Tradeoffs Driving Policy and Research Decisions in Biosurveillance

Howard S. Burkom, Wayne A. Loschen, Zaruhi R. Mnatsakanyan, and Joseph S. Lombardo

In view of concerns over bioterrorism, pandemic influenza, and other public health threats, development of advanced surveillance systems to corroborate and supplement physician sentinel surveillance is a research imperative. Objectives of this article are to describe interrelated decisions underlying the design of a public health surveillance system and to show these decisions' effects on data acquisition and transfer, on analysis methods, and on visualization tools. Some of these decisions are dictated by data limitations, others by goals and resources of the monitoring organization. Most such decisions involve three characteristic tradeoffs: the extent of monitoring for exceptional versus customary health threats, the level of data aggregation for monitoring, and the degree of automation to be used. A fundamental motivation is to extract outbreak information from background noise to empower epidemiologists monitoring public health on a day-to-day basis. This article discusses each of these tradeoffs and illustrates them with three detailed examples.

INTRODUCTION

The purpose of this effort is to show how the goals and capabilities of health-monitoring institutions can shape the selection, design, and usage of tools for automated disease surveillance systems.

Automated surveillance systems to enhance public health monitoring capability have become a priority for both military and civilian institutions since the late 1990s. The biosurveillance program at APL has played a pioneering role in the evolution of these systems. After a proof of concept with several civilian data sources in 1999, this program joined with the military-based Department of Defense Global Emerging Infection System group at the Walter Reed Army Institute for Research to help spawn the multicenter BioALIRT program under the Defense Advanced Research Project Agency.[1] These early efforts produced the Electronic Surveillance System for Notification of Community-based Epidemics (ESSENCE), and versions of ESSENCE

are currently used in both civilian and military monitoring institutions at various jurisdictional levels. Since BioALIRT ended in 2003, the APL program has continued ESSENCE research and development under the U.S. Departments of Defense, Homeland Security, and Health and Human Services, and is currently a Center of Excellence for Bioinformatics sponsored by the Centers for Disease Control and Prevention (CDC).

Biosurveillance systems make opportunistic use of data sources that generally were not designed for prospective surveillance and apply analytic methods adapted from many other disciplines. Their successful application requires the cooperation of epidemiologists, software architects and developers, and analysts such as statisticians. These systems have drawn criticism because of the lack of clear demonstrations of value added to traditional health monitoring.[2]

For example, they have not shown the ability to detect community-level outbreaks. Part of the problem is that few datasets are available with known outbreak signals. It also is true that expectations, objectives, resources, and support for this developing capability vary widely among monitoring institutions. Thus, off-the-shelf solutions do not apply, but local health departments generally do not have the resources to develop customized systems. For advanced biosurveillance systems to achieve widespread acceptance, their roles must be clarified for each application, and they should be designed to fulfill these roles. Clarification of these roles is the subject of this paper.

Considerations for Surveillance System Design

Numerous decisions are faced by a public health institution: What types of outbreak are the target events for detection, and on what scale is the detection needed? What information sources should be used, and how should the data from these sources be filtered and analyzed? How should the results be visualized by users? Some of these decisions are dictated by available data limitations, others by objectives and resources of the organization doing the surveillance. Most such decisions involve three characteristic tradeoffs: the extent of monitoring for exceptional versus customary health threats, the level of aggregation of the monitoring, and the degree of automation to be used.

The first tradeoff results from heightened concern for bioterrorism and pandemics versus ongoing threats that involve endemic disease events such as seasonal outbreaks. A system focused on bioterrorist attacks is scenario-based, concerned with unusual diagnoses or patient distributions, and likely to include attack hypothesis testing and tracking tools. A system at the other end of this continuum has broader syndrome groupings and is more concerned with general anomalous levels at manageable alert rates, where manageability depends on the investigation resources available to the monitoring institution.

Major aggregation tradeoffs are temporal, spatial, and syndromic.[3] Bioterrorism fears have shortened the time scale of health monitoring from monthly or weekly to near real time. The spatial scale of monitoring is a function of the spatial resolution of data recorded and of the monitoring institution's purview and capacity to collect, analyze, and investigate localized outbreaks. Syndromic aggregation involves a decision for each type of information: Is a fine or a coarse filtering of data records more likely to enhance an outbreak signal relative to background noise?

Automation tradeoffs involve the use of data processing to collect information, analyze it for anomalies, and make decisions based on this analysis to investigate or respond to a perceived public health threat. The first of these uses has widespread acceptance, whereas for analysis and response, the degree of automation is a subject of ongoing controversy and research. To what degree can human judgment in alerting/response decisions be automated? What should be the depth and frequency of human inspection and adjustment? Should health indicator data be inspected and analyzed more often during elevated bioterrorism threat conditions? All of these decisions affect monitoring tools and practices as well as funding for related research.

A widely cited article[4] by a CDC Working Group emphasized the importance of understanding these decisions in any surveillance system evaluation as follows:

> The description of purpose should include the indications for implementing the system; whether the system is designed for short-term, high-risk situations or long-term, continuous use; the context in which the system operates (whether it stands alone or augments data from other surveillance systems); what type of outbreaks the system is intended to detect; and what secondary functional value is desired. Designers of the system should specify the desired sensitivity and specificity of the system and whether it is intended to capture small or large events.

The current paper asserts that this information should drive the design of any such system, not just its evaluation, or else the chances of success are haphazard. Key measures cited by those authors, such as sensitivity and timeliness, also are scope- and goal-dependent, or in their words:

> Measuring the validity of a system for outbreak detection requires an operational definition of an outbreak. . . . Operationally, an outbreak is defined by the affected public health jurisdiction when the occurrence of a condition has changed sufficiently to warrant public health attention.

Examples given below elaborate on and quantify these dependences.

For the organization of the rest of the paper, the tradeoffs based on surveillance objectives, on data aggregation, and on the degree of automation are explained in turn, and then three examples are presented to illustrate them and their interrelationship.

DETECTION SYSTEM OBJECTIVES AND RELATED TRADEOFFS

Concerns over bioterrorist threats have driven much of the funding and early development of modern biosurveillance systems. Such concerns lead to a scenario-based approach in which detection efforts are focused on the most plausible perceived attack scenarios. Such a scenario is a deliberate aerosol release of *Bacillus anthracis* spores. However, systems based on such specific scenarios have very limited utility because (*i*) the scenario may never occur, and (*ii*) the scenario must be highly specific to avoid a false-alarm rate that would render the system useless. An actual attack scenario may not adequately reflect the modeled one. In the absence of detailed data on the extent of disease resulting from large-scale bioterrorist attacks, the ability to validate such detectors is limited. For effective use of a public health detection system in a crisis, human monitors must be accustomed to operating the system routinely and interpreting and responding to its outputs. Therefore, the concept of *dual use* for both customary and exceptional threats has become popular, although this concept poses technical challenges discussed below.

