Statistical Challenges in Modern Biosurveillance

Galit Shmueli

Department of Decision & Information Technologies

and The Center for Health Information and Decision Systems

Robert H. Smith School of Business

University of Maryland, College Park, MD 20742

Howard Burkom

The Johns Hopkins University Applied Physics Laboratory, MD

Abstract

Modern biosurveillance is the monitoring of a wide-range of pre-diagnostic 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 have also 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; data labeling is ambiguous in the sense that outbreak periods are not clearly defined or known. Additional challenges arise such as multiplicity in several dimensions, performance evaluation, and practical system usage and requirements. Our focus is mainly on the monitoring of time series for early alerting 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 different statistical challenges in monitoring modern biosurveillance data, describe the current state of monitoring in the field, and survey the most recent biosurveillance literature.

Keywords: disease outbreak; syndromic data; statistical process control; control charts; anomaly detection.

1 Introduction: Modern Biosurveillance Systems

Biosurveillance is the practice of monitoring data for the purpose of detecting disease outbreaks. Traditional biosurveillance has focused on the collection and monitoring of diagnostic medical and public health data retrospectively for determining 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, and even months after the outbreak. By the time this information reaches decision makers it is often too late to treat the infected population or to react in other ways such as stockpiling and dispensing vaccine and medication.

In the last several years, there has been a shift towards biosurveillance systems that would provide early detection of diseases, resulting either from bioterrorist attacks or from "natural" causes such as the avian flu. 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 pre-diagnostic information such as over-the-counter (OTC) and pharmacy medication sales, calls to nurse hotlines, school absence records, web-searches on medical websites, and complaints of individuals entering hospital emergency departments. None of these directly measures 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 their "earliness". Under the assumption that people tend to self-treat and self-medicate before rushing to the hospital, we expect web searching and the purchasing of OTC remedies to precede calls to nurse hotlines and ambulance dispatches, and then to be followed by emergency department visits. Still, this entire continuum is assumed to occur before actual clinical diagnoses can be made (after hospitalization and/or lab tests). In addition to monitoring syndromic data there have been efforts to monitor other types of data that are associated with disease risk factors, such as air and water quality measurements. Although all these evidence sources fall under biosurveillance, we focus on the types of data that are currently being used in biosurveillance systems (Note: The term *syndromic surveillance* was used earlier in the field, but because *syndrome* here has a different meaning than its medical or English use, there has been a movement towards other definitions such as biosurveillance or disease surveillance.).

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 US surveillance systems routinely collect data from multiple sources on a daily basis, and these data are transferred with variable delay to the biosurveillance systems (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 US city. In addition to this mortality series, six additional measures are tracked, all based either on mortality or on 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. pre-diagnostic 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.

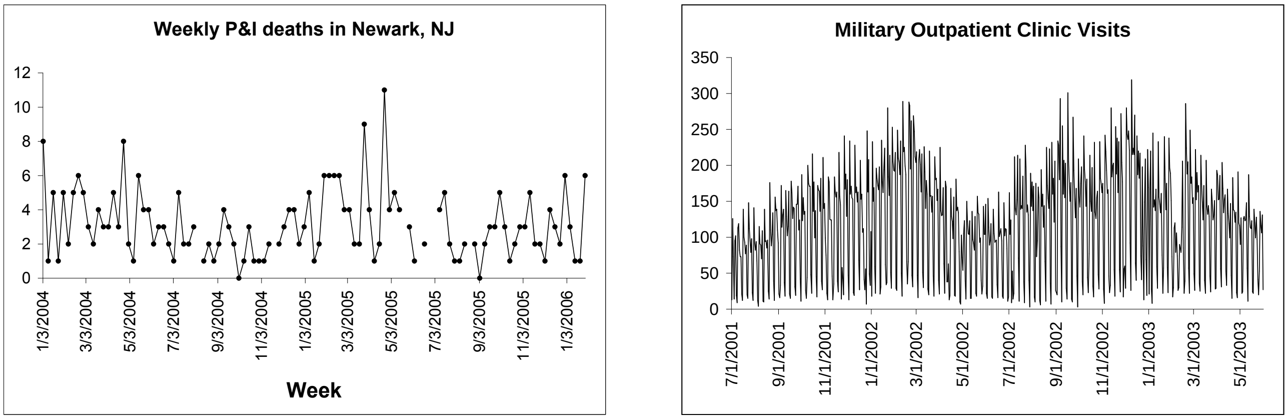


Figure 1: Typical traditional data (left) vs. syndromic data (right), for monitoring influenza.

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 US systems serve a wide geographical region and there is a large and growing number 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. Non-clinical data include over-the-counter 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 classifying these records' textual chief complaints into syndrome groupings. A patient may report several symptoms (e.g., rash and fever), thereby contributing multiple chief complaints. For billing purposes, the emergency departments themselves classify patient records into ICD-9 codes, though 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, etc.

