Statistical Challenges Facing Early Outbreak Detection in Biosurveillance

Galit SHMUELI

Department of Decision,
Operations & Information Technologies and
The Center for Health Information
and Decision Systems
Robert H. Smith School of Business
University of Maryland
College Park, MD 20742
(*gshmueli@rhsmith.umd.edu*)

Howard BURKOM

The Johns Hopkins University
Applied Physics Laboratory
Laurel, MD 20723
(*Howard.Burkom@jhuapl.edu*)

Modern biosurveillance is the monitoring of a wide range of prediagnostic and diagnostic data for the purpose of enhancing the ability of the public health infrastructure to detect, investigate, and respond to disease outbreaks. Statistical control charts have been a central tool in classic disease surveillance and also have migrated into modern biosurveillance; however, the new types of data monitored, the processes underlying the time series derived from these data, and the application context all deviate from the industrial setting for which these tools were originally designed. Assumptions of normality, independence, and stationarity are typically violated in syndromic time series. Target values of process parameters are time-dependent and hard to define, and data labeling is ambiguous in the sense that outbreak periods are not clearly defined or known. Additional challenges include multiplicity in several dimensions, performance evaluation, and practical system usage and requirements. Our focus is mainly on the monitoring of time series to provide early alerts of anomalies to stimulate investigation of potential outbreaks, with a brief summary of methods to detect significant spatial and spatiotemporal case clusters. We discuss the statistical challenges in monitoring modern biosurveillance data, describe the current state of monitoring in the field, and survey the most recent biosurveillance literature.

KEY WORDS: Anomaly detection; Control chart; Disease outbreak; Statistical process control; Syndromic data.

1. INTRODUCTION

Biosurveillance is the practice of monitoring data to detect, investigate, and respond to disease outbreaks. Traditional biosurveillance has focused on the collection and monitoring of diagnostic medical and public health data retrospectively to determine the existence of disease outbreaks. Examples of traditional data are cause-specific mortality rates and daily or weekly counts of selected laboratory results. Although such data are the most direct indicators of the current burden of a disease of interest, in most situations they are collected, delivered, and analyzed days, weeks, or even months after the outbreak. By the time this information reaches decision makers, it may be too late for public health interventions that might avoid or ameliorate early cases or to react in other ways, such as stockpiling and dispensing vaccine and medication.

Disease surveillance research in the late 1990s shifted toward biosurveillance systems that would provide early detection of diseases resulting either from bioterrorist attacks or from "natural" causes, such as the avian flu. This shift meant monitoring information sources not previously used at time scales shortened from weeks or months to days or hours. Modern biosurveillance uses less specific aggregated healthcare-seeking behavior data (also called syndromic data) from opportunistic sources in search of earlier outbreak signals. Syndromic data are derived from prediagnostic information, such as over-the-counter (OTC) and pharmacy medication sales, calls to nurse hotlines, school absence records, searches on medical Web sites, and complaints of individuals entering hospital emergency departments. None of these data directly measure the number of cases of any specific disease, but it is assumed that they contain an outbreak signal earlier than that of traditional sources, because they contain measurable effects of care-seeking behavior before patients experience acute or disease-specific symptoms. The underlying assumption is that data collected from this early care-seeking behavior, such as purchasing OTC remedies, will contain a sufficiently strong and early signal of the outbreak when aggregated across the monitored population. The various data sources fall along a continuum according to both diagnostic specificity and likely detection timeliness. Under the assumption that people tend to self-treat and self-medicate before rushing to the hospital, we would expect Web searching and the purchasing of OTC remedies to precede calls to nurse hotlines and ambulance dispatches, and followed by emergency department visits. Still, this entire continuum is assumed to occur before actual clinical diagnoses can be made (after hospitalization and/or laboratory tests). In addition to monitoring syndromic data, there have been efforts to monitor other types of data associated with disease risk factors, such as air and water quality measurements. All of these evi-

TECHNOMETRICS, FEBRUARY 2010, VOL. 52, NO. 1
DOI 10.1198/TECH.2010.06134

dence sources have been discussed in the context of biosurveillance. We focus here on data sources in current use in existing surveillance systems. (Note: The term *syndromic surveillance* is widely been used to describe infectious disease surveillance using nontraditional data sources in the current context.)

2. BACKGROUND AND CHARACTERISTICS OF SYNDROMIC DATA

Along with the shift in the type of data collected for biosurveillance came a shift in the collection frequency and transfer rate of data. Currently, many U.S. surveillance systems routinely collect data from multiple sources on a daily basis, and these data are transferred with variable delays to the biosurveillance system. (See Fienberg and Shmueli 2005 for a description of this process and examples from several surveillance systems.) Although the data and goals of syndromic surveillance have evolved from those of traditional disease surveillance, many of the traditional monitoring methods remain essentially unchanged in the new context. For example, Figure 1 shows a data series from a traditional source (left) versus one from a modern source that might be used for tracking influenza activity (or detecting an influenza outbreak). The traditional data are weekly counts of pneumonia and influenza-related deaths in a particular U.S. city. In addition to this mortality series, six additional measures are tracked, all based on either mortality or laboratory reports. In contrast, the syndromic series are daily counts of doctor visits related to respiratory complaints in a particular city, before a clinical diagnosis of influenza is made. Thus the two series differ in frequency (daily vs. weekly), in the directness of measuring influenza (confirmed lab reports or mortality vs. prediagnostic indications), and in availability relative to time of diagnosis. A key task is to learn to combine these new data sources with traditional ones so that the new information will clarify, not cloud, the situational awareness of public health monitors.

Several surveillance systems aimed at rapid detection of disease outbreaks and bioterror attacks have been deployed across the United States and in other countries in the last few years. Three of the U.S. systems serve a wide geographical region, and there are increasing numbers of more local systems that collect and monitor data at a county level, city level, or even hospital level. These systems collect clinical data, usually at a daily rate, including emergency department chief complaints and admissions, visits to military treatment facilities, and 911 calls. Nonclinical data include OTC medication and health-care product sales at grocery stores and pharmacies, prescription medication sales, HMO billing data, school/work absenteeism records, and more. The National Bioterrorism Syndromic Surveillance Demonstration Program run by the Harvard Medical School and Harvard Pilgrim Health Care (Yih et al. 2004) tracks data from health plans and practice groups. Of these, the most commonly monitored are records of emergency department patient encounters. Much effort has been devoted to classifying these records' textual chief complaints into syndrome groupings. A patient may report several symptoms (e.g., rash, fever), thereby contributing multiple chief complaints. For billing purposes, the emergency departments themselves classify patient records into International Classification of Diseases 9 (ICD-9) codes, although sometimes this coding takes several days. Some surveillance systems form syndrome groupings based on these derived codes to avoid the complication of parsing chief complaint strings that vary by institution, geographic region, an so on.

In general, data modeling efforts have been focused in two directions: modeling temporal data within a particular geographical region and modeling spatial data at a certain point in time or over time. Often the choice is due to the type of data available. The former setting is similar to statistical quality control, where one or more streams of data are inspected for abnormalities prospectively. In the spatial (or spatiotemporal) applications, methods are aimed at detecting regions whose case distribution is abnormally high compared with other regions. Focusing on the temporal approach, for a particular geographical location, we can view the data in a hierarchical structure. The first level is the data source (e.g., emergency department or pharmacy), and within each data source there are one or more time series, as illustrated in Figure 2. This structure suggests that same-source series should be more similar than series from different sources. This perspective can influence the type of monitoring methods used within a source as opposed to methods for monitoring the entire system, and raises the question of whether a hierarchical model or a flat model is more suitable.

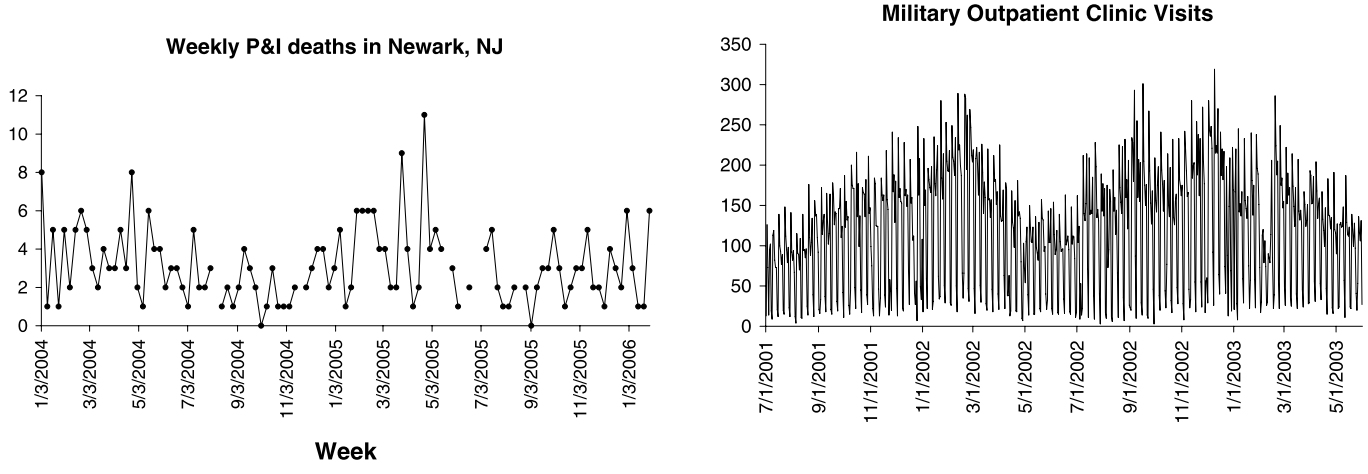

Figure 1. Typical traditional data (left) versus syndromic data (right) for monitoring influenza.

