the local level, with support being provided from state and federal agencies. Because local health departments are limited in terms of fiscal and human resources, public health workers must focus on the most urgent issues with little time to support continuing surveillance activities. Providing local health departments with automated tools to support surveillance activities is currently recognized by most public health agencies as a necessity.

APL's interest in the detection of biological weapons began after Operation Desert Storm with the recognition of Iraq's weapons program and the activities of the former Soviet Union's Biopreparat.[5] Most of the activity was centered on detection of pathogens in the air. In 1997, APL began seeking sponsorship to develop a prototype disease surveillance system that would automate the collection and processing of data containing indicators of disease. In collaboration with the Maryland Department of Health and Mental Hygiene and the Maryland Emergency Management Agency, APL developed a prototype and demonstrated the system for state health officials. The system not only could review collected data but also could apply automated signal-processing techniques to outbreak detection and view the data geographically to locate clusters of infection. In 1999, with the state health department's endorsement, funding for a seedling project was received from the Defense Advanced Research Projects Agency (DARPA) to obtain and effectively use health-indicator data and perform surveillance during the celebrations leading up to and following the change of the millennium.[6] The seedling project acquired health-indicator data from Fort Meade for the winter flu season surrounding Y2K. In early spring 2000, the project learned about a similar effort being conducted by the Walter Reed Army Institute of Research, which focused on the military population in the National Capital Region. The two projects merged; APL and Walter Reed signed a collaborative research and development agreement and submitted a joint proposal to DARPA, seeking funds to continue working on the technology and expanding it to both military and civilian public health authorities across the region. Based on the success of that project and the increasing concern over terrorism, DARPA initiated the Bio-event Advanced Leading Indicator Recognition Technology (Bio-ALIRT) Program. The objective of this program was to identify data sources and analytical techniques that could provide early recognition of disease in the population under surveillance. The terrorist attacks and anthrax-containing letters of 2001 demonstrated the commitment of terrorist groups to use whatever means possible to kill innocent citizens. It also drove home recognition of the urgent need for automation tools to support health departments' disease surveillance mission.

PUBLIC HEALTH INFORMATICS

Public health informatics has been defined as the systematic application of information and computer science and technology to public health practice, research, and learning.[7] It is a relatively new discipline, with just a few major universities beginning to offer training in the field. Public health informatics is different from medical informatics because public health is more concerned with the health of populations rather than delivering care to specific individuals. The APL program currently focuses on the disease surveillance mission within public health, but technology being developed may be more broadly applied in the future to other urgent public health issues that may impact national security.

The mission of APL's disease surveillance program is to reduce the mortality/morbidity from an unexpected outbreak by the insertion of advanced technology into traditional disease surveillance and response. Figure 1 provides a summary of the structure of the APL

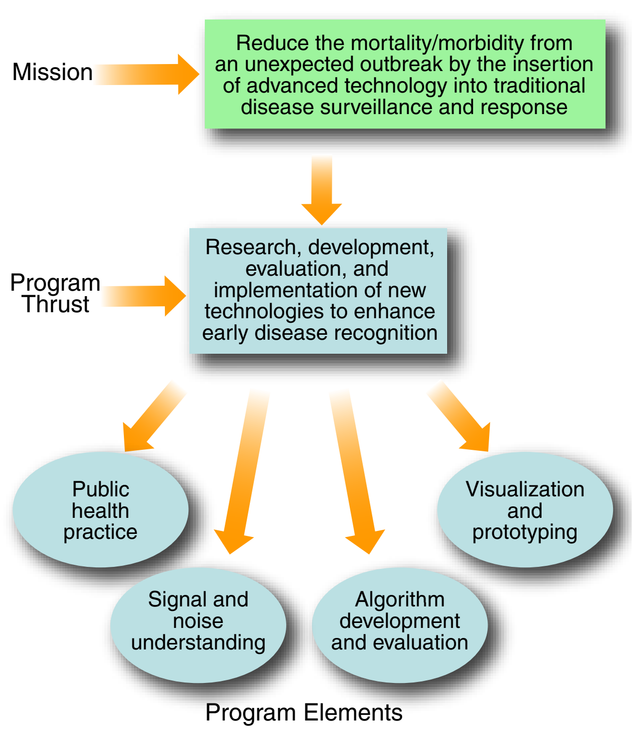

Figure 1. Elements of the APL disease surveillance program.

program. The program is divided into four basic elements. The first element is understanding the data available for surveillance through characterization of the signals and background noise. Because the objective is early disease recognition, many of the current data sources available as health indicators are nonspecific. The second element is the development and evaluation of detection algorithms that can be used to recognize the start of an outbreak. The third element is the efficient interfacing between automation and the epidemiologists who are responsible for disease monitoring. The last element, which may be the most important, is performing the operations research needed to effectively insert an automated surveillance tool into public health practice.

AUTOMATED DISEASE SURVEILLANCE SYSTEMS

To understand the skills that APL has applied to automating disease surveillance, we must first examine the components of a modern disease surveillance system. Disease surveillance systems can be placed into two broad categories: active surveillance and passive surveillance. Active systems require data to be entered specifically for the purposes of public health surveillance. These systems require caregivers or other employees to spend time entering data. The systems have been shown to be useful for special events but have not been sustainable because of the extra time and expense needed to enter data. Passive surveillance systems rely only on data collected for other purposes. Once the process of automating the transfer of data to the surveillance system has been completed, little additional time is required other than for routine maintenance. This article will focus only on passive surveillance systems.

