A Method for Generation and Distribution of Synthetic Medical Record Data for Evaluation of Disease-Monitoring Systems

Joseph S. Lombardo and Linda J. Moniz

There are several initiatives within the health care community to exchange patient medical records across health care facilities. These records hold the potential of providing public health officials with the ability to identify and monitor diseases posing a health risk to their communities. Tools are needed to exploit the wealth of data and information contained within these records. Synthetic medical record data are needed for development and evaluation of new techniques. A hybrid of real background disease levels with injected cases of disease has shown utility for testing systems. The generation of medical records in response to patient chief complaints is provided by the addition of a care delivery model. Once the model is fully developed, synthetic medical record data can be provided as a service, on the National Health Information Network, for developers and public health agencies to support evaluation and training.

INTRODUCTION

During the past 10 years, there has been increased use of information technology to support public health's mission for the monitoring of diseases that can have high mortality and morbidity. The motivation for enhanced surveillance tools has been early identification of a bioterrorist event, such as the widespread release of a biological warfare agent, or of a naturally occurring pathogen, such as a highly virulent form of influenza. Automated disease surveillance systems provide value to health monitors by collecting, archiving, processing, and displaying data containing indicators of the levels of infectious disease within the community. These tools monitor the number of counts of occurrences of a variety of health-seeking behaviors and notify users of abnormal trends in disease levels.

Until recently, only limited health-indicator data have been available to the public health community for surveillance purposes. Many health care facilities supply emergency department chief complaints to public health departments. This information is collected from

the patient or triage nurse resulting from the encounter with the patient. Other sources of data include diagnostic billing codes [per the International Classification of Disease Codes, Ninth Revision (ICD-9)], sales of over-the-counter (OTC) medications from pharmacy chains, 911 calls, school absentee data, and data captured during emergency medical service activities. These data sources lack the specificity needed to identify infectious disease events of concern to public health officials. Data from these sources also tend to be quite noisy. It is doubtful that a bioterrorist event of high concern to the public health community could be discerned in its earliest stages because of the background disease levels creating noise that may mask small signals in incidental data sources.

It is very difficult to anticipate how new outbreaks will begin and progress. This poses a significant challenge to the development and use of automated surveillance tools. Testing the performance of analytical tools is less than ideal when the characteristics of the signals being detected are not well defined. To provide some performance measures, researchers have attempted to evaluate automated systems either by using actual historic data[1] or by creating synthetic data based on a model of how an outbreak may progress.[2,3] Both of these approaches assume that future signals will be similar to those that have appeared in the past. Because there have been few large bioterrorist events, availability of historic data is limited. Current limitations on either historic or synthetic test data will eventually inhibit the development of new techniques.

Recent federal government initiatives[4,5] have resulted in recommendations for data to be supplied by health care organizations for biosurveillance purposes. In addition to the emergency department chief complaint data, these initiatives recommend transfer of medical record data collected from health care facilities. These data include microbiology laboratory and radiology requests and results as well as prescriptions for medications. A pseudonym-modified link is provided that ties the data to an unidentified patient. These data potentially provide support for diagnosing the patient's disease, affording the needed specificity for public health monitoring. With the availability of more types of data, new research into analytical tools is required to link the elements of the medical record data for each patient in a linear fashion that estimates illness and aggregates cases across the population. Additional visualization techniques are needed to expedite analysis for busy health departments monitoring diseases in their community. In addition, health departments need to develop exercises for training for a variety of public health risks. These new requirements all demand the availability of realistic health care data extracted from electronic medical records.

The objective of this article is to present a method both for creating linked medical record data reflecting outbreaks and for distributing these synthetic data to health departments and developers of new methods.

BACKGROUND

Actual Historical Data

A wealth of historical data showing outbreaks of various magnitudes for all high-risk diseases would be ideal for research, development, and evaluation of surveillance applications. However, in the past few decades there have not been many naturally occurring or electronically captured large outbreaks that could be used in the development and evaluation of modern surveillance tools.

Although there is little in the way of historical data reflecting outbreaks, there is ongoing effort to assess the effectiveness of analysis methods using the data that are available. An evaluation of analytical tools has been performed using military data in five cities in the United States.[1] A team of epidemiologists performed a review of historical data of ICD-9 to identify potential outbreaks of respiratory and gastrointestinal illness in those cities. The team identified several events of significant levels that should be detected by an automated system. The data were given to developers of detection algorithms in a blind test to determine the performance of their algorithms. Other examples of real outbreaks used for development are the 1979 accidental release of anthrax in Sverdlovsk[6] and the anthrax-containing letters in 2001.[7] Although these examples provide valuable insights into understanding how a large outbreak of anthrax may progress, neither provide normal background data on illnesses in the area along with the outbreak.

Historically, epidemiologists have captured data on the number of individuals who are ill and not on the initial encounters of those individuals seeking care based on symptoms. These encounters, however, contain valuable information about the background level of disease and population health-seeking behavior that can be expected. This information is essential in determining the severity and extent of unusual outbreaks of disease.

Outbreak Modeling

Researchers have traditionally developed models of diseases based on what limited data are available from occurrences of outbreaks. A historical paper by Sartwell[2] notes the similarity in the distribution of new cases of infectious disease. An early principle taught to students of public health is the progression of disease in populations by dividing individuals into groups who are susceptible, exposed, infected, and then recovered or dead.[8] Individuals move from one group to the next as the outbreak progresses. Many variations of models based on these disease states have been developed. An example is provided by Meltzer et al.,[9] where individuals

with smallpox are modeled as they move through the four major stages of disease and infect others.