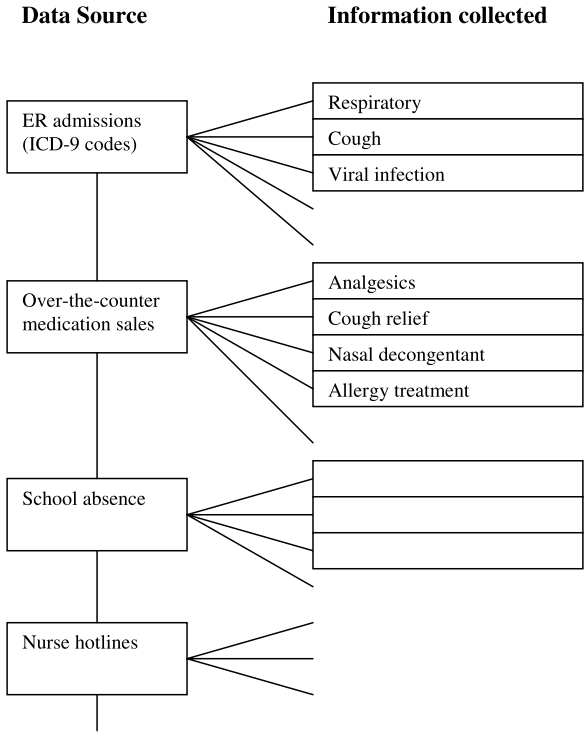

Figure 2. Sketch of data hierarchy. Each data source can contain multiple time series.

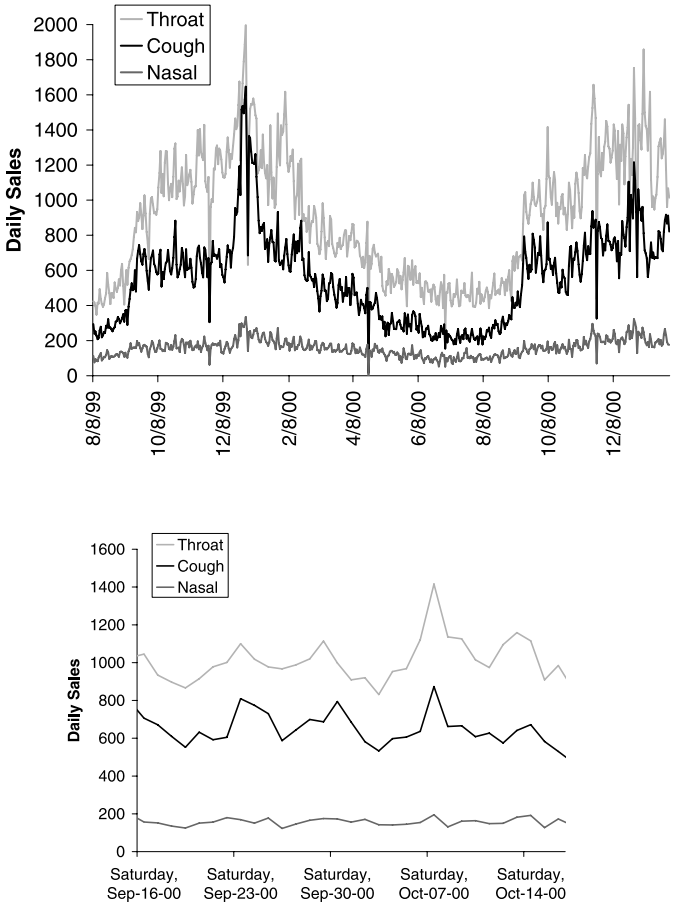

Figure 3. Daily sales of over-the-counter medications from a large grocery chain in the Pittsburgh, Pennsylvania are, by medication subgroup. The bottom panel is a zoom-in on a 1-month period. Two of the series exhibit strong biannual seasonality, and all series exhibit some level of DOW effect. Dips on holidays are due to store closings.

Several features of syndromic data arise from their application context. Unlike diagnostic data, syndromic data are indirect indicators of an outbreak, and most syndromic information is taken opportunistically from data sources developed for insurance billing, inventory management, or some other purpose (e.g., Rolka 2006). Time series derived from such data are subject to sources of variation irrelevant to outbreak detection, such as the correlation of cough medication sales with overall grocery sales. A prominent characteristic of these time series is nonstationarity. Means, variances, and autocorrelation structures tend to change over time, and the degree of nonstationarity changes from series to series; however, these time series display a few general predictable patterns, such as characteristic day-of-week (DOW) behavior. In U.S. emergency department visits, daily counts are typically low on weekends and high early in the work week (Burkom, Murphy, and Shmueli 2007), but also can exhibit other daily patterns (e.g., Reis and Mandl 2003; Brillman et al. 2005), or none (Fricker 2006). On weekends, grocery stores tend to have more traffic, and thus increased medication sales (e.g., Goldenberg et al. 2002). DOW effects can be seen in Figures 3–5, which show daily syndromic data from different sources. These time series also typically display abnormal behavior on holidays and post-holidays (e.g., Fienberg and Shmueli 2005) due to holiday closings (e.g., schools) or limited operation mode (e.g., pharmacies, hospitals), as shown in Figure 3. Annual seasonal population behavior and weather variations also cause characteristic cyclic series features (e.g., some of the series in Figures 3–5). These background features complicate the recognition of the start of an epidemic. The daily data collection frequency also leads to nonnegligible short-term autocorrelation. During the cold season, for example, the number of emergency department visits is usually correlated across successive days. Finally, data quality issues further complicate interpretation of syndromic information. These issues include missing or duplicate records, coding errors, changes in the level of participation of data providers, and inconsistent reporting, and the decision of whether and how to statistically control for each depends on the individual data source.

3. CHALLENGES

The assumption behind syndromic surveillance is that the effect of a disease outbreak will manifest itself as an anomaly in properly filtered population data when expected background behavior is removed or when the data are compared to similar data from unaffected populations. The similarity to the classic quality control setting has led to a widespread use of control charts in public health monitoring (Benneyan 1998; Woodall 2006) and also in temporal biosurveillance. However, the biosurveillance setting is different than the industrial setting in terms of the nature of the collected data, the underlying background behavior, the nature of an outbreak, the evaluation of performance, and the requirements and uses of a biosurveillance system. We discuss each of these in this section.

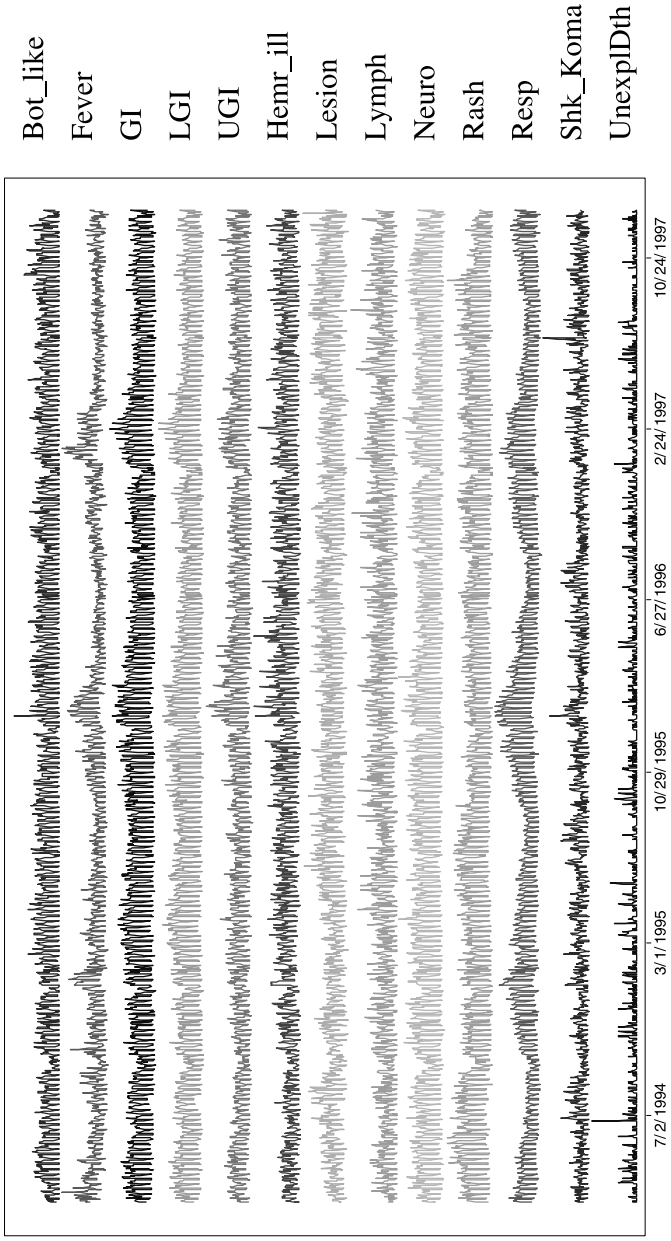

Figure 4. Daily counts of chief complaints at emergency departments in a certain U.S. city, by type of complaint. The first 12 series combine counts of ICD-9 diagnosis codes relevant to a certain syndrome, using the CDC's list of syndrome definitions. The last series is counts of unexplained death.