In the extension of automated surveillance systems to general public health problems, most emphasis has been on infectious disease, such as influenza, as opposed to chronic disease, such as cancer or diabetes. This emphasis results from both bioterrorism and pandemic concerns. For analysis purposes, this emphasis calls for a focus on transient event detection rather than on subtle but lasting data effects. Thus, the small mean shift is not the primary signature of interest for infectious disease surveillance, although detection of such shifts is a classic problem for the statistical process control (SPC) community and would be more relevant in a chronic disease application, such as detecting a persistent increase in cancer risk caused by ongoing industrial pollution. Moreover, in view of the terrorist scenarios, system development has focused on diseases with short-to-moderate incubation periods (i.e., for which symptoms appear within days or at most a couple of weeks of infection), such as influenza or smallpox, as opposed to HIV/AIDS, for which symptoms may not appear for months or years.[5]

Objectives Related to Situational Awareness and Dual-Use Considerations

Most early implementations of biosurveillance systems focused on the capability for detecting the leading edge of an outbreak rather than on threat characterization and tracking,[6] but experience has changed this focus. For example, at an interactive roundtable discussion convened by the U.S. Medicine Institute for Health Studies,[7] there was strong consensus that syndromic surveillance has important uses that do not include early detection. In fact, although early outbreak detections resulting from automated systems do occur, they are rare. Even participants who had found them did not feel that early detection is a sufficient or even a principal justification for these systems. The day-to-day benefit is *situational awareness*, a concept whose dimensions are still unfolding. The following dual or combined uses have been reported:

- Monitoring for new strains of disease. Pandemic influenza, possibly resulting from a human-communicable mutated form of the H5N1 avian flu virus, is of widespread concern; biosurveillance systems have helped identify new strains of lesser seasonal viral infections and of norovirus.[8]
- Seasonal influenza surveillance and tracking. The significant annual influenza morbidity and mortality drive many health departments to use their systems to look for indications of the beginning and the potential local effects of flu season.
- Combination with environmental data to monitor for health problems related to extreme heat, poor air quality, allergy season, or contamination of the food or drinking water supply.
- Disaster management. Several health departments have used automated systems to track societal effects of displaced disaster victims, including those of Hurricane Katrina in 2005.

These applications will develop further as data quality and availability improve. Additional likely uses and enhancements are as follows:

- Combination with individual-based electronic medical records for more specific surveillance. See the article on information fusion and the electronic medical record by Mnatsakanyan and Lombardo elsewhere in this issue.
- Corroboration of warnings from biosensor data. Plans are underway to use biosurveillance systems to corroborate biosensor data indications of the presence of pathogens in indoor ventilation systems or outdoor atmosphere.
- Monitoring of traditionally hospital-acquired infections such as methycillin-resistant *Staphylococcus aureus* that have recently become community-acquired threats, as in generalizations of Morton et al.[9]

General Public Health Anomaly Detection

At the simpler, more general end of the spectrum of surveillance objectives is monitoring for anomalies by using one-sided statistical significance tests for unusually high counts or concentrations of data, whether they be diagnoses, product sales, or syndromic visits. Most health-surveillance systems operational as of 2007 are not scenario-based beyond the syndromic filtering of data records, so that specific signal shapes or anomaly

combinations across data types are not routinely monitored. Commonly used alerting tools are adaptive versions of control charts to signal statistical time series anomalies, typically using "stoplight" coding with red for a highly significant anomaly (e.g., for a *P* value of <0.01) and yellow for lesser significance (e.g., <0.05). More elaborate systems also apply scan statistics or similar methodology to seek localized anomalous disease clusters.[10] These systems do not purport to distinguish seasonal epidemics from unusual public health threats; both are considered part of the detectable signal. To the human monitor is left the task of fusing more syndromic data with laboratory and radiology test results or with other evidence to form and track outbreak hypotheses. The use of these general methods is widespread because of their conceptual simplicity and because they can be implemented without substantial data analysis in institutions that cannot afford extensive development and testing. There is a natural hesitation by local health departments to spend scarce resources on sophisticated published methods that have not been proven effective on local data and that may require expertise to maintain and interpret.

Scenario-Based Detection

At the other end of the surveillance methodology spectrum are algorithms designed to detect specific threats such as bioterrorist aerosol attacks or pandemic influenza. These methods are based on presumed outbreak characteristics and data effects. Research studies have undertaken this approach, sometimes based on considerable underlying research, seeking validation from medical and demographical expert knowledge, historical data analysis, inference based on characteristic data patterns during seasonal outbreaks, and simulations. Scenario-based methods employ assumptions specific to the pathogen causing the outbreak, route of infection, affected population, outbreak symptomatology, health-seeking behaviors of that population, effects on data sources, and delays between these effects on different sources. Some such methods use disease agent-based models in which individual people and their behaviors are represented as software objects[11,12]; these models are expensive, highly dependent on the population of interest, and difficult to validate. A second approach is the application of data-fusion techniques to population-level data. A Bayesian Network (BN) application under development at APL for this purpose is presented in the article by Mnatsakanyan and Lombardo elsewhere in this issue. This approach focuses on unusual distributions of diagnoses or patient ages as well as seasonal and other patterns, all derived from a combination of medical knowledge and data analysis.

Note that unlike general anomaly detection methods, both of these approaches must distinguish usual, seasonal outbreaks from the rare events of interest in the scenario. Any decision support aid also faces the challenges of validation and user acceptance.

Approaches to Combining General and Scenario-Based Surveillance

Most public health systems currently do not use scenario-based analysis, but many departments do use combinations of indicators to address the tradeoff between specific threat types and general surveillance. Although they cannot monitor all possible threat scenarios or potentially interesting subsyndromes, some do look for certain unusual chief complaints or diagnosis codes, such as rare codes corresponding to weaponized diseases such as anthrax or smallpox, while watching the general alerts from syndromic surveillance systems. Thus, the tradeoff problem is addressed with limited, parallel surveillance efforts. Most local health departments cannot support scenario-based analysis routinely but recognize the need to shift effectively to heightened and detailed surveillance measures if a specific threat arises. The BN approach pursued at APL is intended to provide a surveillance umbrella that can monitor for both general and specific anticipated public health problems. Example 2 below gives additional details.

TRADEOFFS RELATED TO DATA AGGREGATION

Having chosen their information sources, health-surveillance system designers must make numerous data-aggregation decisions. Key decision categories are how to group and filter data records, how to aggregate in space and time, and how to manage other covariates such as age and gender.[3] A thematic tradeoff is between (*i*) expanding the collection window size to increase data structure for background modeling and (*ii*) keeping the window size small to avoid masking a potential outbreak signal with the additional counts. For example, the left half of Fig. 1 gives plots of the daily respiratory visit counts for slightly more than 3 years of data on three spatial scales: the entire state, a large treatment facility (here a median respiratory-diagnosis visit count exceeding 100 per day), and a small facility (median visit count ~10 per day). The right half of Fig. 1 shows these counts restricted to a single flu season to clarify local features. At the statewide level, annual cycles and day-of-week effects in the time series are clearly visible and amenable to regression modeling, but these features weaken as the spatial scale decreases, and percentage errors in data modeling grow rapidly.[13] At the ZIP-code level of aggregation, the features cannot be seen at all. Thus, the outbreak signal and the background noise are strongly affected by decisions about how to aggregate the surveillance data. To see a seasonal pattern in the background or a lognormal shape in the signal, one may need to look at week-long counts of county-level data, whereas community counts or 4-hour time intervals might hide such