Figure 2 presents the major components of a typical modern disease surveillance system. The system operates on one or more data streams containing indicators of health for the population. An understanding of the characteristics of a signal representative of the early stages of an outbreak must be known as well as the background noise, which can create false alarms. Surveillance systems currently use civilian and military patient encounter data from hospital emergency departments, billing disease codes from insurance claims, data from private practice office visits, military disease codes from military clinics, information pertaining to emergency medical system (paramedic) services, records of prescription and over-the-counter medication sales, records of school absenteeism, information from poison control center calls, and information from 911 calls. Most data are encrypted before transmission. Once the data are received and decrypted, preprocessing must occur to account for errors in coding or duplicate records. Unstructured textual data must be converted into structured data, and an archive must be created for use by other system components. Data can be grouped into syndromes, or custom groups can be created and analyzed separately.

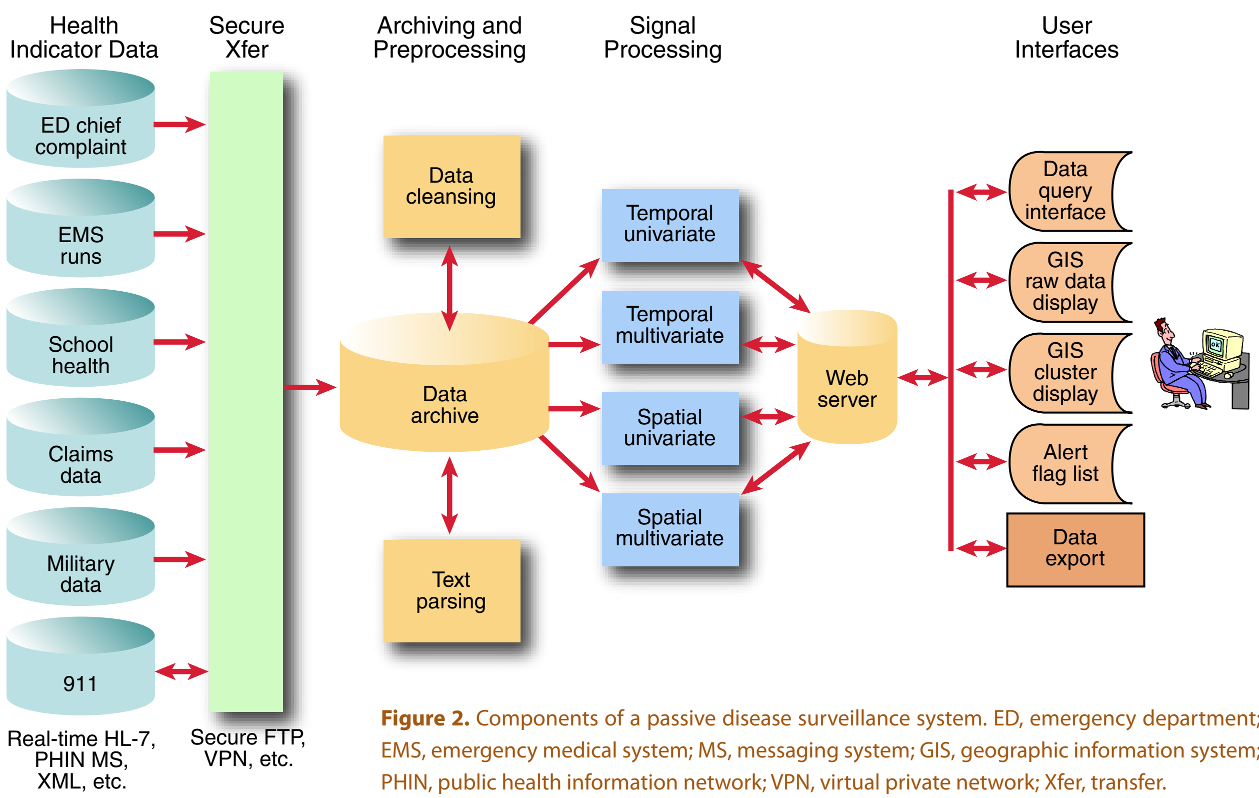

Figure 2. Components of a passive disease surveillance system. ED, emergency department; EMS, emergency medical system; MS, messaging system; GIS, geographic information system; PHIN, public health information network; VPN, virtual private network; Xfer, transfer.

Various different processing algorithms are currently being used to detect the presence of a signal. A few of these algorithms are described by Burkom et al. in the article "Developments in the Roles, Features, and Evaluation of Alerting Algorithms for Disease Outbreak Monitoring" elsewhere in this issue. In many ways, the signal-processing component is similar to detecting radar or sonar signals, except the signals associated with an outbreak can vary widely and detection must be accomplished as early as possible, while the signal is just starting to develop.

The last major component is the interface to the user. Most modern disease surveillance systems use a web server so that users can access and review surveillance data over an encrypted Internet connection. User interfaces consist of inputting parameters and data elements to create custom detection and analysis graphics as well as geographical representations to view both the temporal and spatial characteristics of data clusters. Some of these visualization techniques are described by Loschen in the article "Methods for Information Sharing to Support Health Monitoring" elsewhere in this issue.

Most readers will recognize that the components and data flow are very similar to most data collection and analysis applications. The primary difference is the characteristics of the individual data streams. As a result of decades of experience gained in the development of similar systems in different domains, APL developers were able to rapidly develop a disease surveillance tool for public health.

THE CENTERS OF EXCELLENCE IN PUBLIC HEALTH INFORMATICS

In recognition of the role informatics will play in the future of public health practice, the Centers for Disease Control and Prevention (CDC) established the National Center for Public Health Informatics (NCPHI).[8] NCPHI consists of five divisions internal to CDC and five academic Centers of Excellence where leading research is performed in informatics. To better understand the need for public health informatics research, a quick look at the changing disease surveillance environment is warranted.