3.1 Determining and Modeling Background Behavior

Identifying the presence of an abnormality in data requires defining normal behavior. One complication arises from the intended dual use of biosurveillance systems for detecting natural and bioterror-related or pandemic disease outbreaks, because the data footprint of a seasonal influenza epidemic is a target signal in the former context but part of the background clutter in the latter. In the bioterrorism monitoring context, all usual seasonal influences should be removed for sensitivity even at the peak and aftermath of a usual influenza outbreak. It is nearly impossible to obtain exact dates of local natural disease outbreaks, however. This lack of labeling greatly complicates the evaluation and comparison of detection algorithms. Another challenge is the "time alignment" problem (Rolka et al. 2007); although epidemiologists may surmise the logical sequence of individual care-seeking behavior (e.g., self-medicating before seeing a doctor), the delays from infection to these behaviors and the delays between these behaviors are difficult to quantify. Finally, the population being monitored is very dynamic. Changes in population, data reporting, hospital policies, and other factors lead to a nonstationary, constantly evolving background behavior that is not easy to model using standard techniques. All these factors lead to a lack of well-defined training (phase I) data.

3.2 The Nature of Outbreaks and Their Determination

When considering which monitoring scheme to use, an important factor is the nature of the abnormal behavior to be detected. The behavior includes the magnitude, shape, and expected length of the abnormal behavior. For example, cumulative sum (cusum) charts are more effective than Shewhart X-bar charts in detecting small constant shifts in the process mean. More generally, given a certain signature, one may try to develop the most effective filter to detect it. Biosurveillance involves knowledge about the progression of different diseases in the population derived using theoretical disease epicurve models (e.g., Burkom, Hutwagner, and Rodriguez 2005a) or estimated from historical data, such as the accidental anthrax release in Sverdlovsk, Russia in 1979 (Meselson et al. 1994; Goldenberg et al. 2002; Brookmeyer et al. 2005). For example, Wagner et al. (2001) discussed the footprint of an anthrax outbreak in medical data, and Pavlin (1999) described the difference between the epidemic curve of a deliberate bioterrorism-related disease and that from a natural disease. There has been very little discussion of the expected signature in nonclinical data, especially nontraditional data, however; for example, the manifestation of an anthrax attack in ambulance dispatches or in sales of cough remedies remains unknown. This next step requires inputs from medical and public health experts, as well as domain experts, such as marketers. Such an approach was described by Fienberg and Shmueli (2005). The unknown nature of the outbreak signature means that the task is one of anomaly detection rather than signature identification. Furthermore, it is a nonspecific task; modern biosurveillance systems are intended to detect a wide range of disease outbreaks, ranging from short and intense to gradual and from infectious to noninfectious.

Another major challenge arises from the difficulty in obtaining properly labeled data with exact outbreak periods. This challenge also arises in traditional disease surveillance, where the onset of a local outbreak is difficult to pinpoint. For example, for influenza, the Centers for Disease Control and Prevention (CDC) uses the following national baseline model (*http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5413a2.htm*):

> The expected seasonal baseline proportion of [Pneumonia and Influenza (P&I)] deaths reported by the 122 Cities Mortality Reporting System is projected by using a robust cyclical regression procedure in which a periodic regression model is applied to the observed percentage of deaths from P&I during the preceding 5 years. The epidemic threshold is 1.645 standard deviations above the seasonal baseline.

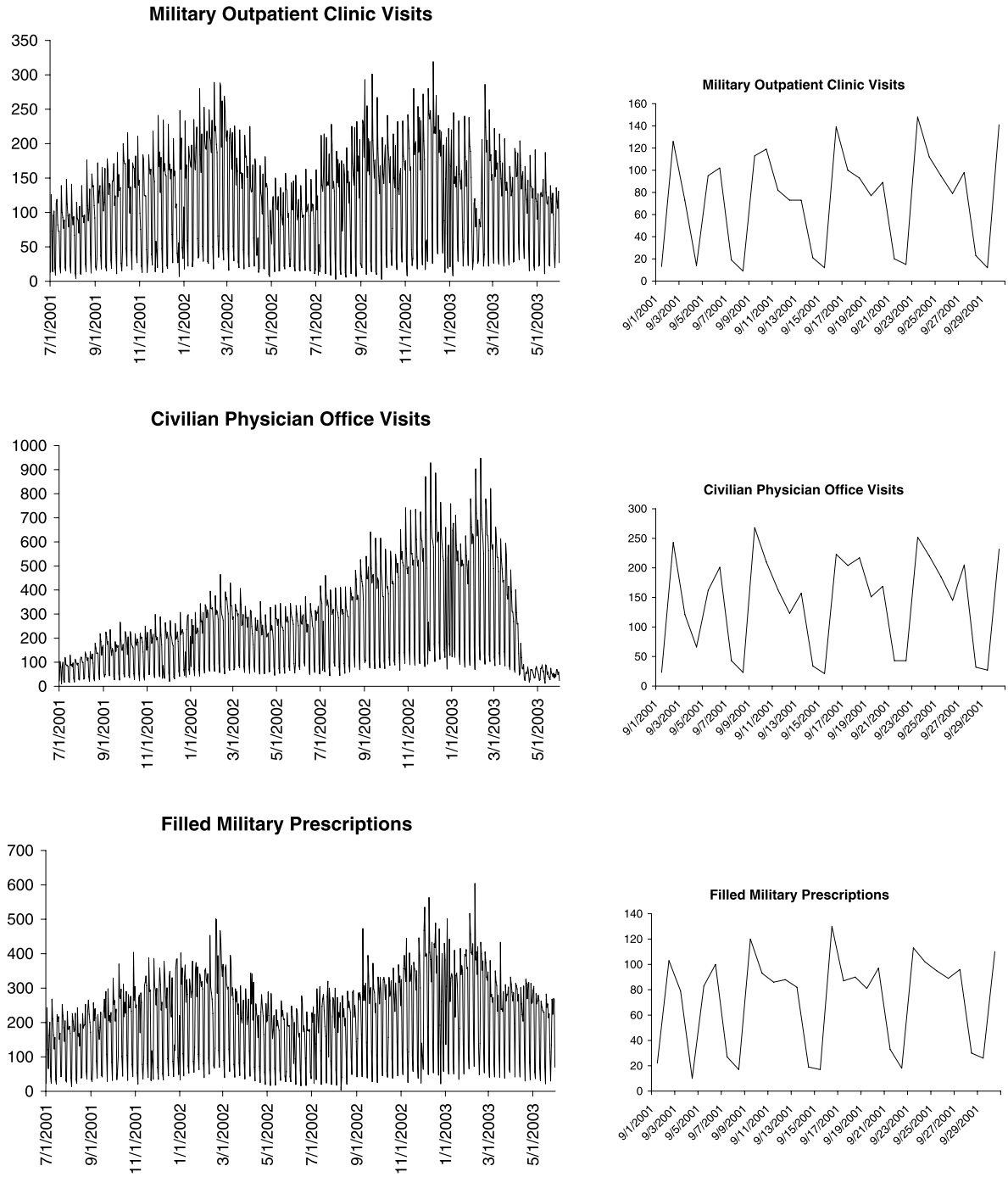

Figure 5. Daily respiratory-related counts in a certain U.S. city from three data sources: visits to military outpatient clinics (top), civilian physician visits (middle), and filled military prescriptions (bottom). Entire 700-day period (left) and single month (right). These are part of a larger data set used in the Bioevent Advanced Leading Indicator Recognition Technology (BioALIRT) biosurveillance program of the Defense Advanced Research Projects Agency (DARPA) conducted between 2001 and 2004 (see Siegrist and Pavlin 2004; Buckeridge et al. 2005, for further details).

National outbreaks are then determined using this baseline (see Figure 6). But the determination of influenza peak activity is done retrospectively; the model assumes a deterministic cyclical behavior where in practice the onset of influenza can occur at different times in different years. Moreover, this national model is the basis for determining national outbreaks, but no solution is offered for local outbreak determination. ("Wide variability in regional data precludes calculating region-specific baselines; applying the national baseline to regional data is inappropriate.")

In the context of modern biosurveillance, the recent BioALIRT biosurveillance program of the Defense Advanced Research Projects Agency (DARPA) was aimed at evaluating different algorithms using a common set of syndromic data in multiple U.S. cities. To determine natural outbreaks in the data, a team of epidemiologists and medical specialists were assigned

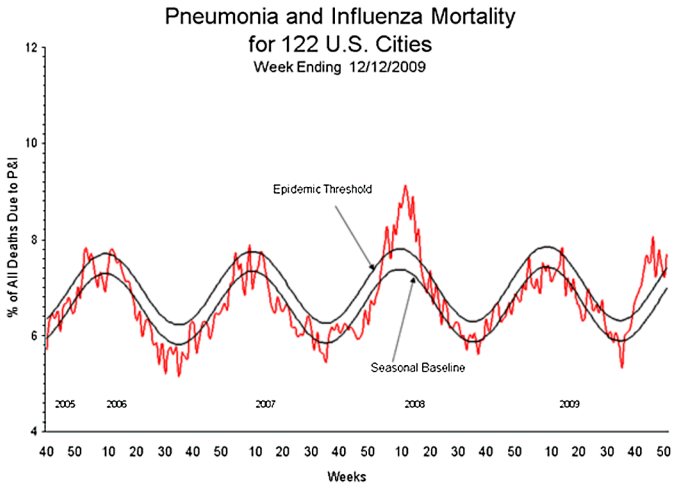

Figure 6. The CDC's national baseline model for determining influenza outbreaks.

the task of identifying outbreaks. According to Siegrist et al. (2005) and as described in Siegrist and Pavlin (2004), the team used three methods to determine "gold standards": documented outbreaks identified by traditional surveillance, visual analysis of the data, and a simple statistical algorithm to identify anomalies in the data. This procedure for algorithm evaluation raises issues surrounding the identification of outbreaks and their dates, the circularity of using statistical guides to help determine outbreaks, and the determination of outbreak-free intervals. A different approach to outbreak-period labeling is to use diagnostic information, such as actual hospital admissions (e.g., Ivanov et al. 2003). Obstacles to this approach are that diagnostic coding is influenced by billing considerations and that classification bias may occur; that is, specific diagnoses may not be coded for earlier cases before laboratory confirmation, but after initial confirmation, diagnoses may be coded freely because of the restrictions, expense, and delay of additional test orders.