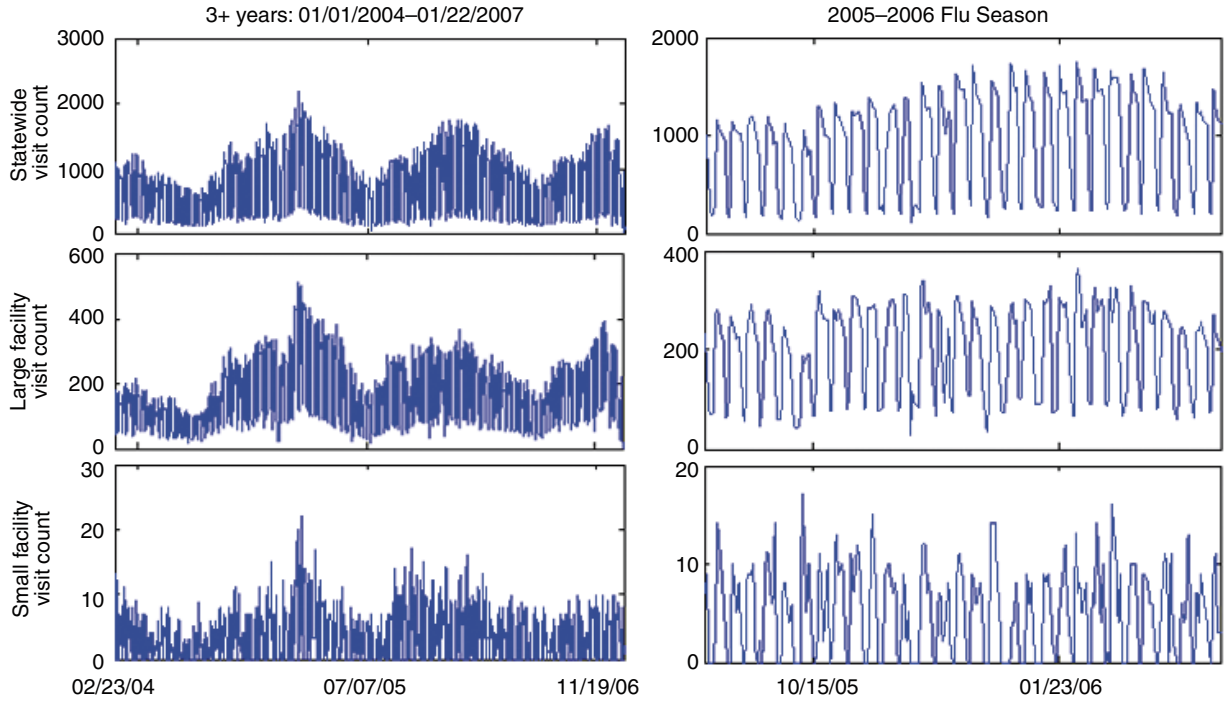


Figure 1. Outpatient clinic daily visit counts for respiratory syndrome for varying spatial aggregation and temporal scale.

features but make small outbreaks visible. These decisions are necessary in the classification of data records according to syndrome grouping, in time, and in space. Finally, we discuss aggregation by syndromic grouping of data records based on physician categorization of illness types.

Grouping and Filtering of Data Records by Syndrome

Limiting a syndrome group to diagnoses closely related to a disease of interest should improve alerting specificity but will likely yield a sparse, unstructured data background, and many such groups may be needed for general public health surveillance. A key feature of the syndromic approach is to monitor larger but noisier syndrome groups for a more general surveillance capability and a reasonable number of data streams to analyze. (Analyzing time series for counts of groups of records representing every possible illness would be prohibitive.) This approach also seeks power to detect the unknown threat; from the viewpoint of bioterrorism monitoring, a genetically engineered pathogen could have a novel presentation to the immune system and thus result in unusual symptom combinations.

Figure 2 illustrates the tradeoff in syndrome grouping. At the more specific, more diagnostic end are narrowly defined syndromes (e.g., localized cutaneous lesions, hemorrhagic illness) and subsyndromes (e.g., influenza-like illness). These narrow classifications are helpful for monitoring of reportable diseases and other specific threats. Analysis of clinical encounters restricted to these subsyndromes must deal with sparse time series. In general, alerting algorithms applied to such series will yield higher specificity (i.e., lower false-alarm rates) than similar methods applied to broader, noisier data classifications.

At the more syndromic end of Fig. 2 are more inclusive data groupings (e.g., febrile or gastrointestinal illness). These groupings are chosen for broader sensitivity to large-scale events. Resulting time series are generally richer in structure, displaying more clearly systematic features such as annual seasonal cycles and day-of-week effects. Sparse time series in this continuum result from finer syndromic filtering, whereas the sparseness in Fig. 1 results from finer geographic subdivision. Regression modeling and other analytic approaches for removing these features to unmask signals of interest are more effective on these time series than on the sparser ones.

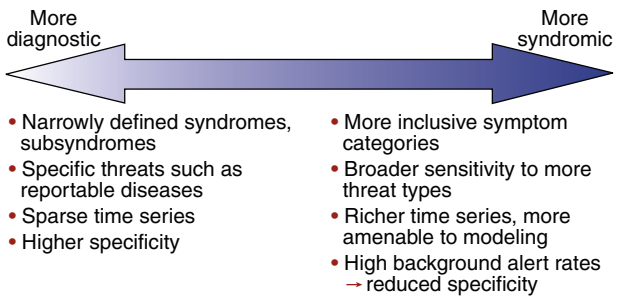


Figure 2. Dual considerations for syndromic classification of data records.

However, these noisier syndrome groupings are more subject to unmodeled systematic data patterns irrelevant to infectious disease surveillance, such as addition of new care facilities or data-transmission problems. Thus, although these groupings arguably permit detection of a wider variety of signals, the resulting alert rates are generally higher.

Grouping of Data in Time

Before 1990, most public health surveillance was done at intervals of weeks or longer. Since that time, health informatics systems have improved considerably, allowing surveillance to be performed on a daily to near-real-time basis. The increased data availability poses challenges to the statistics and data-mining communities. For anomaly detection, should quantities such as clinic visit counts be grouped by week, by day, or, as sample rates and analysis approach real-time capability, by 8-hour or smaller blocks?

Grouping of Data in Space

The spatial scale of monitoring is a function of the spatial resolution of data recorded and allowable for use as well as the monitoring institution's purview and its capacity to collect, analyze, and investigate localized outbreaks. The spatial resolution also may be subject to data-use restrictions (e.g., if exact addresses exist in data records but only ZIP codes or census tracts can be made available for analysis). The signal-to-noise background depends on whether data are monitored at state, city, or local levels. Monitoring of smaller spatial units may allow greater sensitivity to small-scale outbreaks, but only if the system is capable of controlling or managing the resulting multiple alerts and if the algorithms are effective for the scale chosen.

A special case of this tradeoff was considered by Reis and Mandl[14] when they compared (*i*) monitoring respiratory syndrome data for outbreaks from two hospitals separately with (*ii*) monitoring the combined daily counts. They found that the more advantageous strategy depended on whether the outbreak affected one or both hospitals, and their discussion also noted the rich-versus-sparse data-modeling idea above. These results led to the recommendation of a "hybrid approach that maintains both localized and aggregate models, each providing its own complementary advantages in increased detection performance. The data also may be grouped at multiple levels of aggregation to provide additional meaningful perspectives on the data."[14] Grigoryan et al.[15] also invoked this modeling dilemma in their efforts to derive outbreak reference dates from syndromic data. Their comment that "the choice of a data source and its granularity largely depend on the goal of the study and availability of the data" could be equally applied to the choice of data granularity for routine surveillance.

Decisions regarding the grouping of data to form time series for purely temporal anomaly detection often are driven by jurisdictional or logistic considerations, but these decisions can decrease the early-warning advantage of syndromic surveillance, such as when early cases are scattered among the chosen regions. The scan statistics approach[10] has become popular because it avoids preselection bias and can choose the most significant among possible outbreak locations and extents without oversensitivity to multiple testing. When the quality of a data source and the included location information are sufficient, the use of scan statistics can guide spatial aggregation and can direct limited public health resources to localities of anomalous case distributions. In adaptations of scan statistics for routine surveillance by ESSENCE and other biosurveillance applications, temporal aggregation becomes important for the estimation of the data background spatial distribution. In such cases, temporal baseline/test period decisions are necessary to obtain these estimates from historical data. For example, records supplied by a group of sentinel physicians are usually not distributed like the general population according to the most recent census, and surrogate catchment areas are generally unavailable.

TRADEOFFS RELATED TO THE ROLE OF AUTOMATION

The term *automation* here will denote electronic data processing, including collection by means of user interfaces; transmission and distribution; database storage; analysis, graphical or tabular organization and visualization; and any level of decision support, from statistical significance indications to investigation/response guides. Among these functions, those related to input, transfer, and output are widely accepted as essential to advanced disease surveillance. However, there also are tradeoffs among these basic functions. Selection of database technology should be based on the scope of the surveillance system, the range of user platform types, and the intended connectivity with other systems. Additionally, the amount of identifiable information that resides on a system involves weighing the benefit of ready information availability to health monitors against the risk of its access by unauthorized users.