Social modeling efforts have attempted to replicate the behavior of individuals in the population during an outbreak. Each individual is assigned a set of behaviors that can be used to create the health-seeking behavior data needed to understand how a disease can progress in a community.[10] The behaviors include both interactions with others who are susceptible, exposed, infected, or recovered, as well as individual health-seeking behaviors, such as when to stay home because of illness, when to self-medicate, or when to see a physician. Very sophisticated models could have multiple diseases present in the population at the same time, providing a realistic background for disease. However, this method has several disadvantages. When the population is large, the computation time of social models is great, requiring high-performance computing facilities. The behaviors of individuals must be represented statistically, making it difficult to focus on incorporating behavioral changes that may have a significant impact on the outcome. Although in theory it is possible, practically none of the existing social models can create a realistic background that incorporates all of the diseases normally found in a community.

Hybrid Approaches

An alternative approach has been to add cases representative of an outbreak to actual data that have been acquired for the purposes of surveillance. An example of this approach has been provided by Buckeridge et al.[11] The method includes using knowledge and actual data as inputs to a suite of models or processes that express the infection on a population. Figure 1 provides an illustration of the method.

The example provides a sequence of processes for infecting a population from an intentional release of anthrax on an urban population. The type of pathogen, the amount released, and the meteorological factor (that causes the material to spread over a geographical area) must be ascertained to execute the series of processes. A dispersion model calculates the distribution of the material over the area. Census records provide background information on the affected population, but knowledge of the population's behaviors is required to use an infection model. Usual outputs of the infection model include the number of individuals infected, their location and age, and the amount of infectious material received.

Disease progresses through various states that have different incubation periods, signs, and symptoms during the prodromal and fulminant stages. The disease states must be distributed on the exposed population using a disease model. All individuals experience the same symptoms; health-seeking response differs according to a behavior model. This model determines which portion of the population will elect to alter normal activities, self-medicate, seek help from a family physician, or go to an emergency or urgent-care facility. Individuals with signs and symptoms for a disease state following a specific health-seeking behavior are added to existing health-seeking data to create a signal indicative of an outbreak.

The advantages of the hybrid approach are (*i*) that the background data are real and (*ii*) that these data contain all of the characteristics that algorithms and disease monitors must consider, such as day of week, seasonality,

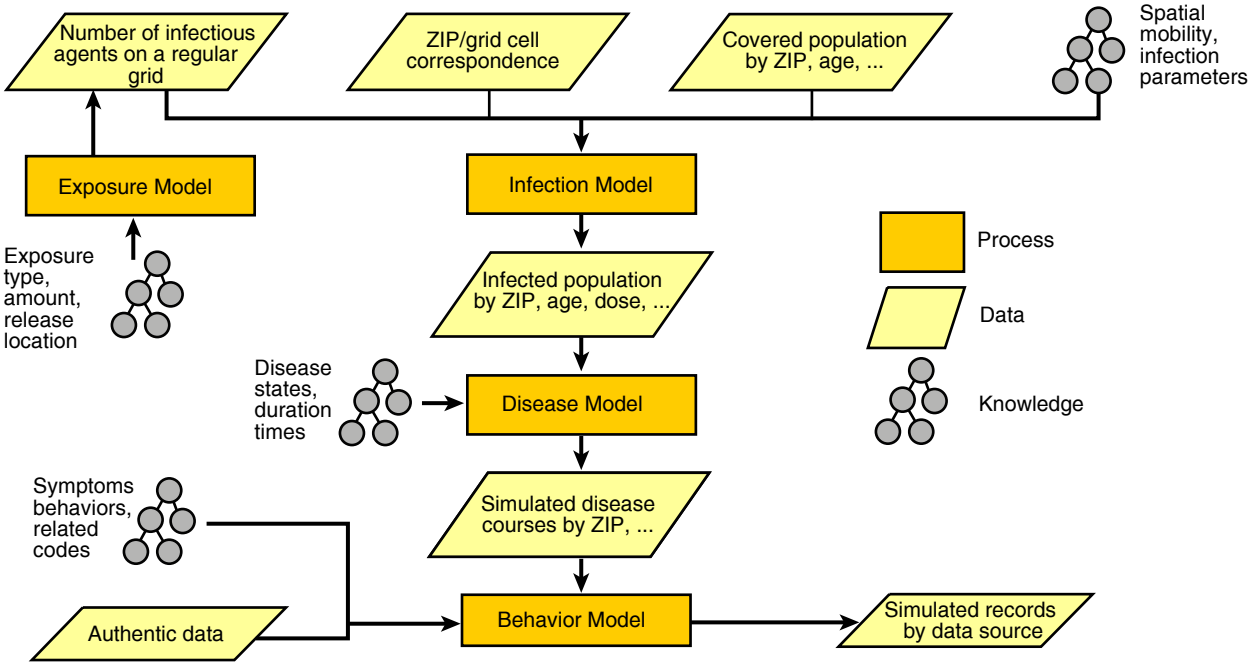

Figure 1. Hybrid method for adding infected cases in real data streams (anthrax example). (Adapted from Buckeridge et al.[11])

and local endemic background diseases. The real disease background allows the imposition of an infinite number of hypothetical outbreaks. The quantity, location, and type of infectious agent can be varied; the health-seeking response model would also be varied according to the outbreak agent. For example, the scenario of smallpox through the transportation network would require a different exposure, infection, disease, and behavior model than the one illustrated in Fig. 1 for an airborne release of anthrax. Figure 2 provides an example of how the processes could be reconfigured for a communicable disease such as smallpox or plague. Infected individuals become the primary data source for new cases of infection. Knowledge of their mobility around the community is needed to determine with which individuals of the susceptible population they will come in contact. Then continuous feedback is used to enhance the spread of disease within and outside of the community.