3.3 Evaluating Algorithm Performance

One of the major challenges in biosurveillance is that of evaluating and comparing the performance of different algorithms. There are technical reasons for this difficulty, including the lack of data sharing across different research groups and the fact that usually the same group that develops and promotes a method also designs the evaluation criteria for assessing the method's performance, thereby providing an opportunity for scientific confounding (Rolka 2006). But more fundamentally, the unlabeled nature of the data causes serious problems when using standard evaluation methods.

The most widely used evaluation metrics in biosurveillance are *sensitivity* (true positive rate), *specificity* (1 − false-positive rate), and *timeliness*. Objective, replicable quantification presents a challenge for each of these metrics. For sensitivity, a sufficient set of target events is needed for a stable estimate of the fraction of true positives. For specificity, it is difficult to prove the absence of an outbreak to count an alarm as false. Measuring timeliness requires accurate determination of the start of an outbreak event. These three metrics are used to compare different algorithms applied to the same data or the same algorithms applied to data with different outbreak patterns. They also are used to set the alerting thresholds (rather than determining the thresholds theoretically). To set these thresholds, the measures are computed and plotted over a range of alerting threshold values, using receiver operating characteristic (ROC) curves and activity monitor operating characteristic (AMOC) curves. ROC curves show the true-positive rate versus the false-positive rate for a range of threshold values, and the area under the curve measures model accuracy (i.e., the larger the area, the better the model). But because in practice, only part of the ROC curve is of interest, some have suggested using areas under "partial" ROC curves (Kleinman and Abrams 2006). Another serious limitation of ROC curves is that they assume stationary performance over the entire time series. When performance is not stationary (e.g., different performance on weekdays vs. weekends or during summer vs. winter), the ROC curve may be deceptive because of the aggregation of times when sensitivity and specificity are high with times when they are low. This problem may occur in more conventional uses of ROC curves because of unknown bias sources, but particular care is needed when residual bias is considered likely. Finally, sensitivity, specificity, and ROC curves do not incorporate the timeliness aspect. AMOC curves (Fawcett and Provost 1999), which are suitable for activity monitoring and are often used in fraud detection, incorporate the timeliness aspect by plotting a timeliness score versus the false-positive rate. This timeliness score should be defined to correctly discriminate between algorithms. For example, the mean time to detection can be deceptive if some target events are completely missed by the algorithms compared; the more robust median or a more complex measure should be considered. Recently, Kleinman and Abrams (2006) proposed a generalized ROC curve that incorporates timeliness by either weighting each point on the ROC curve based on the mean or median timeliness associated with the corresponding threshold, or by creating three-dimensional ROC curves, with the additional axis representing timeliness.

The concepts of run-length distribution and particularly average run length (ARL), the main evaluation metrics used in statistical quality control, are rare in the biosurveillance literature and practice except for a few statistically oriented papers (e.g., Stoto et al. 2006). Another set of statistical predictive measures that appears in more statistically oriented papers are the root mean squared error (RMSE) and mean absolute percentage error (MAPE) (Reis and Mandl 2003; Burkom, Murphy, and Shmueli 2007). These do not directly measure algorithm performance in terms of alarms, but rather provide a measure for assessing the fit of time series models.

The most popular approach to evaluating practical detection performance in the absence of real bioterror-related outbreaks has been to seed real syndromic data with simulated effects of artificial outbreaks (e.g., Goldenberg et al. 2002; Reis and Mandl 2003; Stoto et al. 2006; Burkom, Murphy, and Shmueli 2007, among many others). The types of outbreak signatures that have been used usually span a short number of days, not because the disease is expected to disappear, but rather because *early* detection is evaluated (i.e., detection after 7 days is considered useless). The shapes of signatures include adding a constant number of cases to a period of a few consecutive days (e.g., Reis and Mandl 2003; Brillman et al. 2005), a linearly increasing number of cases (e.g., Goldenberg et al. 2002;

Stoto et al. 2006), and a lognormal shape (e.g., Burkom, Hutwagner, and Rodriguez 2005a). Usually, different magnitudes of this shape are injected to evaluate the sensitivity of the algorithm to the magnitude of the outbreak. An alternative approach is simulating syndromic data and simulating outbreaks (see the discussion in Buckeridge et al. 2005). A recent project by Lotze et al. (2007) created a simulator that can mimic authentic multivariate syndromic data and inject into them simulated outbreak signatures of different nature.

3.4 Handling Alarms and the Problem of Multiplicity

The third important factor in designing a monitoring scheme has to do with the users of the surveillance systems and how they act on its output. Syndromic surveillance systems are currently used by public health officials, who are mostly epidemiologists by training. They examine daily data and the alarms triggered by the system. In theory, in the event of an alarm, they should trace its cause and determine whether indeed an outbreak is occurring. A specific attempt to do this was made during a pilot period of the syndromic surveillance system used by the New York City Department of Health and Mental Hygiene (NYCDOHMH). During this period, the NYCDOHMH reported weekly false alarms, and for each alarm, experts examined the situation to evaluate the likelihood of an actual outbreak (Fienberg and Shmueli 2005). The NYCDOHMH's procedure for investigating alarms included calling the hospitals that generated alarms, consulting with them as to whether that day's conditions were indicative of an outbreak and if so, whether the emergency department staff was alerted (Paladini 2006). These investigations were found to be difficult, however, and the authors proposed a new investigative protocol.

The current reality is that users of some large systems see alarms nearly every day, because of the large number of data streams and regions monitored and also because the methods are not appropriately adjusted for the variations in time series characteristics among these methods. This frequent alerting leads users to ignore alarms and instead use the system to examine the data heuristically. As one user commented, "we hope that when there is a true alarm, we will not ignore it." One statistical reason for this phenomenon is inadequate handling of the nonstationary raw data, for example, not accounting for the day-of-week effect or ignoring autocorrelation. Another reason is multiple testing. Biosurveillance encounters multiple testing issues at several levels. There are multiple data sources, and within each data source there are usually multiple series (as illustrated in Figure 2), and many of the series are further broken down into subseries (e.g., by age group). All series are currently monitored in a univariate fashion, and, finally, multiple detection algorithms are applied to each series. These multiplicities are currently not well recognized and treated, but their effects are highly visible.

Several methods for handling multiple testing have been proposed in the statistics literature, including familywise Bonferroni-type corrections, false discovery rate (FDR) adjustment (Benjamini and Hochberg 1995) and its variants, and Bayesian methods (e.g., Genovese and Wasserman 2002). Each of the methods has its limitations; for example, Bonferroni is considered overly conservative, FDR corrections depend on the number of hypotheses and are problematic with an insufficient number of hypotheses, and Bayesian methods are sensitive to the choice of prior, and how to choose a prior is unclear. The major problem lies in convincing users that these methods are not masking real outbreaks at a univariate level.

Burkom et al. (2005b) and Rolka et al. (2007) used the terms "parallel monitoring" and "consensus monitoring" to distinguish between the problems of testing a hypothesis in separate populations simultaneously ("parallel," or multiplicity in testing) and of monitoring multiple data sources for testing the hypothesis in a single population ("consensus"). According to this distinction, parallel monitoring occurs when creating source-specific hypotheses that are tested separately (or even when stratifying by, e.g., age group), whereas consensus monitoring occurs when multiple series describing the same population are monitored. These authors proposed using familywise error rates for parallel monitoring and multiple univariate corrections (Edgington 1972) for consensus monitoring; however, robust system performance requires detailed operational guidance for the application of these adjustments given a set of surveillance system requirements. Thus a challenge is to clearly determine the hypotheses and their structure and, based on that, to decide how to integrate the results from the different monitoring algorithms and data streams.

Another implication of univariate monitoring is that there might be a loss of important information carried in the relationships between different syndromic series. For example, an anthrax attack would cause many infected individuals to exhibit flu-like symptoms, excluding nasal congestion. Monitoring the sales of nasal remedies and cough remedies jointly might detect the outbreak earlier than monitoring each individually (where nasal remedies' sales would remain unchanged). Discovering such anomalies can also help indicate the type of outbreak.

Although multivariate monitoring might appear to be a better solution, it carries such challenges as opposing trends in different data streams, lags in reporting (which misalign the different series), large differences in data structure across data sources, absence of knowledge regarding the exact relationship among the outbreak signatures in different series, the changing nature of reporting and IT systems in the different data sources, and in general the additional layer of multivariate nonstationarity. In the context of bioterrorism-related outbreaks, Kleinman and Abrams (2006) raised issues related to the definition of timeliness in the absence of clear outbreak labels, as well as the meaning of sensitivity and specificity in the presence of simulated outbreaks in real data. These issues must be tackled directly to construct performance metrics that are meaningful and accurate within the biosurveillance context.