Existing automated disease surveillance systems have been relying on health-indicator data collected for other purposes. Although these data are relatively easy to obtain, the data sources are not very specific in terms of identifying diseases. The past half decade has seen the formation of Regional Health Information Organizations (RHIOs), currently known as Health Information Exchanges (HIEs).[9] These exchanges are making it possible to provide patient medical records to every caregiver in a region that supports the patient, and they provide a one-stop shopping approach to obtaining health-indicator data, making it easy for health departments to acquire patient medical records across the region for the purpose of surveillance. A second change is the expansion of the surveillance role to include tools for continuous monitoring of the population after an event has been identified. This function helps health departments support disease-control measures. Figure 3 is an illustration of the changing regional environment.

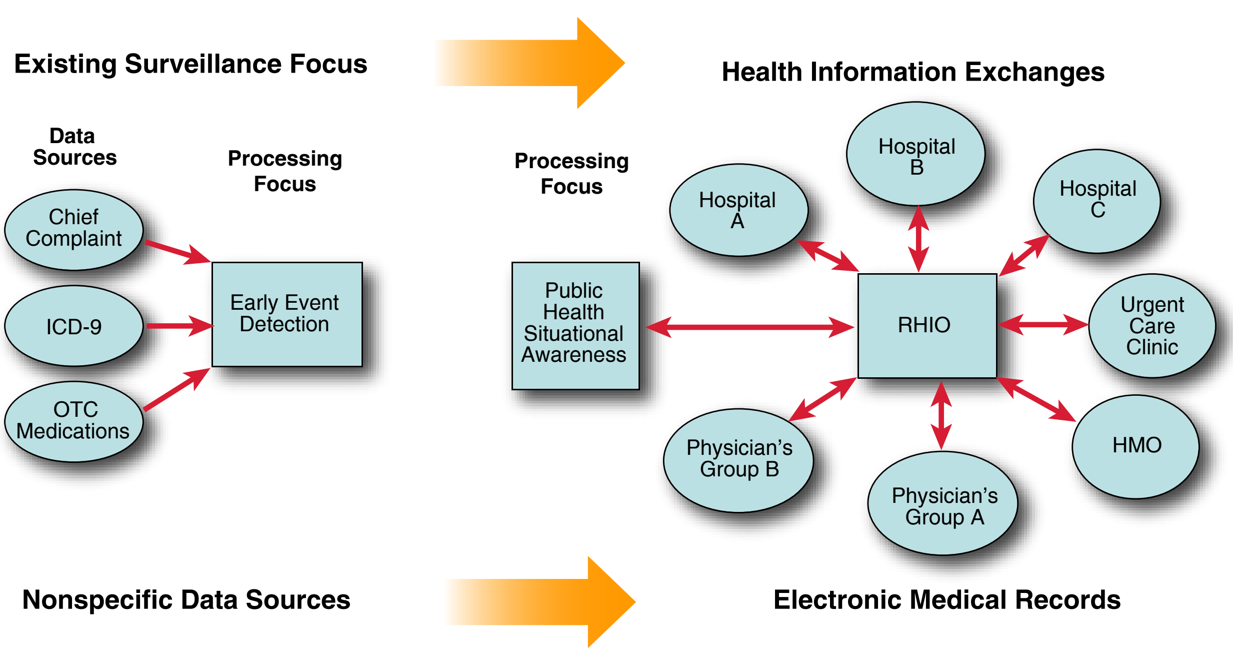

Figure 3. The changing environment in disease surveillance. ICD-9, International Classification of Diseases, Ninth Revision; OTC, over the counter.

At the national level, the formation of the National Health Information Network (NHIN) will make it possible for health departments and HIEs to exchange patient data nationally.[10] This network will facilitate the exchange of information needed to support surveillance among local health departments and between health departments and federal agencies. Figure 4 provides an illustration of the concept of data and information exchange over the NHIN.

CDC'S BIOSENSE PROGRAM

One of the initiatives within NCPHI is the development and operation of a national disease surveillance system known as BioSense. The intent of this program is to obtain real-time medical record data from as many health care facilities across the country as possible, perform automated surveillance within CDC, and make those results available to health departments nationwide. BioSense is currently collecting a wealth of data, but only a small portion of the data is currently being used because new methods are urgently needed to best fuse the data so as not to increase false alarms and burden those monitoring surveillance data.

RESEARCH PERFORMED BY THE CENTERS OF EXCELLENCE

The CDC funded two Centers of Excellence in FY05 at Harvard Pilgrim Health Care and the University of Washington and three Centers of Excellence in FY06 at The Johns Hopkins University, the University of Utah, and the New York City Department of Health and Mental Hygiene. A brief description of the research being conducted at each Center of Excellence is provided.