Modifications to Real Background Data

The Health Insurance Portability and Accountability Act (HIPAA) was enacted by the U.S. Congress in 1996. This law prohibits the use of identified health records for use in research, including the development of disease surveillance systems. To make these data useful, modifications can be made to the data to make them more difficult to identify individuals, while retaining the characteristics needed to test system functionality. The Realistic But Not Real (RBNR) method was developed by R. Picard (personal communication, R. Picard, Los Alamos National Laboratory) for that purpose. The method alters the actual age of the patient while keeping the age within an established bracket. Picard also alters the day of the encounter by exchanging it with another day plus or minus 2 weeks, but the same day of the week. Because this method does not maintain the number of patients with similar symptoms on specific days, one could argue that naturally occurring disease within the background could be lost as a result of the exchanging of patients across a sliding 4-week time window.

However, modification of a large set of data with the RBNR data is a good starting point for synthetic data. A thorough analysis of RBNR data from several hospitals across geographic regions can remove the need for specific and geographically differentiated care models. Using these local attributes for injected data ensures that any injected data follow the standard of care for a particular region.

Our innovative approach to synthetic data depends heavily on the analysis of the data. However, it begins with the hypothetical victims. The victims are chosen according to a demographic model that selects the segment of the population likely to be infected in the injection scenario. The ground truth health states of the victims are tracked according to the physiology of the injected disease. However, these ground truth states are *not* the injected cases as in previous studies; they are the *inputs* to a process that is defined by the analysis step.

The analysis, in effect, defines a function that takes the ground truth health states of the victims injected into the system as its inputs and uses the procedures and models in the data to produce electronic medical records as its outputs. These records mimic the actual medical records, including encounters, laboratory orders, and

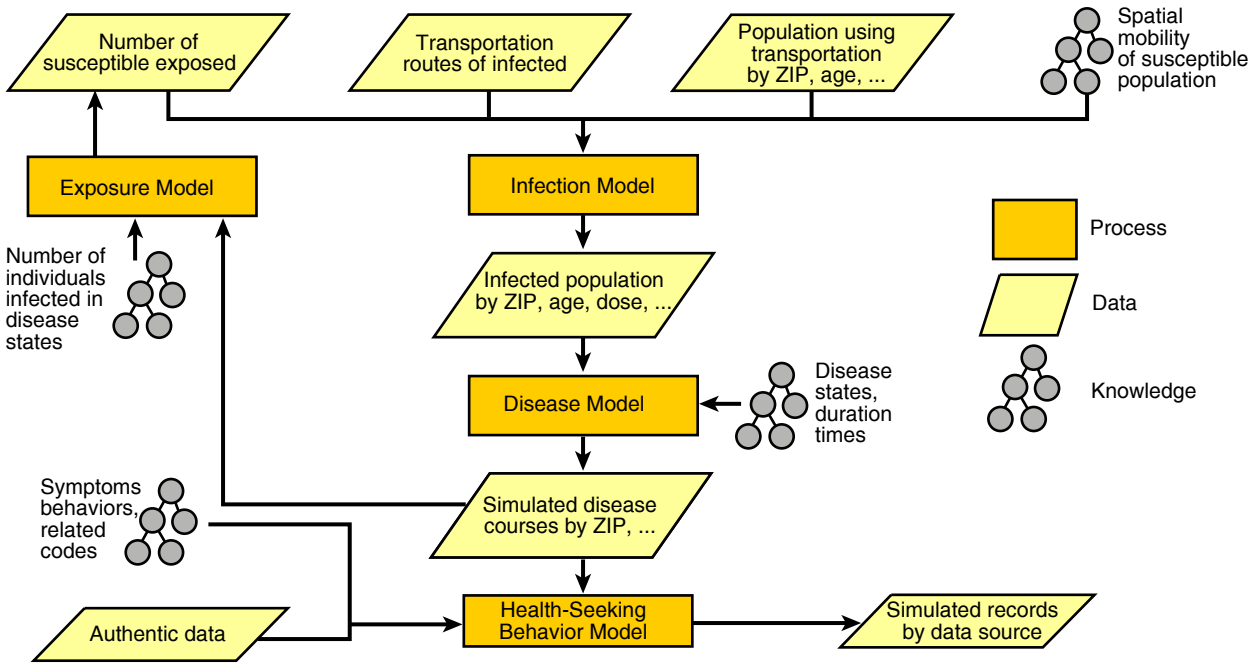

Figure 2. Communicable disease hybrid variant.

laboratory results. The process is designed to be seamless, with the addition of noise and deletion of occasional records at the level inherent in the data.

Analysis of large-scale RBNR data is a serious and intense undertaking. Although hospital protocols will not vary much, patient care can vary according to the type of hospital, insurance concerns, individual preferences of doctors, and patient input. However, the protocols and care models for well defined syndromes should include a core care model with a wide range of variances. The variances can be modeled and used in the synthetic medical records to provide realistic injects that match the existing protocols. For example, if an injection of anthrax is used, the care models for fever and respiratory syndromes could be mirrored, with specific chief complaints that include only those symptoms expected for anthrax. Laboratory tests and radiology requests would be ordered in the same fashion and on the same timeline as similar chief complaints in the background data, with additional tests as symptoms present and urgency varying with the severity of the symptoms.