3.5 Evaluating Overall System Performance

Although there have been many empirical studies of the performance of specific algorithms (e.g., the description of algorithms used in the BioALIRT program and their evaluation, in Buckeridge et al. 2005), there have not been attempts to evaluate the performance of the entire monitoring system. Considering the system as a whole includes multiple algorithms performed on multiple data series. Thus reported performance levels in published empirical studies appear optimistic compared with false-alarm rates in practice. Common acceptable false-alarm rates are much higher than in engineering practices (e.g.,

an average of a false alarm every 2–6 weeks measured in the DARPA BioALIRT program). In some existing systems, however, the overall alert rate is much higher, sometimes with daily statistical anomalies that overwhelm the investigational capability of the monitoring institution. As mentioned earlier, this results not only from inadequate monitoring of raw data (or, equivalently, inadequate adjustment of the monitoring methods), but also from the severe multiple testing of the various data streams, syndrome categories, and monitored subregions on a daily basis. Therefore, a challenge is to move to system-level performance, thereby determining the alarm rates and timeliness of the entire system rather than specific algorithms.

Another issue that is directly related to system performance is cost. Little is known regarding the benefits of operational surveillance systems in terms of their costs or required resources. To design a practical system that yields "reasonable" true and false alarm rates, it is essential to know what is considered reasonable. Eliciting such estimates is difficult, however. Thus, we propose an initial direction of setting a few scenarios of costs and evaluating the different methods, as well as the system, with respect to these scenarios.

In general, to assess the global performance of a system, the definition of "system" must be determined. Components that should be considered part of the system are the collected data (e.g., data quality issues, lag times in data reporting), monitoring algorithms and their related alerting schemes, costs, and the consequences of the alarms to users and perhaps even to decision makers. Without considering all of these aspects (and perhaps others), the real value of the system will remain ambiguous. This issue further underscores the importance of collaboration between the different arms of the system: data providers, developers, implementation teams, users, and decision makers.

4. CURRENT METHODOLOGICAL DEVELOPMENTS

4.1 Temporal Monitoring

The urgency to develop and quickly deploy modern biosurveillance systems has resulted in the use of a small set of monitoring tools and increasing research efforts to improve these tools. Current surveillance systems rely mostly on traditional statistical monitoring methods, such as statistical process control and regression-based methods (e.g., Burkom et al. 2004). The simplicity and familiarity of these methods to the public health community from traditional disease surveillance have led to their continued implementation in the current data environment of multiple prediagnostic data streams at daily and more frequent acquisition rates. In the following sections we describe some of the main monitoring tools currently used in large biosurveillance systems, and then survey new statistical methods proposed in the recent biosurveillance literature.

4.1.1 Methods Used in Practice. Among the monitoring algorithms implemented in current systems, the most widely used are moving average (MA), cusum, and exponentially weighted moving average (EWMA) control charts, as well as Shewhart I-charts applied to regression residuals. We first briefly describe an earlier national system, the Early Aberration Reporting System (EARS; *http://www.bt.cdc.gov/surveillance/ears/*), which has evolved into several new biosurveillance systems. Developed by the CDC (Hutwagner et al. 2003), the analysis portion of EARS comprises adaptations of three control charts: two Shewhart I-charts (called C1 and C2) applied directly to counts, and a truncated cusum chart that accumulates deviations only from the last three days (called C3). The two differences between the EARS charts and standard control charts are aimed at accounting for nonstationarity and detecting a gradual outbreak. The first is the use of a "sliding phase I window" of 7 days for parameter estimation. The second is the use of a "sliding buffer" (Burkom et al. 2004), a window of several days preceding the most recent count that provides a buffer between the data used for parameter estimation (phase I) and that used for alerting (phase II), to avoid contamination of the baseline by a gradual outbreak signal.

These tools have been widely used by public health officials for traditional disease surveillance. But because EARS methods are applied to raw count data, they are more suitable for traditional biosurveillance, where data are commonly aggregated at weekly or longer time scales, analogous to the grouping of classical statistical process control, and thus suffer less from autocorrelation and patterned behavior compared with syndromic data. BioSense initially began with methods based on the EARS system but has recently introduced modifications to adjust the C2 chart for day-of-week effects and for unmodeled features; for example, the W2 method is used to monitor data separately for weekdays and weekends/holidays (see the BioSense bulletin of Sept 30, 2006; *http://www.cdc.gov/biosense/*).

Modern biosurveillance systems use other monitoring tools as well. In particular, ESSENCE uses EWMA charts (applied to the raw data) or Shewhart I-charts applied to residuals from a linear regression model that includes day-of-week, holiday, and postholiday indicators. To determine which of the two types of charts to use, a goodness-of-fit statistic determines whether the regression is useful in explaining the data, and when this test fails, it switches to EWMA. These monitoring schemes also include a 4-week sliding phase-I window and a 1-week sliding buffer. A similar adaptive adjustment (a 28-day sliding phase-I window and a 7-day buffer) was used in the NYCDOHMH system for detecting West Nile virus outbreaks (Mostashari et al. 2003).

Classic control charts are also used in the RODS system with similar adaptations. One of RODS' four monitoring tools is an MA with a 120-day sliding phase-I window (see version 4.2 of the user's manual; *http://rods.health.pitt.edu*); The second tool is a nonstandard combination of cusum and EWMA; EWMA is used to predict next-day counts, and cusum monitors the residuals from these predictions with an empirically chosen threshold of $h = 4.08$. In general, many current algorithms use thresholds determined from empirical experimentation rather than from theoretical design.

The third monitoring tool in RODS is a recursive least squares (RLS) algorithm, which fits an autoregressive model to the counts and continuously updates estimates by minimizing prediction error. A Shewhart I-chart is then applied to the residuals, using a threshold of four standard deviations. This is similar to the regression tool in ESSENCE. A similar approach is used in the CDC's BioSense [called SMART scores, developed by Kleinman, Lazarus, and Platt (2004)], using a Poisson regression of daily counts on the following predictors: (a) a secular (long-term) linear trend over time, (b) sine and

cosine effects for seasonality, (c) month indicators (11 dummies) for nontrigonometric effects of season, (d) day-of-week indicators (six dummies) for day-to-day variability, and (e) holiday and day-after-holiday indicators (*https://btsurveillance.org/btpublic/ri.htm*). Model estimation requires a few months of data for estimating the day-of-week coefficients and a couple of years of data to minimally account for seasonality or monthly effects. The length of daily data streams required for model estimation is a few months for the day-of-week coefficients and a couple of years to minimally account for monthly or seasonal effects. The regression model is then used to predict next-day counts, p values are calculated under the assumption that model residuals are Poisson-distributed, and a SMART score is generated by transforming these p values with a multiple-testing adjustment for the number of geographic subregions considered, the actual count (*http://www.cdc.gov/BioSense/files/CDC_BioSense_User_Guide_VA_DoD_LabCorp_v2.05.pdf*).

The only tool that is not regression-based or a classic control chart applied to the raw data is the WAVELET tool in RODS, which decomposes the time series using Haar wavelets and uses the lowest resolution (the low frequency) to remove long-term trends from the raw series (Zhang et al. 2003). In other words, this is a nonparametric detrending method. The residuals (detrended actual counts) are then monitored using an ordinary Shewhart I-chart with a threshold of four standard deviations.

4.1.2 Methods in the Literature. Aside from the algorithms used within the large systems, there is a growing body of literature on new monitoring algorithms, with empirical studies exploring their performance applied to syndromic data. Because there was no dedicated journal to this field before the recent new journal *Advances in Disease Surveillance*, the literature is dispersed across journals from multiple fields (e.g., medicine, epidemiology, bioinformatics, quality control, public health). A good resource for recent publications is the International Society for Disease Surveillance Web site (*www.syndromic.org*). In what follows we briefly survey some of the proposed methods and the context in which they were tested. The goal of this survey is to demonstrate the relative infancy of modern biosurveillance and to attract further statistical involvement.

Model-Based Approaches. Several efforts have been made to directly model explainable effects (e.g., day-of-week, seasonality, autocorrelation), thereby generating residual series that are approximately iid normal. Regression and ARIMA models have been used for modeling single series; however, ARIMA models are hard to implement in an automated way because of the nonstationary nature of data and its diversity across series. Such fitting requires customized treatment for each series, with the process of preprocessing, identification, and estimation requiring expertise, time, sufficient history, and computational power. An example of this is the work of Reis and Mandl (2003), who used ARIMA models for a single series of 8 years of daily visits to a pediatric hospital. Even if an ARIMA model is applied prospectively, there exists the danger of incorporating gradual outbreaks into the model, thereby masking the outbreak (Reis and Mandl 2003). For this reason, ARIMA models are more likely to serve in retrospective analyses rather in real-time, automated prospective monitoring.

With respect to regression models, linear and Poisson models are currently implemented in several biosurveillance systems using predictors to capture explainable patterns. Several retrospective studies have shown that such models capture these explainable patterns in a variety of syndromic series (e.g., Brillman et al. 2005); however, the main limitation of regression models is the stationarity assumption. Furthermore, modeling long-term patterns requires a long data history, which usually is unavailable or unrepresentative of current behavior due to changes in treatments, coding, population behavior, evolving informatics, faster data rates, changes in reporting practices, and other factors.