Harvard Pilgrim Health Care Center of Excellence

The Harvard Pilgrim Health Care Center of Excellence is a large health maintenance organization (HMO)

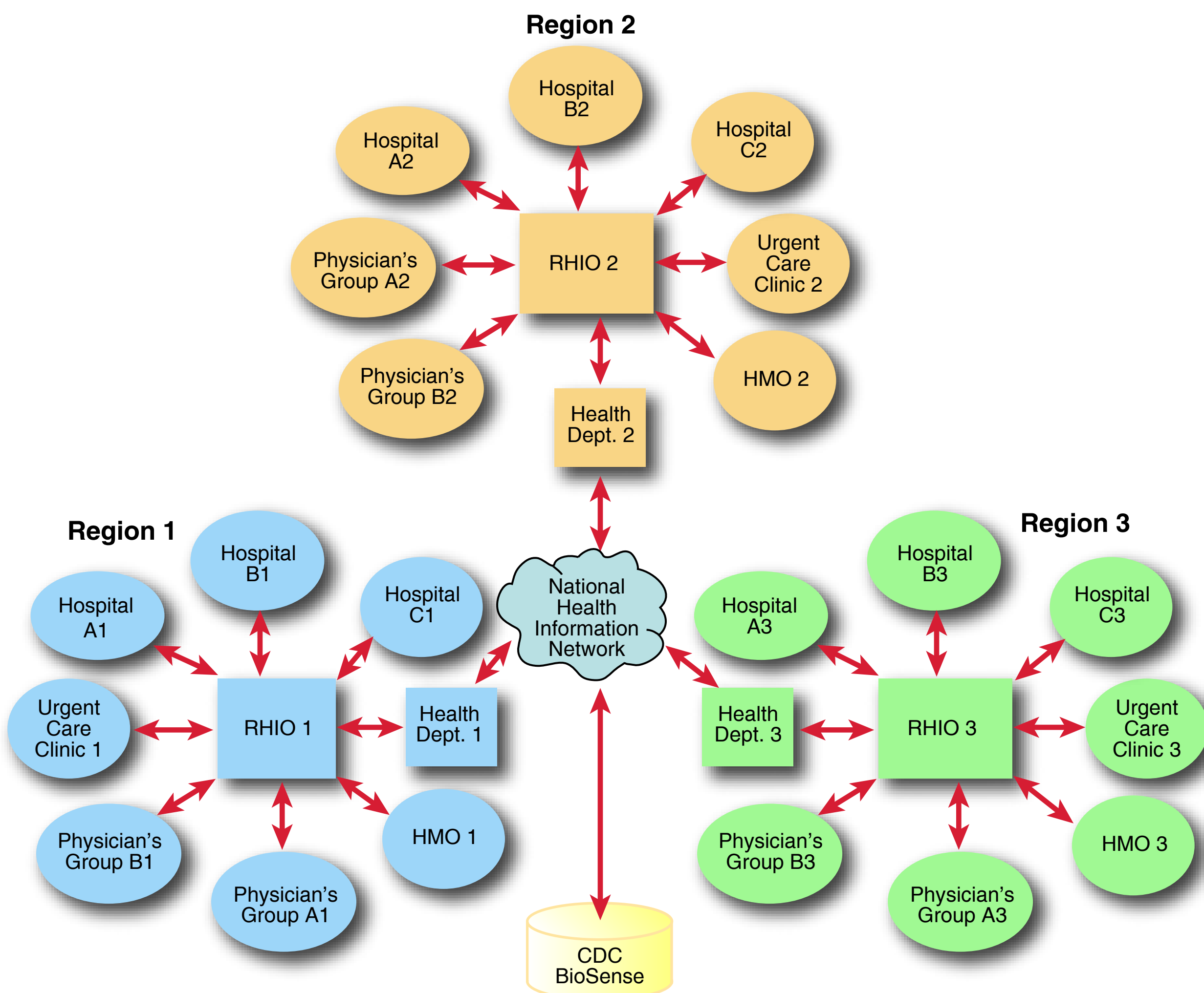

Figure 4. Data exchange over the NHIN.

in the New England region of the United States. This Center of Excellence is performing research on how electronic medical records can be used to enhance the practice of public health. Specific projects include the automated reporting of notifiable conditions.[11] This research is being performed in conjunction with Harvard Medical School. A second focus is the creation of a Personally Controlled Health Record system, which will permit individuals to have and control access to their medical records.[12] This research is being performed in conjunction with the Massachusetts Institute of Technology (MIT) and the Children's Hospital of Boston.

University of Washington Center of Excellence

The University of Washington Center of Excellence is performing research on decision support for disease surveillance. The project studies the value of RHIO data and data-management practices and novel algorithmic approaches to enhance disease surveillance and reporting. The MyPublicHealth project is designing, developing, and evaluating a digital interactive knowledge system to support public health practice.[13]

University of Utah Center of Excellence

The University of Utah Center of Excellence is performing research for automated monitoring and reporting of infectious disease through collaboration with Intermountain Health Care, University Health Care, and the Salt Lake City VA Hospital. Their program is developing decision support tools to help epidemiologists streamline their investigations for infectious diseases. This Center of Excellence also performs research into how to enhance bidirectional communications between health care providers and public health officials.

New York City Department of Health and Mental Hygiene Center of Excellence

The New York City Department of Health and Mental Hygiene Center of Excellence has acquired funding to place new electronic medical records systems into each hospital within the city. These systems will be connected to the health department so that continuous monitoring of the city's population will eventually occur. The New York City Center of Excellence is a collaboration with Columbia University. The research within this Center is focused on the development and evaluation of methods to best exploit the wealth of data that will become available to the health department.

JHU/APL Center of Excellence in Public Health Informatics

The JHU/APL Center of Excellence is focusing its research on improvements needed for automated disease surveillance. There are three efforts within this Center of Excellence. The first project is developing methods to use the electronic medical records for public health surveillance. Medical records are created as a result of encounters with health care providers and from orders given to diagnose and treat illness. These records hold the promise of providing the details needed to improve the specificity of detection processes and minimize the time spent tracking down unimportant alerts.[14]

The second project is examining methods for sharing information among health departments. Privacy laws permit the use of medical records for public health surveillance by health departments. This provision in the law only refers to health care facilities within the jurisdiction of the health department. Because large populations can cover many public health jurisdictions and modern transportation permits rapid movement of contagious persons across many jurisdictions very quickly, health departments must be aware of what is going on around them and what is occurring globally if there is any suspicion that a health risk may move into their jurisdiction. Sharing information and not raw data has a lot of potential. The focus of the project is what information to share and how to share it.