Additional Data to Support Public Health Surveillance

In 2005, the U.S. Department of Health and Human Services (DHHS) chartered the American Health Information Community (AHIC) to recommend a series of information technology standards for the exchange of health care data.[4] Included are recommendations for providing health care data to public health departments or agencies to improve surveillance for infectious diseases that could cause high mortality and morbidity. Another DHHS initiative is the Centers for Disease Control and Prevention's (CDC) BioSense Program. This program has been acquiring hospital care delivery records in near real time for processing to obtain situational awareness of the health status on major population regions across the country.[5] These two initiatives are making it possible to acquire not only emergency department chief complaint data but also virtually every source of health-related data available on the hospital's network.

Data elements available include the status of the hospital's ability to accept and care for new patients as well as a pseudonym-modified patient data linker as an identifier on medical records. The data linker makes it possible to follow the course of care delivery for each patient. The care delivery data elements include the chief complaint, triage notes, vital signs, medications ordered, disposition, physician clinic notes, diagnosis, and procedure codes. In addition, diagnostic supporting data such as microbiology laboratory orders, sample collection status, and results; radiology orders, status, and tests; as well as impressions are also collected. Providing these additional data elements for public health surveillance promises to improve disease reporting, enhancing the specificity of early automated alerting for infectious diseases that pose a health risk as well as overall status of the course of an outbreak.

Although these data offer promise of improved surveillance, they also pose significant challenges for automated collection and analysis tools to exploit their contents. To develop and evaluate tools that take advantage of the full medical record, data are needed to test a wide variety of scenarios. These data must incorporate signals in the data elements of the medical record consistent with the expected treatment protocols used in any of the care facilities that provide data for surveillance.

Significant challenges exist in developing automated tools to exploit the full medical record because individuals with the same illness may present over an extended time period and have a variety of different tests to rule out other illnesses. Because protocols vary widely among treatment facilities, what may be normal for a specific health care facility may be abnormal for another. Automated analysis tools must take into account the differences in care delivery protocols among the various facilities providing data to support public health surveillance.

EMERGE: Electronic MEdical Records GEnerator

The ideal synthetic data generator for creating electronic medical records should have several attributes. It should be able to use simulated patient signs and symptoms to create data for a variety of different infectious diseases at each stage of the disease. The data generator needs to take into account the variability in the relationship among the data elements it creates, accounting for "rule-outs" (tests that are ordered to rule out specific diagnoses), differences in treatment protocols of health care facilities, as well as the preferences of its health care providers. The data generator must also create realistic data errors and drop-outs (missing data) representative of errors in real data.

Figure 3 represents a modification to the hybrid technique, which can be used to create the medical record data. The care delivery model attempts to replicate the delivery of care by requesting services based on chief complaints and symptoms of patients being seen at the facility. The development of this model is known as the EMERGE project.

EMERGE would eventually have the capability to synthesize all of the data elements within the medical record, although not all elements would be available for every patient encounter. These data elements would be linked to the patient by the pseudonym-modified patient reference data element. Details of institutional care delivery protocols are contained within historical data captured for surveillance. Initial versions of EMERGE could be created for large providers of health care. Examples would be the Veterans Health Administration, Department of Defense, or large health maintenance organizations (HMOs).

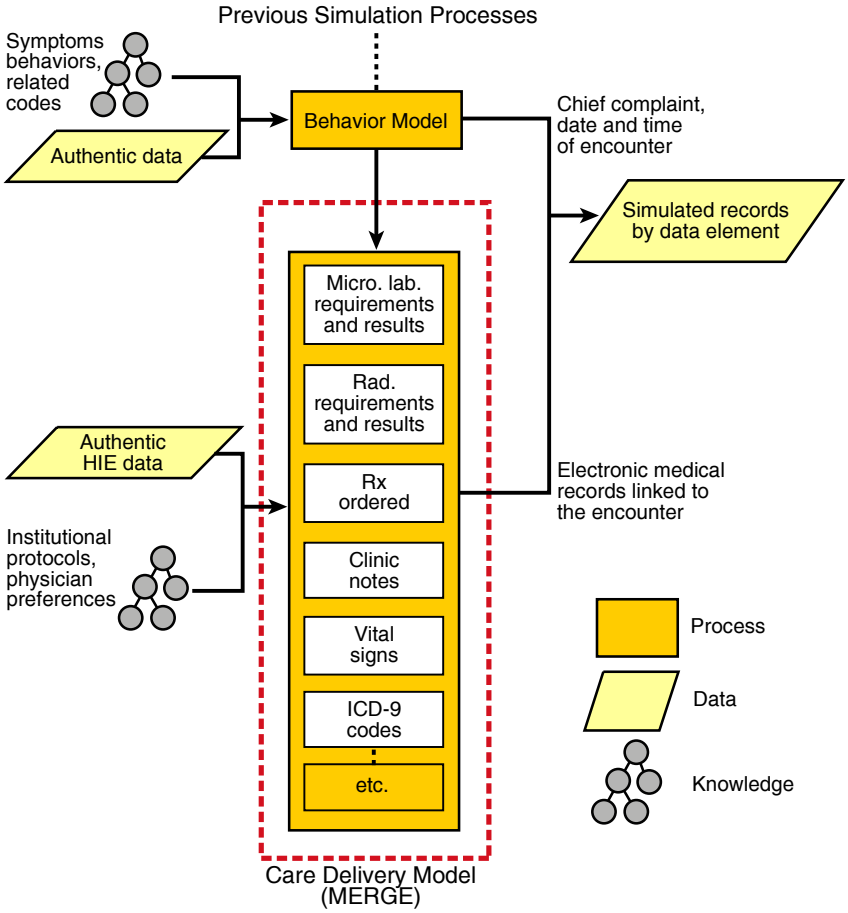

Figure 3. Modification of the hybrid data synthesizer to include EMERGE.