In general, data modeling efforts have gone in two directions: modeling temporal data within a

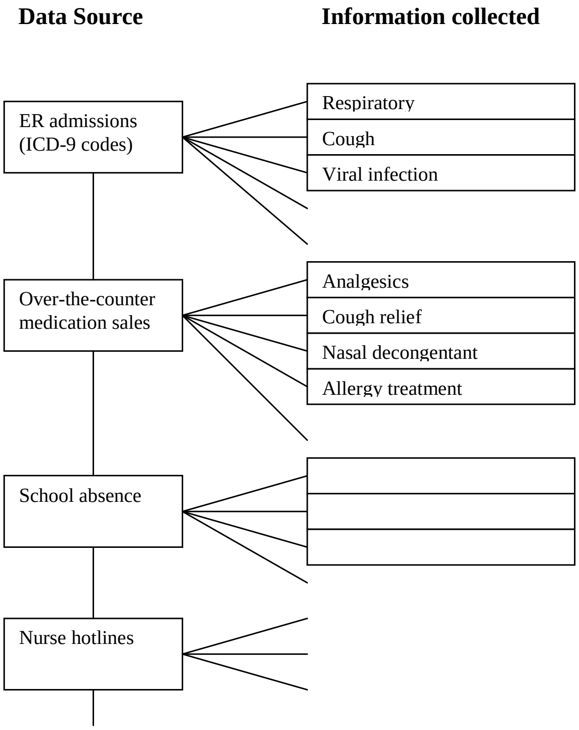


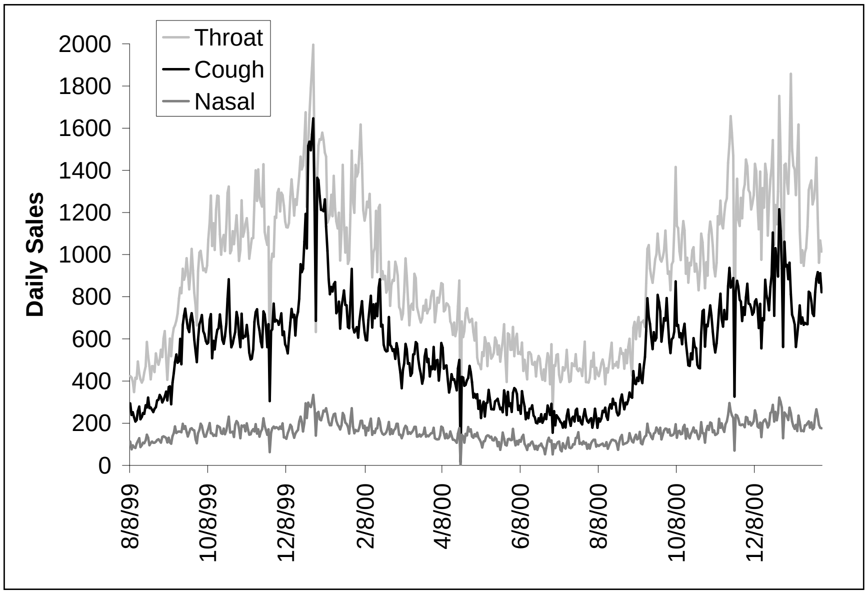
Figure 2: Sketch of data hierarchy: each data source can contain multiple time series

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. In the former, the setting is similar to statistical quality control where one or more streams of data are inspected for abnormalities prospectively. In the spatial (or spatio-temporal) applications, methods are aimed at detecting regions whose case distribution is abnormally high compared to that of other regions. Focusing on the temporal approach, for a particular geographical location we can think of 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 or a flat model is more suitable.

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 other purposes (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 non-stationarity. Means, variances, and autocorrelation structures tend to change over time, and the degree of non-stationarity changes from series to series. However, these time series display a few general predictable patterns, such as characteristic day-of-week (DOW) behavior. In US emergency department visits, daily counts are typically low on weekends and high early in the work week (Burkom et al., 2007), but can also exhibit other daily patterns (e.g., Brillman et al., 2005; Reis and Mandl, 2003), or none (Fricker, 2006). On weekends, grocery stores tend to have more traffic, and therefore 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 non-negligible short-term autocorrelation. During 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