Monitoring model residuals has been mostly done univariately, with a few multivariate exceptions. Burkom et al. (2004) used Hotelling T^2 and multivariate cusum and EWMA charts and compared the results with those from multiple univariate charts. Similarly, Stoto et al. (2006) applied Hotelling T^2 and multivariate cusum charts to hospital count data (see also Rolka et al. 2007). Several important issues arise. First, typical raw syndromic data cannot be used directly in standard control charts, because they are far from multivariate normal and independent over time. Second, because the interest is in identifying increases in disease incidences, charts must be modified to be directionally sensitive [as discussed by Hawkins (1991, 1993); Rogerson and Yamada (2004) and implemented by Burkom et al. (2004); Fricker (2006); Joner et al. (2008)]. Third, the cross-covariance structure is assumed to remain constant although empirical evidence indicates a time-varying structure.

Najmi and Magruder (2005) used a different multivariate formulation to explore the relationship between syndromic and clinical data. They used finite impulse response (FIR) filters to predict clinical data multiple steps ahead using OTC sales as well as clinical data. Shmueli and Fienberg (2006) described several other multivariate schemes that are potentially more suitable for syndromic surveillance because they make less-restrictive assumptions about the underlying data and have proven useful in other fields where similar data and goals are encountered. Achieving a balance between simplicity and performance is an important factor, however.

Data-Driven Approaches. Because of the difficulty in finding a parametric "one model fits all" approach that is sufficiently flexible to accept a wide array of nonstationary input series in an automated fashion, an alternative approach is to use data-driven methods for removing explainable patterns to achieve iid normal residual series. Several studies suggested transformations for achieving normality and accounting for the multiplicative day-of-week effects (Brillman et al. 2005; Fricker 2006; Stoto et al. 2006; Burkom, Murphy, and Shmueli 2007). Muscatello (2004) suggested 7-day differencing to remove day-of-week effects.

Data smoothing also has been suggested as a way to remove day-of-week effects. Forsberg et al. (2006) used a moving average with a 7-day window; however, as Siegrist et al. (2005) concluded, "prefilters using 2- to 7-day averages were also tried, and the detection delay defeated any gain from the data smoothing." In other words, averaging across a week actually dampens the signal and can cause delays in detection. Reis and Mandl (2003) accounted for strong weekly and yearly effects by computing stratified averages; however, this requires a long history that is not typically available, and the resulting residuals tend to be highly autocorrelated.

Other smoothing methods include a cosine transform for denoising (Goldenberg et al. 2002), LOWESS for deseasonalizing (Dafni et al. 2004), ratio-to-moving-average indexes for removing seasonality and day-of-week effects (Shmueli 2005; Lotze et al. 2006), and Holt-Winter's exponential smoothing to account for seasonality, trend, and day-of-week effects that change over time (Shmueli and Fienberg 2006; Burkom, Murphy, and Shmueli 2007). The study by Burkom, Murphy, and Shmueli (2007) found that Holt-Winter's exponential smoothing, with some adaptations, outperforms ordinary regression and adaptive regression models, yet is highly automatable.

Some attempts have been made to use wavelet-based methods in biosurveillance. Wavelets are popular in image denoising and compression, and have been used for these purposes in other engineering fields as well (e.g., Jin and Shi 1999). They are computationally efficient and are "general detectors" in the sense of not being tuned to a particular anomaly pattern. They have been used much less frequently in process prediction or monitoring, and in general for prospective tasks, however. Goldenberg et al. (2002) used a redundant spline-based wavelet for decomposing series of OTC medication sales to produce next-day forecasts. This was done by fitting autoregressive models at each of the wavelet scales. As mentioned in the previous section, RODS uses wavelets to detrend data and to remove postholiday dips, by subtracting the low-frequency scale from the original series (Zhang et al. 2003).

Shmueli (2005) proposed a modified scheme of the approach of Aradhye et al. (2003), where coefficients within each scale are thresholded using 3-sigma limits and the original series is then reconstructed from the thresholded scales (thereby highlighting abnormal patterns). Shmueli (2005) discussed the challenges in using wavelet transforms for biosurveillance and the required modifications, including computing wavelet coefficients in a prospective manner, adjusting for multiple testing, and accounting for the dependence structure that arises in redundant wavelets where the downsampling stage is not performed. Lotze et al. (2006) performed a thorough empirical study of this wavelet-based method, including a comparison with regression-based methods. Finally, there are some promising methods for multivariate wavelet-based monitoring in other fields, such as that of Bakshi (1998), that have not yet been explored in biosurveillance.

4.2 Spatial and Spatiotemporal Monitoring

Although the focus of this article is on temporal monitoring, we must mention the complementary area of spatiotemporal monitoring, in which patient, customer, or clinic location information in data records is used to identify localized case clusters. An issue of *Statistics in Medicine* (edited by Lawson, Gangnon, and Wartenberg 2006) was devoted to disease cluster detection. Currently the most widely used method is the spatiotemporal scan statistic (Kulldorff 2001), which searches for statistically significant clusters by comparing daily counts in a certain geographical region with its neighboring regions and with past days. The method is based on computing a likelihood ratio-based statistic (assuming a Bernoulli or Poisson model) and using randomization to obtain p values.

The main focuses of subsequent research have been (a) comparing the empirical performance of the spatiotemporal scan statistic with other algorithms (e.g., Kulldorff, Tango, and Park 2003; Kedem and Wen 2007), (b) improving the scan statistic in terms of computational time [see, e.g., Neill, Moore, and Cooper (2006), who developed a faster Bayesian alternative that does not require randomization], and (c) improving the scan statistic's ability to treat more general cluster shapes than circular forms to define a geographical region (e.g., Kulldorff et al. 2006).

A common practical obstacle is that the spatial distribution of syndromic data does not agree with census or other general population distributions. Reasons for this disagreement include the locations of care providers or clinics available in a data set, unknown or changing catchment area of a health maintenance organization or pharmacy chain, and varying patterns of health care utilization among neighboring demographic groups. Thus an important challenge that requires attention is estimation of the spatial background distribution. Kleinman et al. (2005) demonstrated that the rate of significant cluster determination may be reduced and irrelevant/nuisance clusters avoided by modeling data features, such as seasonality and day-of-week effects, to improve this estimation, and that further work is needed to determine appropriate estimation procedures for various data sets. Another drawback to the identification of meaningful spatial clusters is that the data record field most often used for case geolocation is patient residence address. Because data sets that provide a work address are rare, workplace-based clusters are not likely to be found using space-based algorithms like scan statistics. The workflow scan statistic of Duczmal and Buckeridge (2006) provides an approach to using demographic data to find these clusters; more such work is needed to improve the utility of available data for cluster detection. A neglected but related and important additional challenge is that of determining whether attributable cases in identified clusters are linked and worthy of investigation; clearing this hurdle will require close cooperation among care providers, informaticists, and statisticians.

Spatiotemporal monitoring also involves many of the temporal monitoring challenges, such as multiplicity (Kulldorff et al. 2007; Rolka et al. 2007) and performance evaluation (Kleinman et al. 2006).

5. CONCLUSIONS AND FUTURE DIRECTIONS

The goals of this article are to introduce the important area of modern biosurveillance and to enumerate the challenges that it poses to traditional statistical monitoring. Statisticians have had little influence in the design, implementation, or evaluation of operational systems, and there is a pressing need to develop improved biosurveillance systems. Opportunities exist for developing statistical methodologies for improved monitoring and evaluation of biosurveillance systems. Multiple components make biosurveillance challenging statistically. First, syndromic data are less specific but arguably timelier than exact diagnoses for detecting disease outbreaks, and filtering the data records to maximize the signal-to-noise ratio is an ongoing challenge requiring elicitation of imprecise and often intuitive medical domain expertise. Second, for the time series and other data objects derived for routine monitoring, conventional data

assumptions of statistical process control, such as temporal independence and stationarity, are commonly violated. This challenge is not specific to biosurveillance and is apparent in chemical processes, geophysical data, and other areas. Third, the outbreak data signature depends on both the characteristics of the underlying pathogen, such as the distribution of incubation periods and the outbreak symptomatology, and the data source details, such as coding practices and recording delays. Outbreak signature shapes are useful only in scenario-based surveillance. In the absence of syndromic data that contain bioterrorism-related outbreaks, the task is one of anomaly detection rather than of signature identification. Fourth, the lack of labeled data that arises from the ambiguity of outbreak definitions and periods is a serious obstacle to evaluating system performance. The implications are a lack of proper phase-I data. When the goal is to detect bioterrorism-related outbreaks, we can (luckily) assume that the data are clean of effects of such attacks, but the presence of natural outbreaks creates more background noise, which is difficult to model if not specified as an outbreak. Thus, the current approach to evaluation is to seed real data with artificial outbreaks. Although this approach provides simulated events to detect, the ambiguity of the presence of additional true outbreaks remains. Furthermore, outbreak simulation is challenging, because of the unknown nature of an outbreak signature in syndromic data. Therefore, simulating a certain type of outbreak intrinsically determines the most efficient monitoring algorithm (e.g., a cusum for detecting a small step function change). Finally, the issue of multiplicity in testing raises serious questions that should be carefully addressed.

An important, seemingly nonstatistical, challenge is the actual use of biosurveillance systems by public health officials. The current disconnect among algorithm developers, implementers, and users has led to systems with uncontrolled alert rates that depend on various epidemiologic and informatics issues. Such experiences foster distrust in statistical monitoring and in biosurveillance itself. It is our responsibility to use all available statistical ammunition to create adequate yet simple methods that will aid expert decision making rather than confuse the user.