Eventually the EMERGE model could contain the unique care delivery protocols for each facility providing data for surveillance. EMERGE data outputs could be modified and augmented by the standard of care as it is manifested in a particular facility. In this way, any synthetic data for any kind of outbreak can be tuned both to the facility that provides the care and to the range of symptoms that are associated with the disease. Alternatively, the data could be used to determine which protocols provide the optimal care. This would assist with continuously measuring and refining the delivery of care.

Enabling Technologies, Health Information Exchanges, and the National Health Information Network

Recent information exchange initiatives at the local level have resulted in the creation of Regional Health Information Organizations (RHIOs) or Health Information Exchanges (HIEs), which have made health records available for improved delivery.[12] An excellent example is the HIE assembled by the Regenstrief Institute of the University of Indiana.[13] Hospitals in and around Indianapolis send data to Regenstrief for distribution among the hospitals in order to improve care delivery. This data feed is also provided to state and local health departments for surveillance. These data-capture initiatives have facilitated the collection of medical records not only for early identification of infectious disease events but also for the continuous flow of data during outbreaks. This flow of data helps health departments monitor the execution of their containment plans and creates public health situational awareness. Figure 4 provides an illustration of the opportunities available to health departments resulting from initiatives in data capture and exchange.

On a larger scale, the federal government is creating a National Health Information Network (NHIN) to enhance the flow of data and information across the country. Distributed networks will make it possible to automatically exchange surveillance data and information among federal partners and state and local health departments. Eventually, data and information obtained outside of the region will be able to be ingested and analyzed along with local data. Figure 5 represents a surveillance model where analysis is performed at the CDC and at many state and local health departments using data and information from several sources available on local HIEs connected in the future to the NHIN.

It is ideal to analyze data locally where events and conditions that influence health and health reporting are known. Privacy concerns may prohibit sharing of raw data outside of local health departments; in this case, only locally processed information can be shared. Automated surveillance systems have the potential for using information from outside of their region to complement local surveillance and create better situational awareness. The value of automatic processing of information from outside of the region will eventually need to be evaluated, requiring both data and information from inside and outside of the region to be available or created. The NHIN provides the connectivity to share data and information at all levels of aggregation to protect privacy and exploit local knowledge of the environment and the population.

EMERGE as a Web Service

The use of automated disease surveillance tools will become more sophisticated and valuable to the knowledgeable user with advances in technology and the

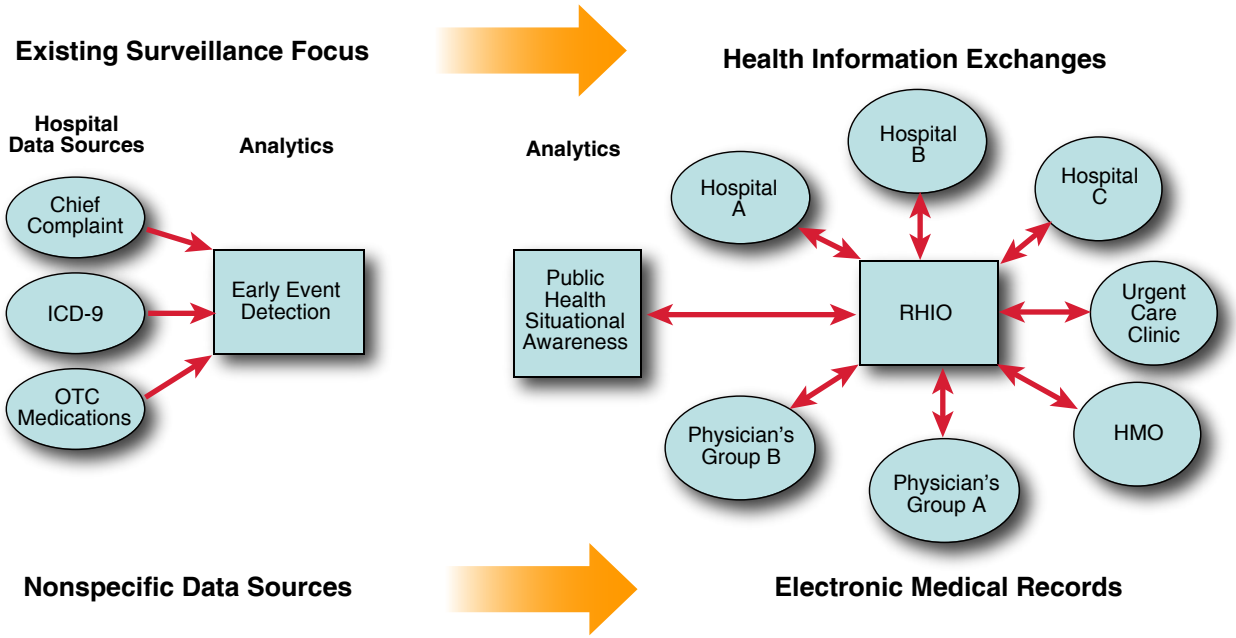

Figure 4. The changing data-provisioning environment for public health surveillance.