The last objective is forming collaborations with researchers at CDC or other Centers of Excellence to leverage each others' talents for streamlining research. So far we have had successful collaborations in the development of an advanced query tool to support the analysis and development of synthetic medical records data to be used for system development, evaluation, and training.

Incorporating Electronic Medical Records into Automated Disease Surveillance Systems

A brief introduction to medical records is provided so that the reader can obtain a better understanding of the research being performed within the Center. Some of the data elements contained within the electronic medical record that are of value to public health surveillance are shown in Table 1. Most modern electronic medical record systems contain data in multiple formats. As an example, when a patient is seen in a clinic or emergency department, the reason for the encounter or chief complaint is recorded as a text field, whereas a prescription for a medication can be selected from a pull-down list and can appear in the record in a structured format. The format for requesting a test from the microbiology laboratory would also be structured, but the content of the element is very different from a prescription medication data element.

When the data become available also varies with the data element. Taking an x-ray while a patient is in the emergency department may only take an hour, but obtaining the report from the radiologist could take several hours, if not longer.

Table 1. Typical data elements contained within an electronic medical record.

Data element	Format	When available
Time and date	Structured	Patient encounter
Medical record number	Structured	Patient encounter
Patient chief complaint	Text	Patient encounter
History of present illness	Text	Patient encounter
Laboratory test request	Structured	Patient encounter
Imaging request	Structured	Patient encounter
Medication orders	Structured	Patient encounter
Clinic notes	Text	After encounter
Laboratory test results	Structured	After encounter
Imaging results	Text	After encounter
ICD-9 billing codes	Structured	After encounter

Another issue with using electronic medical record data for public health surveillance is that many of the tests or imaging requests are done to rule out diseases of concern by the physician seeing the patient. The requests are available in the medical record soon after the encounter and are only an indication of the level of concern of the physician and not necessarily an indication of the disease of the patient. Health care workers are accustomed to viewing the data and information present in the record in all of the various formats. The challenge for an automated disease surveillance system is for a series of automated processes to replicate the data and information fusion process being performed by these workers. The approach being taken by the APL project is to use data provided by the electronic medical records to obtain a measure of severity of the illness of interest for the patients being seen at each facility, thereby filtering out many of the encounters previously counted in broad categories, such as respiratory or gastrointestinal, which have little significance with regard to the health of the community. An additional filter is performed by age group before a second level of fusion is performed with aggregated counts of less-specific health-indicator data. The process is depicted in Fig. 5. The method includes a hybrid of both statistical and probabilistic algorithms to accomplish the fusion across several data types to identify a population with severe illness.

A more detailed description of the analytical development being performed to support incorporation of medical records into disease surveillance is provided elsewhere in this issue by Mnatsakanyan and Lombardo in the article "Decision Support Models for Public Health Informatics."

Information-Sharing Research Project

Blum and Duncan[15] describe objects that are processed by computers as data, information, or knowledge.

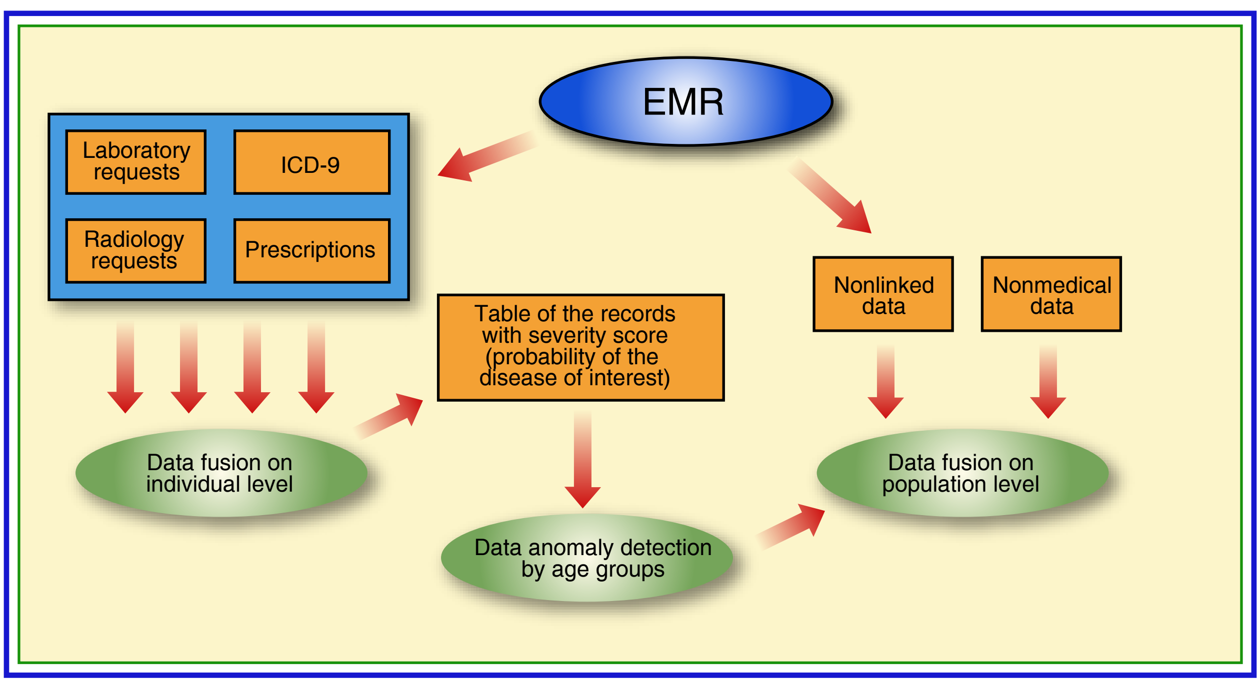

Figure 5. Process for obtaining increased specificity for population health risks from electronic medical records (EMR).