All of these issues highlight the similarity between bioterrorism-related outbreak detection and other event detection tasks, such as fraud detection in accounting (e.g., Bay et al. 2006) and network security and intrusion detection. In such tasks, events tend to be rare; detecting them can have a significant impact, the signature of an event is hard to define (and new types of events constantly evolve), and evaluating algorithms is difficult (Dash et al. 2006). Another similarity when considering bioterrorism is the presence of an event generator that is aware of the monitoring system and tries to game it. In contrast, when considering natural disease outbreaks, the task is closer to software monitoring or traffic incident monitoring via sensors (Singliar and Hauskrecht 2006), where many events are present but their exact time is difficult to determine. Moreover, the event generator is "innocent" in the sense of not trying to game the monitoring system.

Current temporal biosurveillance practice relies on heuristic adaptations of classical control charts applied mostly to raw count data. For the aforementioned reasons, we believe that these tools are not always adequate for the purpose and requirements of biosurveillance. However, the biosurveillance literature contains methods and adaptations that have been shown empirically to outperform current practice. In addition, the literature contains new methods that are likely to improve performance but currently are not applied directly to modern biosurveillance [e.g., hidden Markov models for influenza surveillance by Ozonoff and Sebastiani (2006), temporal scan statistics for monitoring weekly national reports of brucellosis by Wallenstein and Naus (2004), neural networks for multivariate health surveillance by Adams, Saithanu, and Hardin (2006), and the moving-F chart not affected by reduction in variability by Riffenburgh and Cummins (2006)]. The lack of application to modern biosurveillance calls attention to one of the main barriers to statistical involvement in this field: data access. Currently syndromic data are available only to researchers affiliated with a particular biosurveillance system or research group, for reasons of data confidentiality and nondisclosure agreements. This is a major obstacle to scientific progress in both temporal and spatiotemporal biosurveillance. Hopefully some data will be made available to academic researchers. Various attempts have been made to draw statisticians to biosurveillance research, such as the work group on Anomaly Detection in National Defense and Homeland Security by the Statistical and Applied Mathematics Sciences Institute (SAMSI, *http://sisla06.samsi.info/ndhs/ad/*) and the growing number of conference sessions in statistics conferences devoted to biosurveillance. A second barrier to entry into this research field is the dispersion of the relevant literature across journals in a variety of fields, along with the lack of detailed descriptions of methods implemented in practice. We hope that this survey sheds light on these aspects. We have especially attempted to spotlight the various statistical challenges that leave room for contributions in this important emerging field.

In this article we have concentrated on statistical challenges to early detection in biosurveillance using aggregated syndromic time series. The obstacles discussed range from data acquisition and quality issues to adaptations and combinations of classical methods from related disciplines. Looking beyond these challenges, the experience of the first 10 years of automated biosurveillance systems has led to a shift away from early detection based solely on aggregated syndromic data. This shift has resulted in part from a perception that in most situations, syndromic data alone do not contain sufficient information to justify a resource-intensive public health response. Instead, a question widely asked in the public health community is how to combine aggregated syndromic data with more individual-based diagnostic data for general situational awareness, as well as for early detection and corroboration. Thus, the problem of combining evidence from disparate data sources is crucial. It involves dealing with differences relevant to the threat of interest, the relative timeliness of the signal, the reliability of the data, and many other factors. But as electronic health record data become more available and the fusion of evidence types becomes more realistic, robust improvements in surveillance will depend on effectively managing these obstacles. Will multivariate versions of the aforementioned methods be made sufficiently robust for routine health monitoring? Will less transparent machine learning approaches to data fusion gain acceptance? Acceptance of any analytical fusion of evidence will require a systematic treatment of the evaluation of issues discussed earlier.

Close collaboration of statisticians with computer scientists and especially with the public health practice community will be essential to the necessary advances and their successful implementation.

ACKNOWLEDGMENTS

The authors thank the following persons for invaluable comments and suggestions that greatly improved this manuscript: Sean Murphy from the Johns Hopkins Applied Physics Laboratory, Dr. Ken Kleinman from Harvard Medical School and Harvard Pilgrim Health Care, Dr. Bill Woodall from Virginia Tech, Dr. Ron Fricker from the Naval Postgraduate School, Dr. Karen Kafadar from Indiana University, and the associate editor and three anonymous referees. Permission to use the data was obtained through data use agreement 189 from TRICARE Management Activity. The work was partially supported by National Institutes of Health Grant RFA-PH-05-126.

[Received November 2006. Revised April 2009.]

REFERENCES

Adams, B. M., Saithanu, K., and Hardin, J. M. (2006), "A Neural Network Approach to Control Charts With Applications to Health Surveillance," invited talk at *the 2006 Joint Statistical Meeting*, Seattle, WA. [49]

Aradhye, H. B., Bakshi, B. R., Strauss, R. A., and Davis, J. F. (2003), "Multiscale Statistical Process Control Using Wavelets—Theoretical Analysis and Properties," *AIChE Journal*, 49 (4), 939–958. [48]

Bakshi, B. R. (1998), "Multiscale PCA With Application to Multivariate Statistical Process Monitoring," *AIChE Journal*, 44, 1596–1610. [48]

Bay, S., Kumaraswamy, K., Anderle, M., Kumar, R., and Steier, D. (2006), "Large Scale Detection of Irregularities in Accounting Data," in *6th International Conference on Data Mining (ICDM)*, IEEE, pp. 75–86. [49]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300. [45]

Benneyan, J. C. (1998), "Statistical Quality Control Methods in Infection Control and Hospital Epidemiology, Part I: Introduction and Basic Theory," *Infection Control and Hospital Epidemiology*, 19, 194–214. [41]

Brillman, J. C., Burr, T., Forslund, D., Joyce, E., Picard, R., and Umland, E. (2005), "Modeling Emergency Department Visit Patterns for Infectious Disease Complaints: Results and Application to Disease Surveillance," *BMC Medical Informatics and Decision Making*, 5 (4), 1–14. [41,44,47]

Brookmeyer, R., Johnson, E., and Barry, S. (2005), "Modelling the Incubation Period of Anthrax," *Statistics in Medicine*, 24 (4), 531–542. [42]

Buckeridge, D. L., Burkom, H., Campbell, M., Hogan, W. R., and Moore, A. W. (2005), "Algorithms for Rapid Outbreak Detection: A Research Synthesis," *Journal of Biomedical Informatics*, 38, 99–113. [43,45]

Burkom, H. S., Elbert, Y., Feldman, A., and Lin, J. (2004), "Role of Data Aggregation in Biosurveillance Detection Strategies With Applications From Essence," *Morbidity and Mortality Weekly Report*, 53 (Suppl.), 67–73. [46,47]

Burkom, H. S., Hutwagner, L., and Rodriguez, R. (2005a), "Using Point-Source Epidemic Curves to Evaluate Alerting Algorithms for Biosurveillance," in *2004 Proceedings of the Statistics in Government Section, Toronto*, American Statistical Association [CD-ROM]. [42,45]

Burkom, H. S., Murphy, S., Coberly, J., and Hurt-Mullen, K. (2005b), "Public Health Monitoring Tools for Multiple Data Streams," *Morbidity and Mortality Weekly Report*, 54 (Suppl.), 55–62. [45]

Burkom, H. S., Murphy, S. P., and Shmueli, G. (2007), "Automated Time Series Forecasting for Biosurveillance," *Statistics in Medicine*, 26 (2), 4202–4218. [41,44,47,48]

Dafni, U. G., Tsiodras, S., Panagiotakos, D., Gkolfinopoulou, K., Kouvatseas, G., Tsourti, Z., and Saroglou, G. (2004), "Algorithm for Statistical Detection of Peaks—Syndromic Surveillance System for the Athens 2004 Olympic Games," *Morbidity and Mortality Weekly Report*, 53 (Suppl.), 86–94. [48]

Dash, D., Lane, T., Margineantu, D., and Wong, W.-K. (2006), "Opening Remarks," in *Workshop on Machine Learning Algorithms for Surveillance and Event Detection, 23rd International Conference on Machine Learning*, Pittsburgh, PA. [49]

Duczmal, L., and Buckeridge, D. L. (2006), "A Workflow Spatial Scan Statistic," *Statistics in Medicine*, 25 (5), 743–754. [48]

Edgington, E. S. (1972), "A Normal Curve Method for Combining Probability Values From Independent Experiments," *Journal of Psychology*, 82, 85–89. [45]

Fawcett, T., and Provost, F. (1999), "Activity Monitoring: Noticing Interesting Changes in Behavior," in *5th ACM SIGKDD International Conference*, New York: ACM, pp. 53–62. [44]

Fienberg, S. E., and Shmueli, G. (2005), "Statistical Issues and Challenges Associated With Rapid Detection of Bio-Terrorist Attacks," *Statistics in Medicine*, 24 (4), 513–529. [40-42,45]

Forsberg, L., Jeffery, C., Ozonoff, A., and Pagano, M. (2006), "A Spatio-Temporal Analysis of Syndromic Data for Biosurveillance," in *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, Springer. [47]

Fricker, R. D., Jr. (2006), "Directionally Sensitive Multivariate Statistical Process Control Methods With Application to Syndromic Surveillance," *Advances in Disease Surveillance*, 3, 1. [41,47]

Genovese, C., and Wasserman, L. (2002), "Operating Characteristics and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society, Ser. B*, 64, 499–451. [45]

Goldenberg, A., Shmueli, G., Caruana, R. A., and Fienberg, S. E. (2002), "Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales," *Proceeding of the National Academy of Sciences*, 99, 5237–5240. [41,42,44,48]

Hawkins, D. M. (1991), "Multivariate Quality Control Based on Regression-Adjusted Variables," *Technometrics*, 33, 61–75. [47]