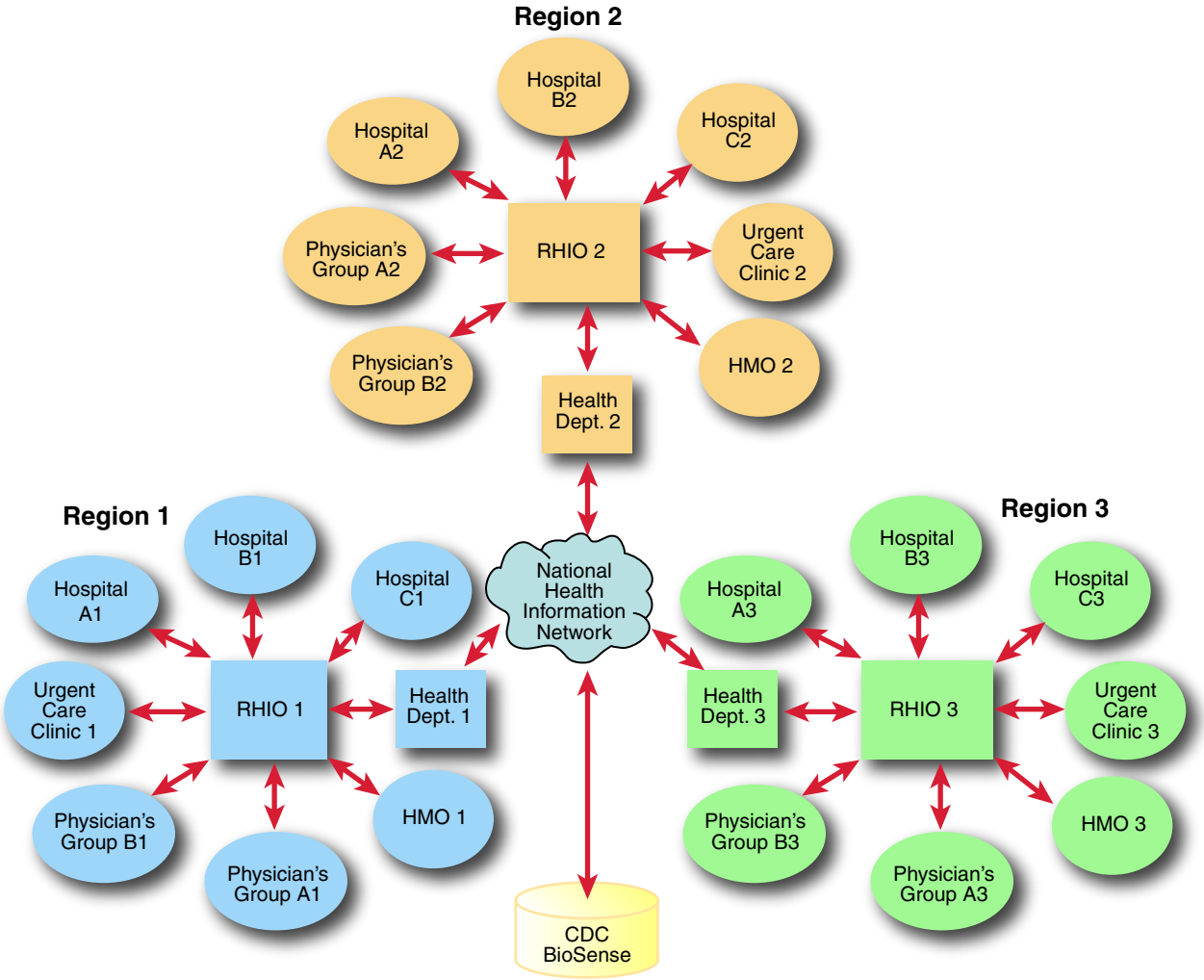

Figure 5. Data capture via the NHIN requires expanded analytics to be used in automated surveillance systems.

availability of data sources to public health. Data will be needed to perform development, evaluation, training, and exercises at the local, regional, and national levels. Given the wide variety of data/information needed to perform inter-regional exercises and the cost of creating synthetic data for the large number of disease and outbreak scenarios that pose a risk to public health, it makes sense to create a vision for a data generator that can provide the public health community with this commodity.

The proposed approach is to provide synthetic data as a service over the NHIN using local data obtained for surveillance for background levels and to inject data for patients (as described above) with signs and symptoms indicative of an emerging health event on top of the background. EMERGE would supply the medical records created for the patients whose data are to be injected. Figure 6 illustrates the creation and distribution process for the medical records.

A request for data would be made by one or more authorized public health organizations for the purpose of testing system performance, training, or conducting exercises. Creating the data would consist of obtaining current or historical background data and adding the event to the background. The specific background data would come from the jurisdiction requesting the