The "data" designation refers to a single measurement, element of demographics, or physical condition made available to the computer application or analyst. "Information" is a set of data with some interpretation or processing to add value. "Knowledge" is a set of rules, formulas, or heuristics applied to the information and data to create greater understanding. Figure 6 applies these principals to the indicators of health and the outputs of processes applied to them. At the lowest level are data in the form that they were created. At the highest level is the interpretation of the epidemiologists and analysts who use the surveillance tools. Interpretation of results and conditions is best performed by those who are responsible for the health of their populations. The data on the lower levels could provide elements that might be used to identify one or more individuals. The Health Insurance Portability Accountability Act enacted by the U.S. Congress in 1996 prohibits sharing such identifying information without the prior consent of the individuals about whom the data are being shared. Sharing population health status information without data elements that could be used to identify individuals is more acceptable because there is a mutual benefit to health departments to know what is going on in neighboring communities. Automated processes can be developed to use information outside of the community to determine risk within the community. This tool could be of great potential benefit in resource-limited settings that do not have the time or personnel to monitor what is going on outside of their community.

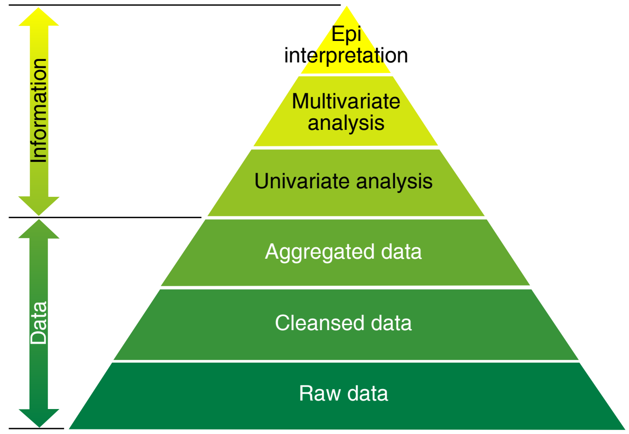

Figure 6. Comparison among data and health-indicator knowledge. Epi, epidemiologist.

One potential solution is to provide interpreted data as a web service to an application that displays national or global summaries of health status securely to other public health entities. This information would then be available for use within local disease surveillance systems. Figure 7 presents the information-exchange concept using web services.

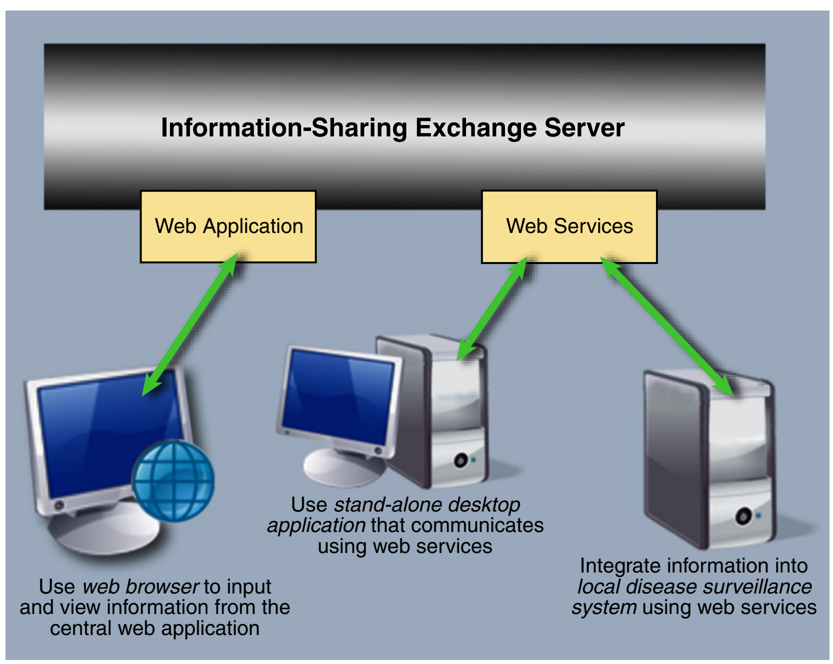

Figure 7. Disease surveillance information exchange using web services.

COLLABORATIVE RESEARCH PROJECTS

APL is currently collaborating with CDC researchers on two projects. The first is the development of a query tool to support surveillance analysis using any of the data elements available within the database. With the addition of medical records in surveillance systems, there are more types of data available to perform analysis that better defines the cases under study. Most public health analysts do not have the time or skills to perform complex data queries that group patients by combining or excluding them based on their chief complaints and results of diagnostic tests performed. The advanced query tool provides an intuitive approach to extracting groups of individuals that satisfy the query. Figure 8 provides an illustration of the process. The user is guided through a series of pull-down selections to create a query, which is then verified and executed.

The increased surveillance system functionality provided by electronic medical records and information exchanges must be developed and evaluated. System users must be trained to efficiently exploit these new functions. To perform these tasks, there is a need for a wealth of data containing indicators of outbreaks with components of the signal in each of the elements of the medical record.