The role of automation in analysis, investigation, and response may range from strictly subjective human decisions, based on visual data inspection of raw data and external knowledge, to relying completely on informatics tools for decision support. Most operational systems are much closer to the subjective extreme. Pending the acceptance of automated analysis and response algorithms, tools to facilitate the investigation process are restricted in practice to (*i*) specialized queries for various in-depth data views and (*ii*) online communication tools for sharing user interpretations, investigation history, and other external knowledge.

Factors that stimulate research and development toward increased utilization of automated analysis, especially because of limited human resources for monitoring, are the increasing number of data sources and the volume of the data in those sources, potentially important symptom and age groupings, and geographic regions of interest.

A first step toward automated analysis is the use of general alerting algorithms to decide the statistical significance of increased data counts, proportions, or other outcome variables. These algorithms are often derived from the control charts of SPC, but with two key differences. First, many of the time series derived for monitoring do not reach or maintain a steady state, and their distributions continue to evolve for a variety of reasons related to both population behavior and data capture. Second, the signal to be detected is not a persistent mean shift as in many SPC charts, and precise authentic signals are rare in these data sources, so there is no accepted definition of a target signal. Because of the vagueness caused by these issues and the application to many disparate data streams, false-alarm problems are common. Tradeoffs related to statistical alerting algorithms include the following:

- *Use of empirical versus theoretical thresholds*. Standard SPC charts require that successive observations be independent and identically distributed Gaussian, but many surveillance time series violate these assumptions. Many empirical and composite strategies have been implemented for deciding when to alert. See the article on developments in temporal alerting algorithms by Burkom et al. elsewhere in this issue.
- *Length of baseline for training periods*. The evolving nature of surveillance time series is not conducive to classical analysis using years of historical data, which may be unavailable anyway. A short, sliding baseline captures recent behavior, whereas a longer one provides more stable values for mean, variance, and other needed parameters. For daily data predictions, baselines ranging from 4 to 8 weeks have proven effective.
- *Alerting strategy*. Alerting algorithms typically borrow from the control charts successfully used to determine when industrial processes go out of control in SPC. Methods based on Shewhart charts are more suitable for the detection of single-interval spikes and other relatively impulsive signatures, whereas methods based on exponentially weighted moving average (EWMA) and cumulative summation charts are preferable for more gradual signatures.[16] Algorithms used in ESSENCE systems as of 2007 use an adaptive version of a combined Shewhart–EWMA strategy developed for hospital infection control.[9]
- *Frequentist versus Bayesian statistics*. Regarding the paradigm for statistical inference, applications using informative prior probabilities and adaptation based on posterior distributions have been rare in operational systems. For prospective, scenario-based approaches (see the accompanying article by Mnatsakanyan and Lombardo elsewhere in this issue),[17] Bayesian enhancements are essential, and further development of these enhancements is an important aspect of the effort to reduce false-positive signals.

Research to resolve these issues continues, but the gap between public health practitioners and theoreticians must be closed with an operations-research approach in which public health goals and constraints are clarified. In the meantime, use of statistical alerts is mixed among health monitors; some depend on them or combine them, whereas others ignore them.

The use of automation for more advanced decision support has been a subject of academic research, but little has been put into daily practice. When location data are available and spatial distribution is stable, the popularity of scan statistics owes to the fact that they can be used to determine location and approximate spatial extent of potential outbreaks. However, cluster significance is not clear, the problem of determining expected distributions can produce frequent irrelevant clusters, and tools need to be developed to determine the likelihood of linkage of cases in a cluster. For additional discussion of practical decision-support issues and their treatment, see Example 2 below and the article by Mnatsakanyan and Lombardo elsewhere in this issue. As in the discussion of scenario-based methods, important obstacles to the use of analytic decision-support tools are scientific validation, complicated by widely disparate data environments and the absence of sufficient labeled data, and limited epidemiological acceptance.

System designers also must choose appropriate visualization tools to expedite daily monitoring. These choices involve a tradeoff between large, scalable, web-based, multi-tiered server systems and smaller, single-user desktop applications. This tradeoff affects the potential complexity of analysis tools. Some models or simulation methods require dedicated servers and are too computationally intensive for an individual's desktop machine. Also, the functionality of the user interface can affect the modes of visual analysis available to the user.

The requirement for visual analysis capabilities may determine the use of thin-client or thick-client applications. Thin-client applications (i.e., those that can be implemented within a web browser) are sufficient for many analysis and visualization approaches. In these applications, the browser may get information by calling server programs or by running its own scripts. By contrast, thick-client applications run desktop programs that must be installed and maintained separately on the machine of each user who needs them. For example, in mapping applications based on the widely used ESRI

software, a developer could choose between ArcIMS as a thin-client tool to provide layered maps through a web browser or ArcView as a thick-client approach[18] to provide access to a desktop application with a richer set of geographic information system (GIS) tools. The thin-client/thick-client decision may be uniform across all system utilities and users, or a combination approach could be employed. Factors influencing this tradeoff are the desired capabilities, maintenance costs, licensing considerations, and availability and organization of developers.

Additional tradeoff questions related to user interface development are as follows:

1. Should all users have access to all system data? If not, how should layers of privilege be implemented? Which users should be enabled to choose configuration options, such as chart types, syndrome groupings, or alerting thresholds?
2. Can raw data be shared among users? If some cannot, what summary information can be shared? For example, hospitals or state health departments prevented from sharing patient counts may be able to share knowledge that certain diagnosis groups have reached epidemic levels.
3. Should the system maps, charts, and other visualization products be preselected or customizable?

A basic tradeoff related to these questions is that of centralization, or of how much functionality and control are to be assigned to a central monitoring entity—a person, facility, or agency, depending on the scope of the system—and how much should be distributed among other users. This tradeoff applies to decisions of data filtering, analysis methodology, and visualization modes. Opposing factors in this tradeoff are the need for uniformity in classification and description and the ability of users to choose the most appropriate methods corresponding to their local data knowledge.

Example 1: Web-Based Visualization Tools

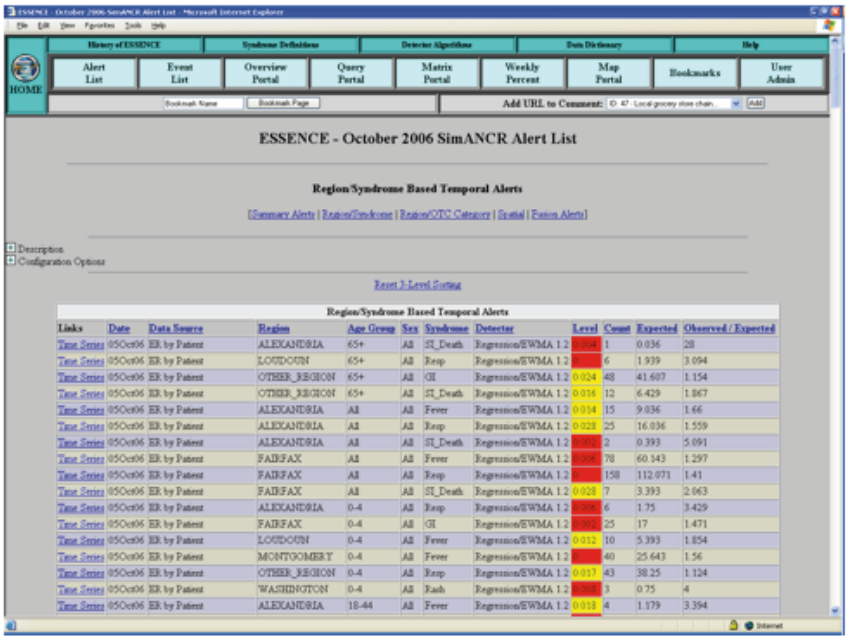
The tradeoff decisions discussed above differ across and within monitoring organizations such as state and local health departments and military treatment facilities. A biosurveillance system for widespread use must accommodate these differences. Our first example illustrates how ESSENCE developers have designed website screen displays with the required versatility. Figure 3 shows two of these displays, with the upper one (a) showing a region/syndrome-based alert list, and the lower one (b) showing a summary alert list. Other display types also have been designed for users who want to look only at geographic maps, compare certain raw data plots, follow a potential developing threat, make weekly reports, etc.