——— (1993), "Regression Adjustment for Variables in Multivariate Quality Control," *Journal of Quality Technology*, 25, 170–182. [47]

Hutwagner, L., Thompson, W., Seeman, G., and Treadwell, T. (2003), "The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS)," *Journal of Urban Health*, 80 (2) (Suppl.), 89–96. [46]

Ivanov, O., Gesteland, P. H., Hogan, W., Mundorff, M. B., and Wagner, M. M. (2003), "Detection of Pediatric Respiratory and Gastrointestinal Outbreaks From Free-Text Chief Complaints," in *AMIA Annual Fall Symposium*, Madison, WI: Omni Press, pp. 318–322. [44]

Jin, J., and Shi, J. (1999), "Feature-Preserving Data Compression of Stamping Tonnage Information Using Wavelets," *Technometrics*, 41 (4), 327–339. [48]

Joner, M. D., Jr., Woodall, W. H., Reynolds, M. R., Jr., and Fricker, R. D. (2008), "A One-Sided MEWMA Chart for Health Surveillance," *Quality and Reliability Engineering International*, 24 (5), 503–518. [47]

Kedem, B., and Wen, S. (2007), "Semi-Parametric Cluster Detection," *Journal of Statistical Theory and Practice*, 1 (1), 49–72. [48]

Kleinman, K. P., and Abrams, A. M. (2006), "Assessing Surveillance Using Sensitivity, Specificity and Timeliness," *Statistical Methods in Medical Research*, 15 (5), 445–464. [44,45]

Kleinman, K., Abrams, A., Kulldorff, M., and Platt, R. (2005), "A Model-Adjusted Space-Time Scan Statistic With an Application to Syndromic Surveillance," *Epidemiology and Infection*, 133, 409–419. [48]

Kleinman, K., Abrams, A., Yih, W. K., Platt, R., and Kulldorff, M. (2006), "Evaluating Spatial Surveillance: Detection of Known Outbreaks in Real Data," *Statistics in Medicine*, 25 (5), 755–769. [48]

Kleinman, K., Lazarus, R., and Platt, R. (2004), "A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, With an Application to Biological Terrorism," *American Journal of Epidemiology*, 159, 217–224. [46]

Kulldorff, M. (2001), "Prospective Time-Periodic Geographical Disease Surveillance Using a Scan Statistic," *Journal of the Royal Statistical Society, Ser. A*, 164 (1), 61–72. [48]

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006), "An Elliptic Spatial Scan Statistic," *Statistics in Medicine*, 25 (22), 3929–3943. [48]

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W., Kleinman, K., and Platt, R. (2007), "Multivariate Scan Statistics for Disease Surveillance," *Statistics in Medicine*, 26 (8), 1824–1833. [48]

Kulldorff, M., Tango, T., and Park, P. J. (2003), "Power Comparisons for Disease Clustering Tests," *Computational Statistics & Data Analysis*, 42 (4), 665–684. [48]

Lawson, A., Gangnon, R., and Wartenberg, D. (2006), "Special Issue on Developments in Disease Cluster Detection," *Statistics in Medicine*, 25 (5), 721–916. [48]

Lotze, T., Shmueli, G., Murphy, S., and Burkom, H. (2006), "A Wavelet-Based Anomaly Detector for Early Detection of Disease Outbreaks," in *Workshop on Machine Learning Algorithms for Surveillance and Event Detection, 23rd International Conference on Machine Learning*, Pittsburgh, PA. [48]

Lotze, T., Shmueli, G., and Yahav, I. (2007), "Simulating Multivariate Syndromic Time Series and Outbreak Signatures," Technical Report RHS-06-054, Smith School of Business, University of Maryland. [45]

Meselson, M., Guillemin, J., Hugh-Jones, M., Langmuir, A., Popova, I., Shelokov, A., and Yampolskaya, O. (1994), "The Sverdlovsk Anthrax Outbreak of 1979," *Science*, 266 (5188), 1202–1208. [42]

Mostashari, F., Kulldorff, M., Hartman, J., Miller, J., and Kulasekera, V. (2003), "Dead Bird Clusters as an Early Warning System for West Nile Virus Activity," *Emerging Infectious Diseases*, 9, 641–646. [46]

Muscatello, D. (2004), "An Adjusted Cumulative Sum for Count Data With Day-of-Week Effects: Application to Influenza-Like Illness," presentation at *Syndromic Surveillance Conference*, Boston, MA. [47]

Najmi, A., and Magruder, S. (2005), "An Adaptive Prediction and Detection Algorithm for Multistream Syndromic Surveillance," *BMC Medical Informatics and Decision Making*, 12, 5–33. [47]

Neill, D. B., Moore, A. W., and Cooper, G. F. (2006), "A Bayesian Spatial Scan Statistic," in *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, pp. 1003–1010. [48]

Ozonoff, A., and Sebastiani, P. (2006), "Hidden Markov Models for Prospective Surveillance," presented at *The Anomaly Detection Group in National Defense and Homeland Security*, SAMSI, Research Triangle Park, NC. [49]

Paladini, M. (2006), "From Data to Signals to Screenshots: Recent Developments in Nycdohmh Emergency Department Syndromic Surveillance," presentation at *DIMACS Working Group on BioSurveillance Data Monitoring and Information Exchange*, Piscataway, NJ, available at *http://dimacs.rutgers.edu/Workshops/Surveillance/slides/paladini.ppt*. [45]

Pavlin, J. A. (1999), "Epidemiology of Bioterrorism," *Emerging Infectious Diseases*, 5, 528–565. [42]

Reis, B., and Mandl, K. (2003), "Time Series Modeling for Syndromic Surveillance," *BMC Medical Informatics and Decision Making*, 3 (2). [41,44,47]

Riffenburgh, R., and Cummins, K. (2006), "A Simple and General Change-Point Identifier," *Statistics in Medicine*, 25 (6), 1067–1077. [49]

Rogerson, P. A., and Yamada, I. (2004), "Monitoring Change in Spatial Patterns of Disease: Comparing Univariate and Multivariate Cumulative Sum Approaches," *Statistics in Medicine*, 23 (14), 2195–2214. [47]

Rolka, A., Burkom, H., Cooper, G. F., Kulldorff, M., Madigan, D., and Wong, W.-K. (2007), "Issues in Applied Statistics for Public Health Bioterrorism Surveillance Using Multiple Data Streams: Research Needs," *Statistics in Medicine*, 26 (8), 1834–1856. [42,45,47,48]

Rolka, H. (2006), "Emerging Public Health Biosurveillance Directions," in *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, New York: Springer, pp. 101–107. [41,44]

Shmueli, G. (2005), "Wavelet-Based Monitoring for Modern Biosurveillance," Technical Report RHS-06-002, University of Marlyand, Robert H. Smith School of Business. [48]

Shmueli, G., and Fienberg, S. E. (2006), "Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Bio-Surveillance," in *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, New York: Springer, pp. 109–140. [47,48]

Siegrist, D., and Pavlin, J. (2004), "Bio-Alirt Biosurveillance Detection Algorithm Evaluation," *Morbidity and Mortality Weekly Report*, 53 (Suppl.), 152–158. [43,44]

Siegrist, D., McClellan, G., Campbell, M., Foster, V., Burkom, H., Hogan, W., Cheng, K., Buckeridge, D., Pavlin, J., and Kress, A. (2005), "Evaluation of Algorithms for Outbreak Detection Using Clinical Data From Five U.S. Cities," technical report, DARPA Bio-ALIRT Program. [44,47]

Singliar, T., and Hauskrecht, M. (2006), "Towards a Learning Traffic Incident Detection System," in *Workshop on Machine Learning Algorithms for Surveillance and Event Detection, 23rd International Conference on Machine Learning*, Pittsburgh, PA. [49]

Stoto, M., Fricker, R. D., Jain, A., Davies-Cole, J. O., Glymph, C., Kidane, G., Lum, G., Jones, L., Dehan, K., and Yuan, C. (2006), "Evaluating Statistical Methods for Syndromic Surveillance," in *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, New York: Springer, pp. 141–172. [44,45,47]

Wagner, M. M., Tsui, F. C., Espino, J. U., Dato, V. M., Sittig, D. F., Caruana, R. A., McGinnis, L. F., Deerfield, D. W., Druzdzel, M. J., and Fridsma, D. B. (2001), "The Emerging Science of Very Early Detection of Disease Outbreaks," *Journal of Public Health Management and Practice*, 7, 51–59. [42]

Wallenstein, S., and Naus, J. (2004), "Scan Statistics for Temporal Surveillance for Biologic Terrorism," *Morbidity and Mortality Weekly Report*, 53 (Suppl.), 74–78. [49]

Woodall, W. H. (2006), "The Use of Control Charts in Health-Care and Public-Health Surveillance," *Journal of Quality Technology*, 38 (2), 89–104. [41]

Yih, W., Caldwell, B., Harmon, R., Kleinman, K., Lazarus, R., Nelson, A., Nordin, J., Rehm, B., Richter, B., Ritzwoller, D., Sherwood, E., and Platt, R. (2004), "The National Bioterrorism Syndromic Surveillance Demonstration Program," *Morbidity and Mortality Weekly Report*, 53 (Suppl.), 43–49. [40]

Zhang, J., Tsui, F., Wagner, M., and Hogan, W. (2003), "Detection of Outbreaks From Time Series Data Using Wavelet Transform," in *AMIA Annual Symposium*, Madison, WI: Omni Press, pp. 748–752. [47,48]