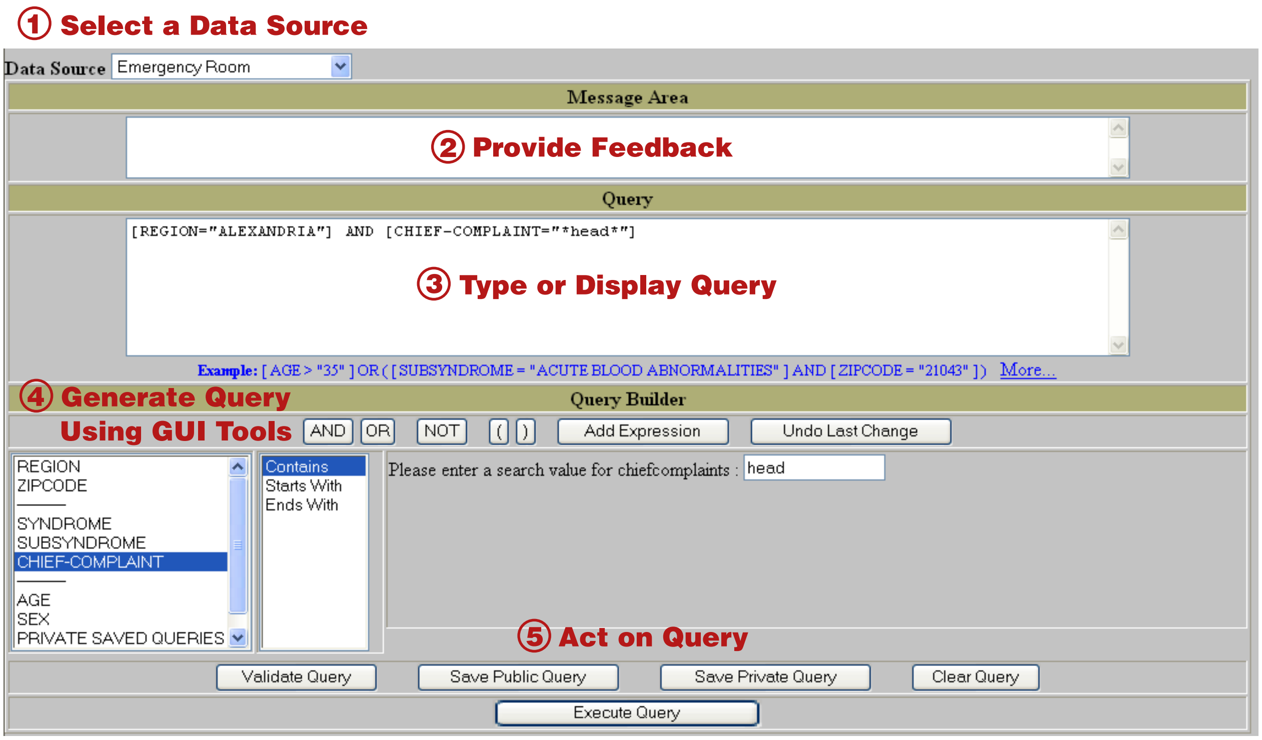

Figure 8. Advanced query tool user interface. GUI, graphic user interface.

Real medical record data containing indicators of outbreaks caused by bioterrorism do not exist. A second project within the JHU/APL Center of Excellence in collaboration with personnel from CDC is the creation of a model to synthesize medical records reflecting these outbreaks. The model is being developed by extracting the protocols for diagnosis and treatment from the electronic medical records and determining what tests, medications, and procedures are ordered for patients with specific chief complaints. The model will then create these data elements for artificial patients that have been added to represent an outbreak. The approach builds on the previous work of Buckeridge et al.[16] in developing a hybrid model for patients presenting at emergency departments with specific chief complaints. Figure 9 illustrates the process of creating synthetic medical records that have been injected for patients with specific chief complaints. This approach for creating synthetic electronic medical record

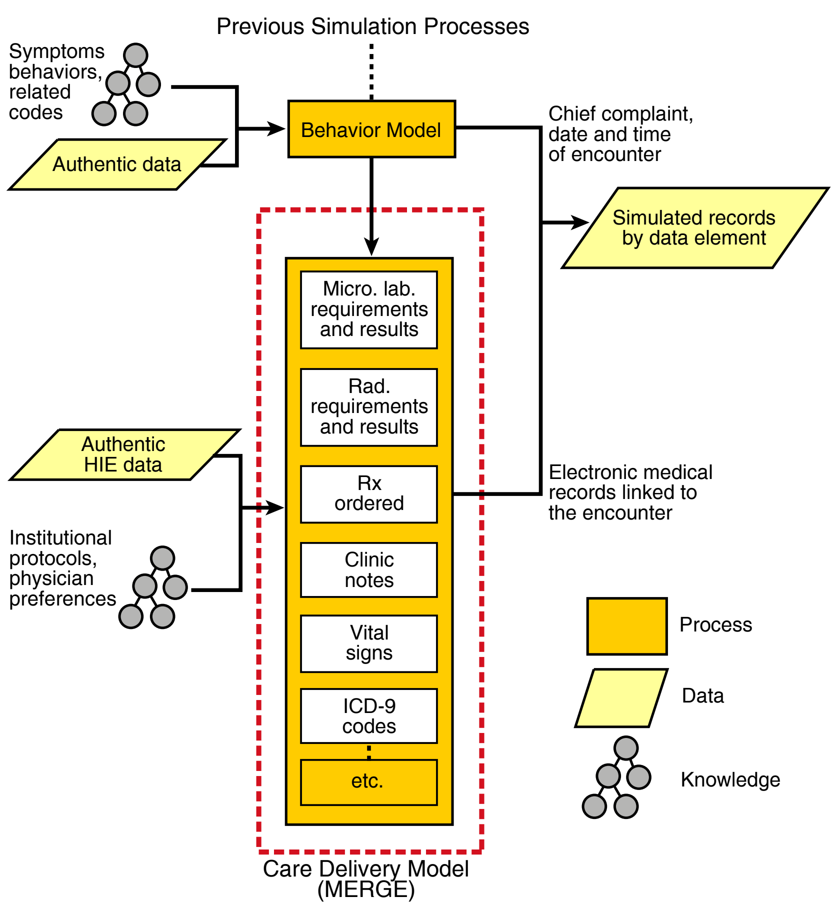

Figure 9. Medical records data generation using a care delivery model.

data will be described by Lombardo and Moniz in the article "A Method for Generation and Distribution of Synthetic Medical Record Data for Evaluation of Disease-Monitoring Systems" elsewhere in this issue.