The display of individual alerts in Fig. 3a shows details of significant threshold crossings of univariate algorithms applied to individual stream time series. These data streams are organized by health district or county-based geography. Each statistical alert is for a time series filtered according to the data type, region, syndrome grouping, age category (or all ages), and patient sex (or both sexes). Each entry gives these qualifiers with the date, observed count, expected count, and a P value measure of anomaly. The user may navigate with the mouse to drill down for additional detail, and the alerts may be sorted, filtered, or pasted into spreadsheets. For a local department whose data derive only from a few hospitals, this screen may be adequate for daily monitoring. However, for a state-level department monitoring 6–12 syndrome groups in dozens of hospitals and also monitoring over-the-counter sales, nurse hotline calls, and other data types, a multiple testing problem is inevitable. The number and variety of alerts can be overwhelming, especially in departments where monitoring is left to one to two staff members with other responsibilities.

The summary alert list in Fig. 3b was designed to help the monitor see the integrated health picture. This display is divided into results for two data sources: daily counts of syndromic encounters for (*i*) emergency department visits and (*ii*) physician's office visits. For each source, results are organized in rows for geographic regions and columns for syndrome groups. These syndrome groups range from broad categories such as gastrointestinal (GI) to narrower diagnostic categories such as lymphadenitis (LYMPH). For each source/region/syndrome combination, two rows of asterisks are displayed. The asterisk rows depict alert levels for successive days, with the most current day on the right. The symbols in the upper row are shaded red to indicate significant alert levels and yellow for mild alerts. The symbols on the bottom row are shaded red, orange, yellow, blue, and green to describe the user concern level for alerts in the upper row. These respective concern levels indicate: responding/intervening, investigating, monitoring, not concerned, and for informational purposes only. Symbols in the upper row are statistical algorithm results, whereas symbols in the lower row represent reactions and findings of users, possibly investigation results or rule-outs from specific local knowledge. Thus, this display gives the user a composite population health overview informed by both data-based mathematical measures and external knowledge across regions, data types, and illness categories.

These displays accommodate the range of surveillance levels and objectives discussed in this paper. For the tradeoff of surveillance objectives, the user looking at outbreak scenarios can hypothesize about combinations of regions and illness categories suggested by the summary alert list and can look for more specific indications such as at-risk age groups or degrees of anomaly in region/syndrome alert lists. For more general health monitoring, these displays provide as broad a picture as the data sources allow, given the medical expert-based syndrome categories.

(a)



(b)

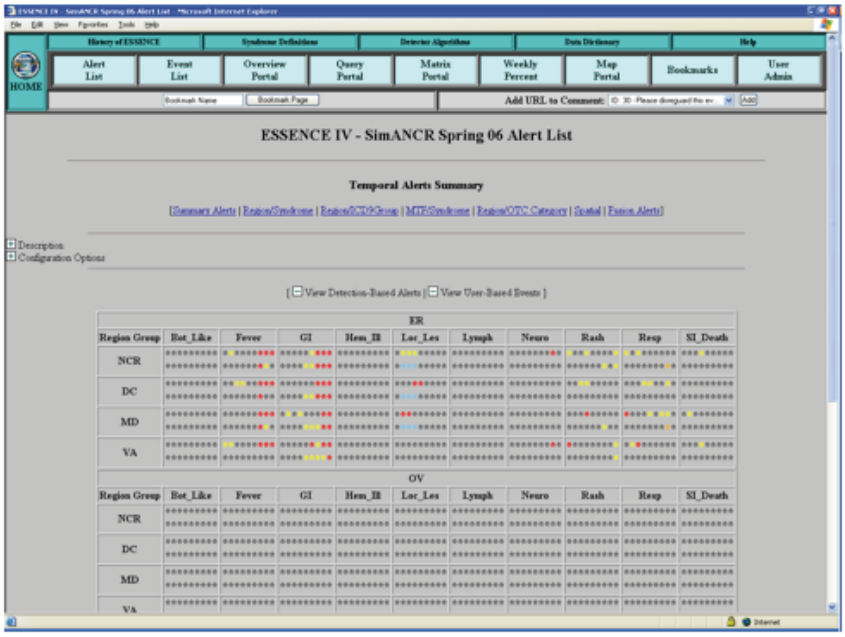


Figure 3. Visualizations for detailed and summary views. (a) The region/syndrome alert list shows details of significant threshold crossings of univariate algorithms applied to individual stream time series. Sorting, filtering, and copying are enabled. (b) The summary alert list presents results across regions, data sources, and syndrome groups for situational awareness over space and time. Both algorithm results and user concern levels are shown.

The value of these displays for addressing the aggregation tradeoff is that they give views at summary and detailed levels. State or regional users may routinely scan the overall picture and, when summary alerts or other information sources cause concern, may drill down to the level of raw data counts for more detail. The alert-list displays also may be filtered to concentrate on a specific region, syndrome type, or time interval.

Regarding the automation tradeoff, the obvious benefit of these displays is that they extract a relatively small number of alerts from hundreds or thousands of possible combinations of data sources, syndrome categories, age groups, and regions. In the summary alert displays, they also address the degree of dependence on statistical alerting algorithms by allowing the human monitor to rule out or emphasize the importance of the alerts.

Example 2: Bayesian Networks for Decision Support

This example addresses the problems of applying automated systems for decision support and for combining general and scenario-based detection. As noted above, use of decision-support tools requires validation, or scientific evidence of efficacy, and acceptance in the public health culture. The evidence must show that automated tools can fuse information from multiple data streams to improve situational awareness. The data streams are derived from clinical sources, such as emergency department visits and nurse call hotlines, as well as from nonclinical ones such as over-the-counter remedy sales and absentee rates. A decision-support tool must fuse the available data, taking into account differences in acquisition rate, reliability, and relevance to the perceived health threats. Public health acceptance requires the further demonstration that such a tool can provide understandable guidance by using routinely available data and can be implemented at reasonable cost and effort. The BN methodology is sufficiently flexible and transparent for this purpose.

A BN is a compact representation of the joint probability distribution function of a problem domain. It is typically visualized as a directed acyclic graph, with nodes representing Boolean random variables and directed edges representing conditional dependencies. For each child

node, a conditional probability table (CPT) quantifies the parent node dependencies. The nodal structure and the probabilities may be specified with a combination of expert knowledge and data analysis. See Lin et al.[17] for a further explanation and concept demonstration. For BNs constructed at APL for health surveillance, epidemiologist judgment has determined the structure, e.g., which data sources and subsyndromes show effects of a waterborne disease outbreak. The CPT probabilities have been derived by using simulations and data analysis constrained by guidance elicited from medical or epidemiological experts. To keep computations manageable as the number of data sources grow and the BN gets more complex, the data inputs to the BN are alerting algorithm outputs instead of the raw data, so that algorithmic preprocessing can handle systematic data patterns.