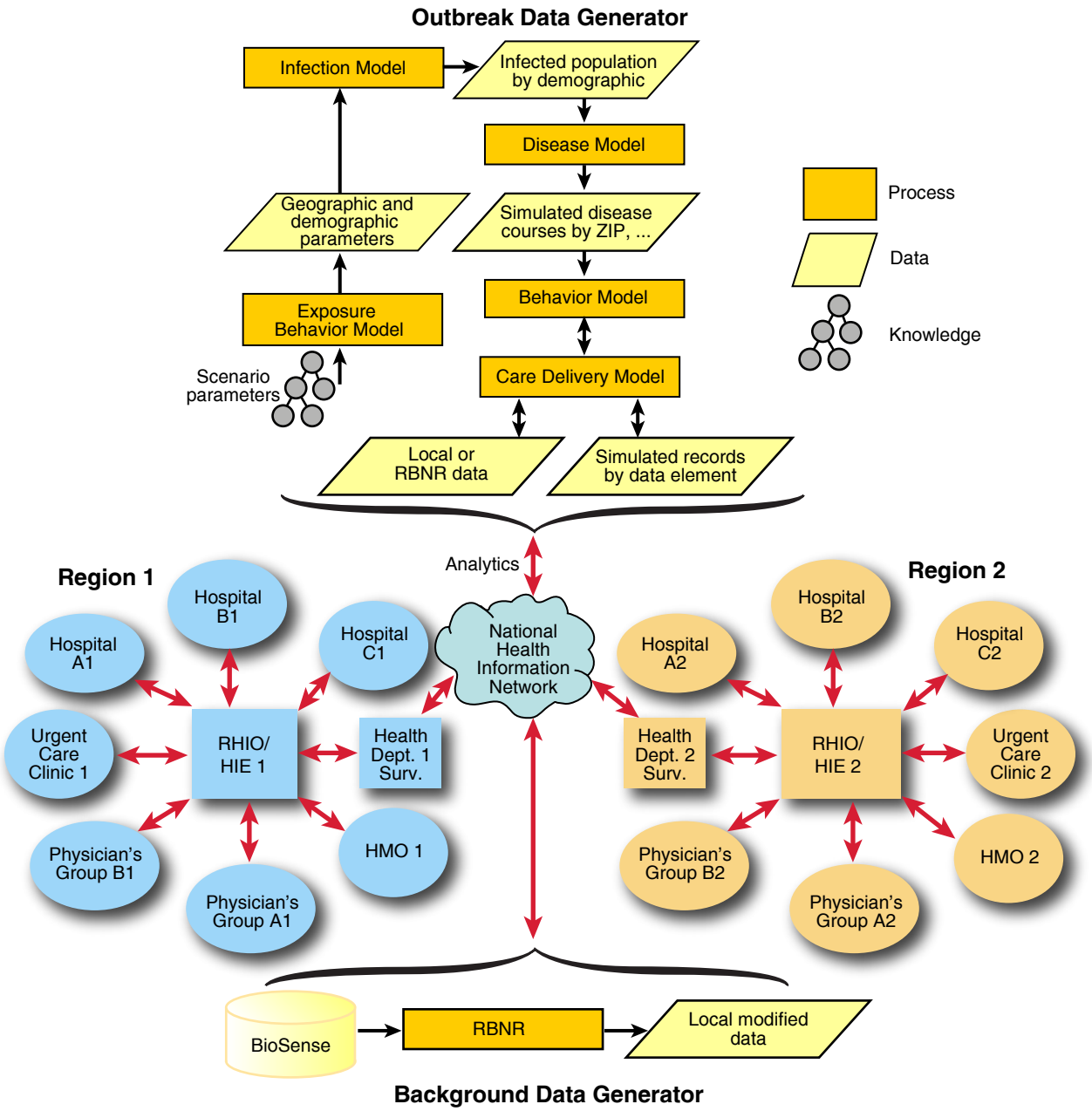

Figure 6. Model for providing synthetic data as a service on the NHIN.

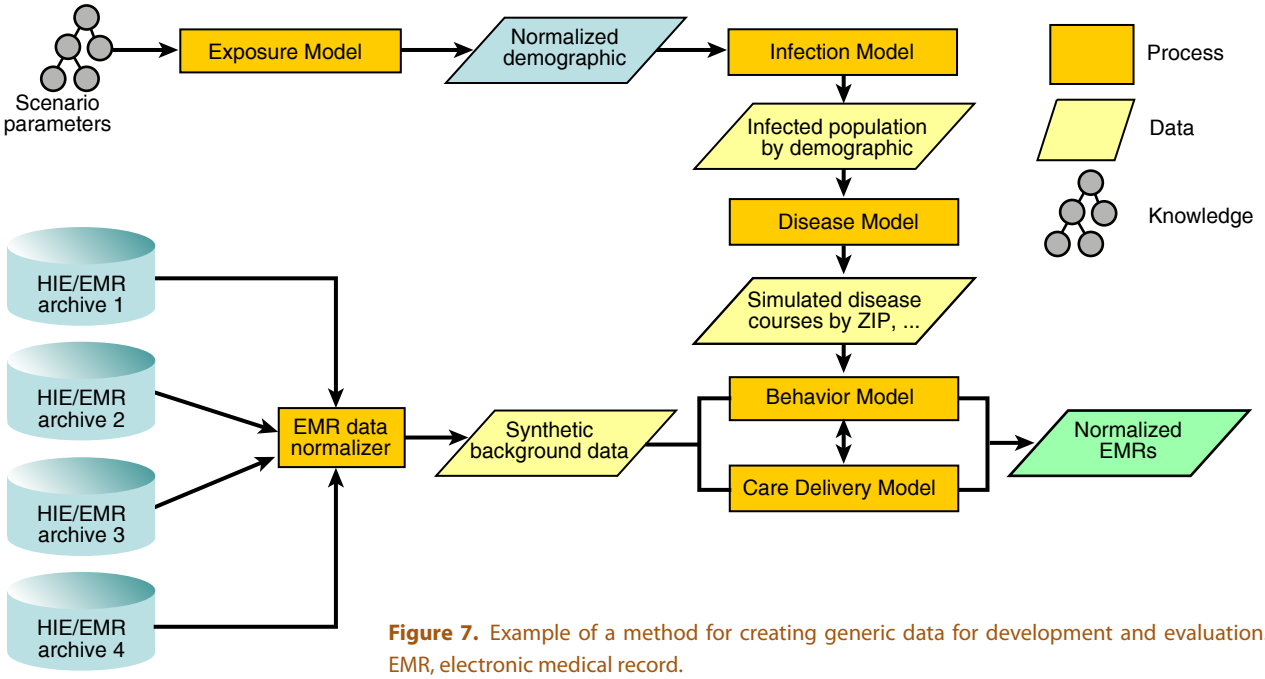

Figure 7. Example of a method for creating generic data for development and evaluation. EMR, electronic medical record.