CONCLUSIONS

Rapid identification and control of highly infectious diseases that cause high mortality and morbidity are a national security priority.[17] Rapid identification of these health risks is becoming an ever-increasing challenge because of the presence of highly virulent strains, rapid population movement, and potential bioterrorist activities. For automated disease surveillance systems to support this mission, they must increase their sensitivity and specificity for detecting a wide variety of public health risks. The JHU/APL Center of Excellence in Public Health Informatics is exploring electronic medical records as a more specific source of data. The Center is exploring methods to share information across jurisdictional boundaries and developing new analytical tools to fuse data and information. The articles that appear in this issue of the *Johns Hopkins APL Technical Digest* provide insight into the work being performed by the Center.

ACKNOWLEDGMENTS: This article was supported by Grant P01 HK000028-02 from the CDC. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CDC.

REFERENCES

[1]Duffy, J., *Epidemics in Colonial America*, Louisiana State University Press, Baton Rouge, LA (1953).

[2]Duffy, J., "Smallpox and the Indians in the American Colonies," *Bull. Hist. Med.* **25**, 324–341 (1951).

[3]Kilbourne, E. D., "Influenza Pandemics of the 20th Century," *Emerg. Infect. Dis.* **12**(1), 9–14 (2006).

[4]Gold, H., *Unit 731 Testimony*, Yenbooks, Singapore (1996).

[5]Alibek, K., and Handelman, S., *Biohazard*, Random House, New York (1999).

[6]The Johns Hopkins University Applied Physics Laboratory, *ESSENCE: Protecting Public Health*, http://www.jhuapl.edu/newscenter/stories/st050928.asp (accessed 28 Dec 2007).

[7]Yasnoff, W. A., O'Carroll, P. W., Koo, D., Linkins, R. W., and Kilborne, E., "Public Health Informatics: Improving and Transforming Public Health in the Information Age," *J. Public Health Manage. Pract.* **6**(6), 67–75 (2000).

[8]Centers for Disease Control and Prevention, *National Center for Public Health Informatics (NCPHI)*, http://www.cdc.gov/ncphi (accessed 20 Dec 2007).

[9]Brailer, D., Augustionos, N., Evans, L. M., and Karp, S., *Moving Toward Electronic Health Information Exchange: Interim Report on the Santa Barbara County Data Exchange, California HealthCare Foundation, July 2003*, http://www.chcf.org/documents/ihealth/SBCCDE InterimReport.pdf (accessed 20 Dec 2007).

[10]MedTech Publishing Company, *NHINWatch*, http://www.nhinwatch.com (accessed 18 Sept 2007).

[11]Klompas, M., Lazarus, R., Daniel, J., Haney, G. A., Campion, F. X., et al., "Electronic Medical Record Support for Public Health (ESP): Automated Detection and Reporting of Statutory Notifiable Diseases to Public Health Authorities," *Adv. Dis. Surveill.* **3**(3), 1–5 (2007).

[12]Simons, W. W., Mandl, K. D., and Kohane, I. S., "The PING Personally Controlled Electronic Medical Record System: Technical Architecture," *J. Am. Med. Inform. Assoc.* **12**(7), 47–54 (2005).

[13]University of Washington, *Center for Public Health Informatics*, http://www.cphi.washington.edu/projects (accessed 20 Dec 2007).

[14]Lehmann, H. P., Abbott, P. A., Roderer, N. K., Rothschild, A., Mandell, S., et al. (eds.), *Aspects of Electronic Health Record Systems*, Springer, New York, 2nd Ed. (2006).

[15]Blum, B. I., and Duncan, K., *A History of Medical Informatics*, ACM Press, New York (1990).

[16]Buckeridge, D. L., Burkom, H., Moore, A., Pavlin, J., Cutchis, P., et al., "Evaluation of Syndromic Surveillance Systems–Design of an Epidemic Simulation Model," *MMWR Morb. Mortal. Wkly. Rep.* **53**(Suppl.), 137–143 (2004).

[17]White House Office of the Press Secretary, Homeland Security Presidential Directive (HSPD), *HSPD-21: Public Health and Medical Preparedness*, 18 Oct 2007.

The Author

Joseph S. Lombardo has been employed by APL for the past 38 years performing research on various forms of surveillance, particularly sensors, signal coherence, background noise analysis, and data presentation. For the past 10 years, he has focused on developing and improving automated tools to enhance disease surveillance. He has led the development of the ESSENCE disease surveillance system, which is currently being used widely by the Department of Defense, the Department of Veterans Affairs, and several state and local health departments. Mr. Lombardo has degrees in engineering from the University of Illinois at Urbana–Champaign and from The Johns Hopkins University. He was the William S. Parsons Informatics Fellow with The Johns Hopkins University School of Medicine. His e-mail address is joe.lombardo@jhuapl.edu.

Joseph S. Lombardo