Figure 4 shows a sample BN built to detect outbreaks of influenza. The data sources represented in this BN are civilian emergency department visits, military outpatient clinic visits, and over-the-counter sales. The reader interested in structural details should consult the article by Mnatsakanyan and Lombardo elsewhere in this issue. For the parent node, marked "influenza," the output probability is the degree of belief that an influenza outbreak is in progress. The second-level nodes are marked as "source factor," "age factor," and "self-care factor," and these nodes in turn depend on specific data sources by age-group representation and severity. Severity is represented by hospital "discharge factor," which includes anomalies in the number of emergency department encounters that result in admissions. Subnodes are included to represent expert knowledge and observed data behavior regarding especially susceptible or early indicator age groups such as infants and the elderly.

The transparency of BN modeling is understandable from Fig. 4. A high probability of influenza epidemic may be traced to the probabilities of component nodes. The BN indication of the onset of a flu outbreak obtains credibility from a consensus among algorithms applied to data sources and likely age groups and from the subnode indication that severe cases are peaking. Furthermore, certain combinations of anomalies among the subnodes can be used for hypotheses of epidemic scenarios. For example, activation of the nodes corresponding to military clinic visits only could suggest an outbreak imported by new trainees from another location or redeployment of troops from overseas, or perhaps a biological attack on a military base. Indication of an influenza outbreak in which the only age groups showing significant anomalies are young adults with many hospital admissions, as in the second wave of the 1918 pandemic, could stimulate investigation of an unusually virulent outbreak.[19] For scenario surveillance, unusual public health threats could thus be inferred from subnodes of general BNs, or scenario-based BNs could be implemented. The BN approach can satisfy requirements for both general health surveillance and the monitoring of potentially disastrous threats.

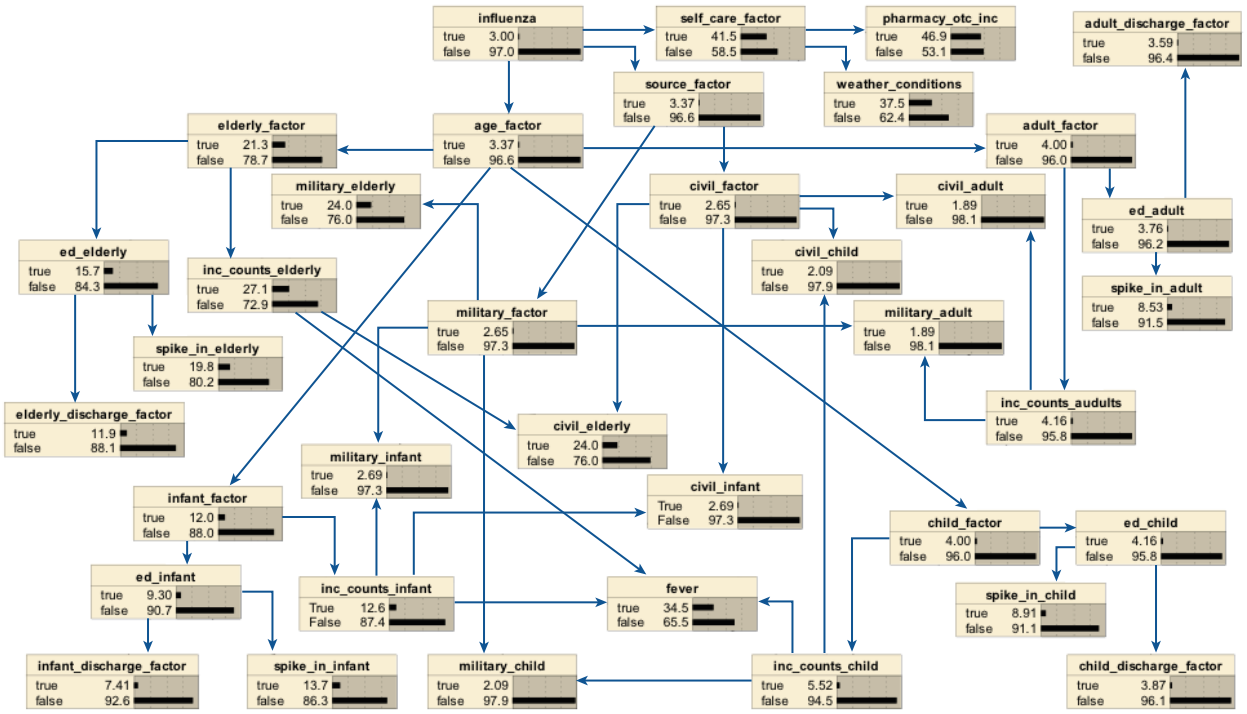


Figure 4. BN combining epidemiological knowledge and data analysis for automated decision support and increased specificity.

Regarding the aggregation tradeoffs, the BN approach can help with both syndromic and spatial issues. Nodes representing regions and syndrome groups that are sparse in data counts may be added to a BN, but the higher-level node CPTs may require corroboration from other evidence nodes that will reduce the overall false-alarm rate.

Example 3: Significant Case Cluster Determination

The final example concerns aggregation of spatial data for detection of disease clusters but also concerns the tradeoff in surveillance objectives. The issue behind this example was the spatial unit for cluster detection in a dataset of outpatient clinic visit counts illustrated in Fig. 1. In the data records, each outpatient visit record contains ZIP codes for both patient residence and the clinic where treatment occurred. Respiratory syndrome data for the 3+-year time period contained 32 clinic ZIP codes and 1095 residence ZIP codes. However, the distribution of records among these ZIP codes is highly skewed, as seen in Fig. 5. The upper and lower histograms show the numbers of clinics and residence ZIP codes, respectively, binned by median daily count. At the facility level, the median count is zero in 20 of the 32 facilities, and the great majority of cases are seen at only 6 facilities. The skewness is even clearer at the residence ZIP-code level, with median counts of zero for 89% of the 1233 ZIP codes. What the skewness means for cluster detection is that, near the large facilities and the ZIP codes with many cases, a large number of cases is required for a cluster to become statistically significant. In regions with sparse data representation, a much smaller outbreak is detectable. These statements are quantified by applying scan statistics to detect simulated signals representing outbreak data effects.

The procedure used for simulating the signals is as follows:

1. Choose a central subregion to be used as the location of origin for a point-source outbreak.
2. Generate a stochastic, lognormal epidemic curve using a median incubation period of 3 days.[20]
3. Given the number of outbreak-attributable cases on a particular day, choose the subregion for each case with a random draw assuming that the number of cases decays exponentially with the distance of the subregion from the outbreak center.

The signal injection methodology was:

1. To estimate background cluster rates, run the scan statistics algorithm M times, once for each day of the 3+ years of background data except for a warm-up period of 8 weeks for an initial estimate of the spatial distribution. An estimate of the probability of false alarm per day at threshold α is then $PFA_\alpha = MF_\alpha/M$, where MF_α is the number of days with P value $< \alpha$ for some cluster. (One cannot infer the background alert rate from the P value alone because the empirical data typically violate the underlying statistical assumptions.)
2. Conduct 100 trials, each with a unique stochastic outbreak signal with its own distribution in space and time. In each trial, for each day of the simulated signal, add the attributable counts to the authentic background counts for each affected subregion. Choose the signal start date for the initial trial early in the dataset, and for each subsequent trial, move the injected signal forward 8 days to vary the background noise and the day-of-week and seasonal effects.
3. For each trial, run the scan statistics algorithm for each day of the signal, and find the minimum P value for each day, but only for those clusters that intersect the simulated outbreak region. Those clusters will be treated as detections if the P value is below threshold α.
4. The estimate of the probability of detection (representing sensitivity) at threshold α is then $PD_\alpha = ND_\alpha/100$, where ND_α is the number of trials with a P value below α on signal injection days.