synthetic data. The background data used could be current data or historical data covering approximately the same time of year. Background data may be held in a federal archive such as BioSense or locally within the RHIO or HIE. For the case where data is requested by developers, background data could be modified with an RBNR algorithm or other processes to retain the characteristics of local data and the relationships that exist among the data elements of the medical record.

The synthetic data generator could also be used to create standard datasets for the development of surveillance algorithms and interfaces using medical record data. These datasets could include normalized background data obtained from a variety of locations modified by an enhanced RBNR process. This background could then be used with a variety of different scenarios and diseases to form an evaluation dataset to be used to determine the performance of different surveillance tools. Figure 7 illustrates a series of processing steps that could create generic data to be used for development and evaluation of surveillance tools.

Several datasets could be created for different scenarios and distributed openly for the development of tools. Other datasets would be created strictly for a blind evaluation.

CONCLUSIONS

AHIC's recommendations for data to support biosurveillance, the CDC BioSense Program, as well as the creation of RHIOs/HIEs are providing public health officials with the resources to improve the specificity of early alerting for outbreaks. To fully exploit these resources, the algorithms and methods require additional development, testing, and evaluation. The development, testing, and evaluation process requires datasets containing signals in data contained within the electronic medical record. Existing hybrid data-generation methods can be expanded to include models for the delivery of care, which would add signals onto the elements of the medical record. Once the processes have been assembled, synthetic data could be provided as a service over the NHIN or as generic development and evaluation data for the development of algorithms or applications.

ACKNOWLEDGMENTS: This article was supported by Grant P01 CD000270-01 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]Siegrist, D., and Pavlin, J., "Bio-ALIRT Biosurveillance Detection Algorithm Evaluation," *MMWR Morb. Mortal. Wkly. Rep.* **53**(Suppl.), 152–156 (2004).

[2]Sartwell, P. E., "The Distribution of Incubation Periods of Infectious Diseases," *Am. J. Epidemiol.* **51**, 310–318 (1950).

[3]Daley, D. J., and Gani, J. M., *Epidemic Modeling, an Introduction (Cambridge Studies in Mathematical Biology, No. 14)*, Cambridge University Press, New York (1999).

[4]American National Standards Institute Standards Healthcare Information Technology Standards Panel website, http://www.ansi.org/standards_activities/standards_boards_panels/hisb/hitsp.aspx?menuid=3 (accessed 18 Sept 2007).

[5]Centers for Disease Control and Prevention, *BioSense: For Public Health Departments*, http://www.cdc.gov/biosense/publichealth.htm (accessed 18 Sept 2007).

[6]Meselson, M., Guillemin, J., Hugh-Jones, M., Langmuir, A., Popova, I., Shelokov, A., et al., "The Sverdlovsk Anthrax Outbreak of 1979," *Science* **266**, 1202–1208 (1994).

[7]Jernigan, J. A., Stephens, D. S., Ashford, D. A., Omenaca, C., Topiel, M. S., et al., "Bioterrorism-Related Inhalational Anthrax: The First 10 Cases Reported in the United States," *Emerg. Infect. Dis.* **7**(6), 933–944 (2001).
[8]Gordis, L., *Epidemiology*, 2nd Ed., W. B. Saunders Company, Philadelphia, PA (2000).
[9]Meltzer, M. I., Damon, I., LeDuc, J. W., and Millar, J. D., "Modeling Potential Responses to Smallpox as a Bioterrorist Weapon," *Emerg. Infect. Dis.* **7**(6), 959–969 (2001).
[10]Barrett, C. L., Eubank, S. G., and Smith, J. P., "If Smallpox Strikes Portland…," *Sci. Am.* **292**(3), 42–49 (March 2005).
[11]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).
[12]NHINWatch website, http://www.nhinwatch.com (accessed 18 Sept 2007).
[13]"Indiana Health Information Exchange, "Regenstrief Institute Part of Consortia to Develop Nationwide Health Information Network; Indiana's Participation Reinforces Status as National Leader in Healthcare Information Technology," *Business Wire* (16 Nov 2005), http://findarticles.com/p/articles/mi_m0EIN/is_2005_Nov_16/ai_n15800673.

The Authors

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. **Linda J. Moniz** has been employed by APL since August 2007. She received her B.A. from Wellesley College and her M.S. from the University of North Carolina (Chapel Hill), both in mathematics. She worked in industry as an applied mathematician for 10 years, concentrating on data fusion and analysis and computational tools. She received her doctorate in mathematics at the University of Maryland in 2001. Her postdoctoral research, at the Naval Research Laboratory and the United States Geological Survey Patuxent Wildlife Research Center, centered on applications of dynamical systems, primarily data analysis in chaotic systems, structural health monitoring, embedding of experimental data, synchronization of chaotic systems, and monitoring of ecological systems. For further information on the work reported here, contact Mr. Lombardo. His e-mail address is joe.lombardo@jhuapl.edu.

Joseph S. Lombardo

Linda J. Moniz