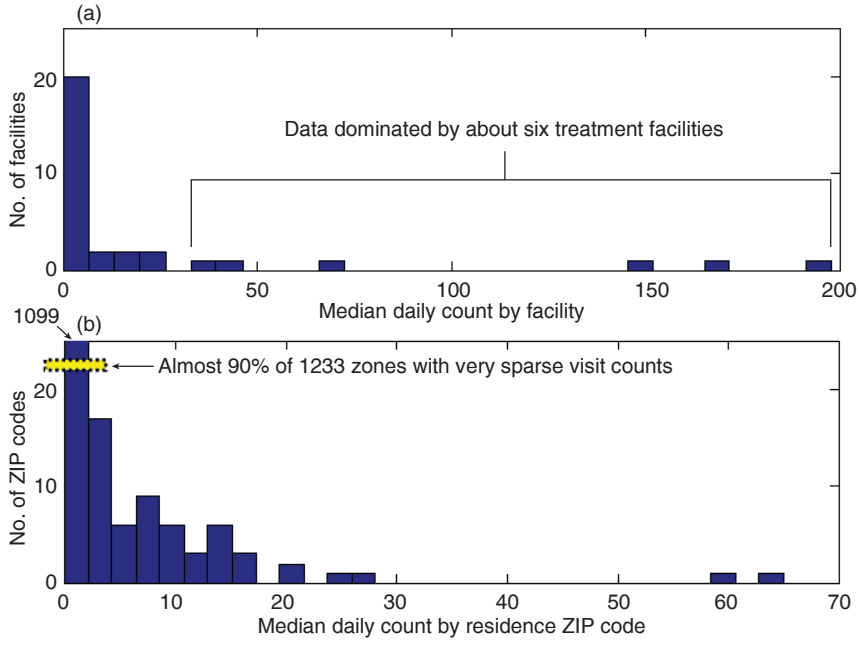


Figure 5. Histograms of median daily respiratory syndrome visit counts by treatment facility (a) and residence ZIP code (b).

A number of these experiments were conducted to examine the effect of tradeoffs in spatial aggregation, and two such experiments were compared for which the center of the outbreak was placed in one of the sparse subregions not close to any large clinic. The first experiment sought clusters using only the 32 clinic ZIP codes, while the second used the 1233 patient residence ZIP codes. Sensitivity was measured in two ways: the first was to credit a detection if a significant cluster was found within the first 3 days of the signal, and the second was to credit a detection at any day of the signal. The receiver operating characteristic (ROC) curves of Fig. 6 allow comparison of the results for both levels of aggregation using both measures.

The *x* axis of Fig. 6 gives the estimated background alert rates per day for alert rates of <0.15 (an upper bound of approximately one per week) to focus on alarm frequencies that are practical for public health surveillance. The *y* axis gives the two empirical sensitivity measures described above. It is characteristic of this dataset that, for these comparisons, lower alert rates can be achieved by using subregions at the facility level. However, for scenarios with a small outbreak away from large facilities, much higher sensitivity can be achieved by using residence ZIP code data, and the advantage increases as the algorithm threshold is permitted to increase (i.e., as the health monitor can tolerate false alarms).

To understand these performance comparisons relative to pure time series monitoring with fixed region sizes, temporal detection algorithms removing the seasonal and day-of-week effects were run for time series at the state, large-facility, and small-facility levels on the same 100 trial outbreak signals. For each trial, the same attributable signal counts were added to background data projected onto each of these time series, whose scales may be reviewed in Fig. 1. Not surprisingly, the 40-case signal spread over 9 days was not detected in the statewide or large-clinic time series. To better compare the methods, we varied the total outbreak size for the same signal shapes and reran the temporal algorithms. The summary in Table 1 shows the number of total cases required to achieve 90% and 75% detection sensitivity at the different aggregation levels, where sensitivity was again measured for the entire signal and for detection by day 3. This table gives an idea of the signal sizes, spread over 7–10 days, that are detectable at high sensitivity for these aggregation levels. For the large-clinic and statewide data series, these sizes are one and two orders of magnitude greater than the signal detectable by using scan statistics and residence ZIP codes for a moderate detection threshold. The small-facility table entries do give results that are comparable with the 40-case outbreak. However, monitoring 32 (in this dataset) time series for the separate facilities introduces a multiple-testing problem if they are all monitored at one central location. For a detailed treatment of the problem of monitoring multiple univariate series, see Marshall et al.[21]

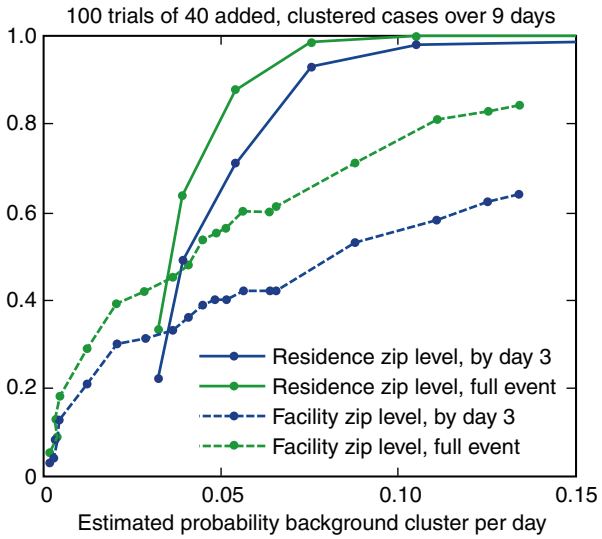


Figure 6. ROC curve comparison results for spatial signal detection at home versus facility aggregation level.

To choose methods and aggregation levels, a monitoring institution should interpret results such as Figs. 4 and 5 in terms of its objectives and resources. The question of which method to use and the level of spatial resolution depends on the thematic considerations of this paper:

Table 1. Outbreak sizes required for 75% and 90% detection sensitivity by using temporal methods alone, by level of spatial aggregation.

Aggregation level	Size of outbreak effect (number of data cases required) for desired temporal detection probability					
	State (median 829)		Large facility (median 166)		Small facility (median 3)	
Detection probability (%)	75	90	75	90	75	90
By day 3	660	1100	170	246	23	30
By day 7	1100	1500	308	407	33	46

1. If the objective is bioterrorism, interest may be limited to outbreaks near the large clinic and the methods with smaller background alert rates.
2. The utility of monitoring a large number of small data streams may depend on whether a health department has distributed surveillance and investigation capability. If an investigation requires a half day, alerts occur several times per week purely because of multiple testing, and the health department has only one to two staff members responsible for surveillance, investigations cannot keep pace with indicated alerts.
3. The scan statistics paradigm can help, but for datasets such as the outpatient visit records in the example, a small spatial resolution will require additional investigation capability. Data analysis should be done before operational use to examine likely cluster rates and significant cluster characteristics.

CONCLUSIONS

In view of public health concerns regarding bioterrorism and the inevitability of pandemic influenza, development of advanced surveillance systems to corroborate, supplement, and close gaps in physician sentinel surveillance is a research imperative. Early implementations of these systems have yielded uneven results in acceptance and in demonstrated utility. Public health users have reported benefits from these systems, but not always the benefits conceived in system design.[7] Given the multiple uses of health surveillance systems for a variety of purposes envisioned only after system design and implementation, the small number of reported successes and the excessive false alarms experienced by some users are understandable. Requirements analysis should be driven by an understanding of decisions regarding each of the tradeoffs discussed above, involving surveillance objectives, data aggregation, and the roles of automation.

The interrelationships among these tradeoffs also must be understood. Levels of syndromic, temporal, and spatial aggregation should be compatible with goals and response capabilities. Database and visualization tools should be chosen and adjustable to the appropriate complexity for intended use. For scenario-based detection to be accepted and successfully used by health monitors, the methodology must be sufficiently transparent, with user-friendly automated tools to clarify the basis for decision support.

Once the tradeoff decisions are understood, they can drive choices regarding data source selection and filtering, anomaly detection methods, visualization tools, and the resultant investigation and response protocols. Research initiatives related to public health surveillance should be guided by well defined objectives given at the level of scenarios and required detection performance. Improved, context-sensitive requirements analysis will determine appropriate choices for these tradeoffs and will sharpen the roles and improve the utility of advanced surveillance systems.

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

REFERENCES

[1]Lombardo, J. S., "The ESSENCE II Disease Surveillance Test Bed for the National Capital Area," *Johns Hopkins APL Tech. Dig.* **24**(4), 327–334 (2003).

[2]Eban, K., "BioSense or Biononsense?," *The Scientist* **21**(4), 32 (2007). Available at http://www.the-scientist.com/2007/4/1/32/1/ (accessed 19 Dec 2007).

[3]Burkom H. S., Elbert, Y. A., Feldman, A., and Lin, J., "Role of Data Aggregation in Biosurveillance Detection Strategies with Applications from the Electronic Surveillance System for the Early Notification of Community-Based Epidemics," *MMWR Morb. Mortal. Wkly. Rep.* **53**(Suppl. 1), 67–73 (2004).

[4]Buehler, J. W., Hopkins, R. S., Overhage, J. M., Sosin, D. M., and Tong, V., "Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks," *MMWR Morb. Mortal. Wkly. Rep.* **53**(RR05), 1–11 (2004). Available at http://www.cdc.gov/mmwR/preview/mmwrhtml/rr5305a1.htm (accessed 19 Dec 2007).

[5]"Medical Management of Biological Casualties Handbook," U.S. Army Medical Research Institute of Infectious Diseases, Ft. Detrick, MD, Fourth Ed. (Feb 2001).

[6]Kauffman, A. F., Meltzer, M. I., and Schmid, G. P., "The Economic Impact of a Bioterrorist Attack: Are Prevention and Postattack Intervention Programs Justifiable?," *Emerg. Infect. Dis.* **3**(2), 83–94 (1997).

[7]U.S. Medicine Institute for Health Studies, "Roundtable Discussion: Assessing Syndromic Surveillance: Costs, Benefits, Future," Executive Summary (19 Oct 2007). Available at http://www.usminstitute.org/roundtables.html (accessed 19 Dec 2007).

[8]Baker, M., Smith, G. E., Cooper, D., Verlander, N. Q., Chinemana, F., et al., "Early Warning and NHS Direct: A role in community surveillance?" *J. Public Health Med.* 25:362–368 (2003).

[9]Morton, A. P., Whitby, M., McLaws, M.-L., Dobson, A., McElwain, S., et al., "The Application of Statistical Process Control Charts to the Detection and Monitoring of Hospital-Acquired Infections," *J. Qual. Clin. Pract.* **21**, 112–117 (2001).

[10]Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., and Mostashari, F., "A Space-Time Permutation Scan Statistic for the Early Detection of Disease Outbreaks," *PLoS Med.* **2**(3), 10.1371/journal.pmed.0020059 (2005).

[11]Buckeridge, D. L., Owens, D. K., Switzer, P., Frank, J., and Musen, M. A., "Evaluating Detection of an Inhalational Anthrax Outbreak," *Emerg. Infect. Dis.* **12**(12), 1942–1949 (2006).

[12]Hogan, W. R., Cooper, G. F., Wallstrom, G. L., Wagner, M. W., and Depinay, J.-M., "The Bayesian Aerosol Release Detector: An Algorithm for Detecting and Characterizing Outbreaks Caused by an Airborne Release of *Bacillus anthracis*," *Stat. Med.* **26**(29), 5225–5252 (2007).

[13]Burkom, H., Murphy, S. P., and Shmueli, G., "Automated Time Series Forecasting for Biosurveillance," *Stat. Med.* **26**(22), 4202–4218 (2007).

[14]Reis, B. Y., and Mandl, K. D., "Integrating Syndromic Surveillance Data Across Multiple Locations: Effects on Outbreak Detection Performance," in *AMIA Annu. Symp. Proc.*, Washington, DC, pp. 549–553 (2003).

[15]Grigoryan, V. V., Wagner, M. M., Waller, K., Wallstrom, G. L., and Hogan, W. R., *The Effect of Spatial Granularity of Data on Reference Dates for Influenza Outbreaks*, Real-Time Outbreak and Disease Surveillance Technical Report (2005). Available at http://rods.health.pitt.edu/LIBRARY/2005%20AMIA-Grigoryan-Reference%20dates%20for%20flu-submitted.pdf (accessed 19 Dec 2007).

[16]Ryan, T. P., *Statistical Methods for Quality Improvement*, Wiley, New York (1989).

[17]Lin, J. S., Burkom, H. S., Murphy, S. P., Elbert, Y., Hakre, S., et al., "Bayesian Fusion of Syndromic Surveillance with Sensor Data for Disease Outbreak Classification," Chap. 6, in *Life Science Data Mining*

Science, Engineering, and Biology Informatics, Vol. 2, S. Wong and C.-S. Li (eds.), World Scientific Publishing, Hackensack, NJ (2007).
[18]East, R., Goyal, R. K., Haddad, A., Konovalov, A., Rosso, A., et al., "The Architecture of ArcIMS, a Distributed Internet Map Server," in *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases*, Redondo Beach, CA, Lecture Notes in Computer Science, Vol. 2121, pp. 387–403 (2001).
[19]Taubenberger, J. K., and Morens, D. M., "1918 Influenza: The Mother of All Pandemics," *Emerg. Infect. Dis.* **12**(1), 15–22 (Jan 2006).
[20]Burkom, H., Hutwagner, L., and Rodriguez, R., "Using Point-Source Epidemic Curves to Evaluate Alerting Algorithms for Biosurveillance," in *2004 Proceedings of the American Statistical Association, Statistics in Government Section*, Toronto, Canada, on CD-ROM (Jan 2005).
[21]Marshall, C., Best, N., Bottle, A., and Aylin, P., "Statistical Issues in the Prospective Monitoring of Health Outcomes Across Multiple Units," *J. R. Stat. Soc. Ser. A* **167**, 541–559 (2004).

The Authors


Howard S. Burkom


Wayne A. Loschen


Zaruhi R. Mnatsakanyan


Joseph S. Lombardo

Howard S. Burkom received a B.S. degree from Lehigh University and M.S. and Ph.D. degrees in mathematics from the University of Illinois at Urbana–Champaign. He has 7 years of teaching experience at the university and community college levels. Since 1979, he has worked at APL developing detection algorithms for underwater acoustics, tactical oceanography, and public health surveillance. Dr. Burkom has worked exclusively in the field of biosurveillance since 2000, primarily adapting analytic methods from epidemiology, biostatistics, signal processing, statistical process control, data mining, and other fields of applied science. He is an elected member of the Board of Directors of the International Society for Disease Surveillance. **Wayne A. Loschen** holds an M.S. in computer science from The Johns Hopkins University and a B.S. in computer science from the University of Illinois. He has worked at APL since 1999 as a software engineer, database administrator, system administrator, GIS developer, and web developer, and he has extensive experience in the management and visualization of large databases of environmental and health-related data. His work on the ESSENCE disease surveillance project has involved the development of novel detection algorithms, user interfaces, and visualization techniques for effectively understanding complex, multifaceted, and multistream temporal and spatial data. **Zaruhi R. Mnatsakanyan** is leading research activities in the field of decision support and data fusion for The Johns Hopkins University Center of Excellence in Public Health and Disease Surveillance Program in the National Security Technology Department at APL. Dr. Mnatsakanyan has a Doctor of Engineering degree in computer science and machinery engineering from the State Engineering University of Armenia as well as master's degrees in computer science and in computer science and automated engineering from the same university. Her research interests include sensor data fusion, Bayesian Networks, pattern recognition, machine learning, and cognitive science. **Joseph S. Lombardo** has been employed by APL for the past 38 years performing research in 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. For further information on the work reported here, contact Dr. Burkom. His e-mail address is howard.burkom@jhuapl.edu.