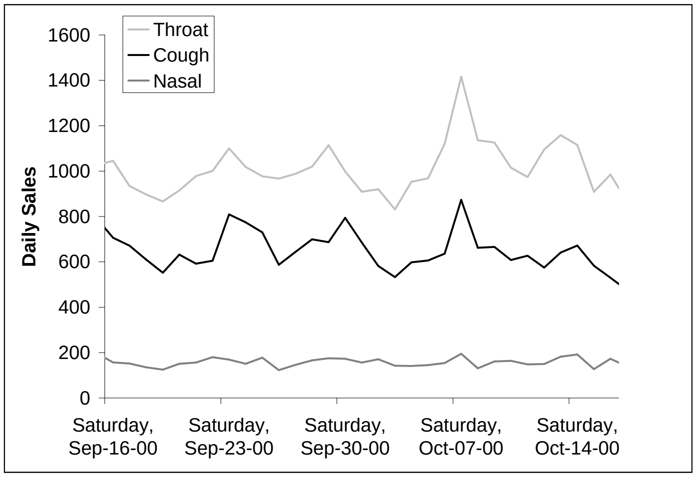


Figure 3: Daily sales of over-the-counter medications from a large grocery chain in the Pittsburgh, Pennsylvania are, by medication subgroup. Bottom panel is a zoom-in on a one-month period. Two of the series exhibit strong bi-annual seasonality, and all series exhibit some level of day-of-week effect. Dips on holidays are due to store closings.

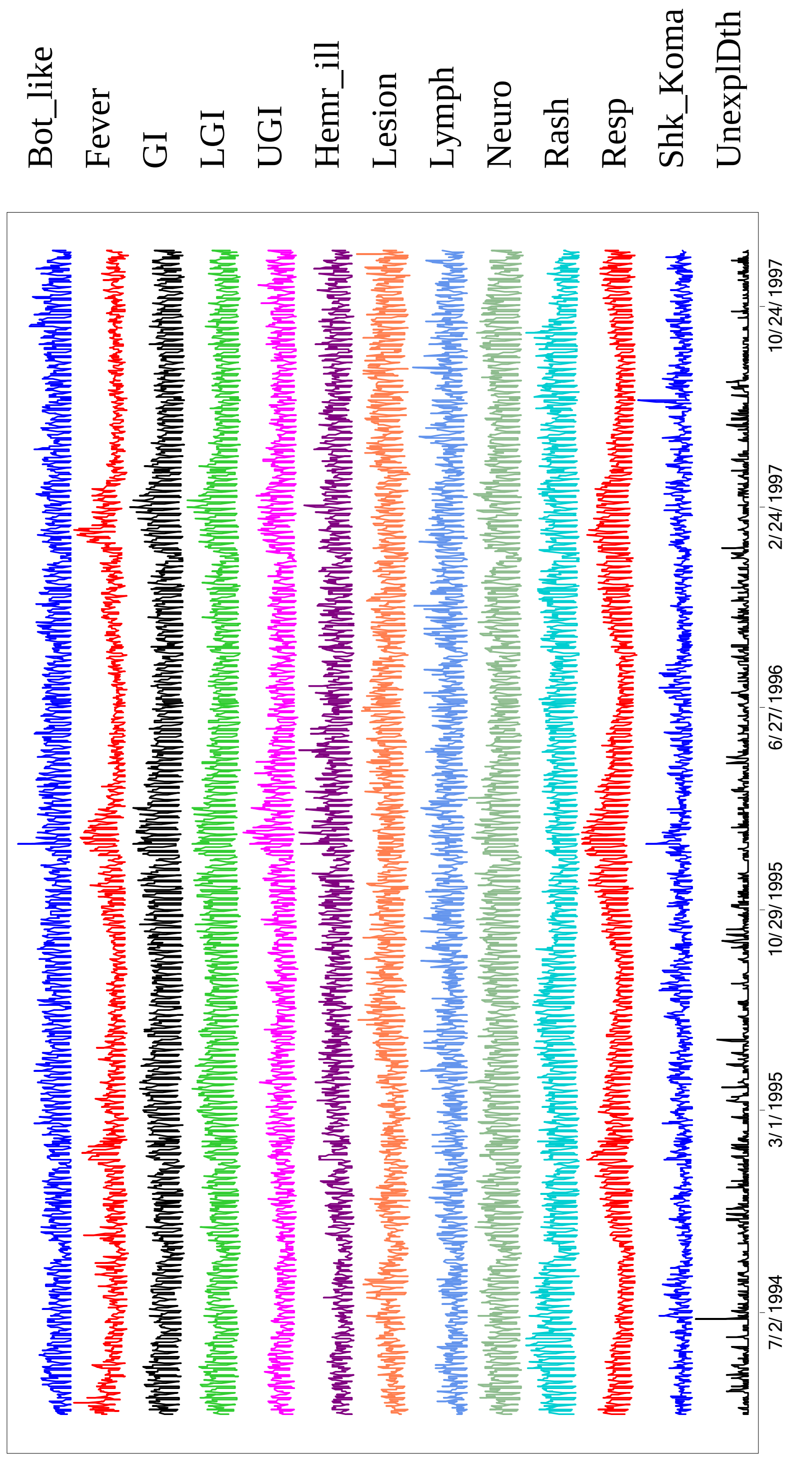


Figure 4: Daily counts of chief complaints at emergency departments in a certain US city, by type of complaint. The first 12 series combine counts of ICD-9 diagnosis codes relevant to a certain syndrome, using the CDC's list of syndrome definitions. The last series is counts of unexplained death.

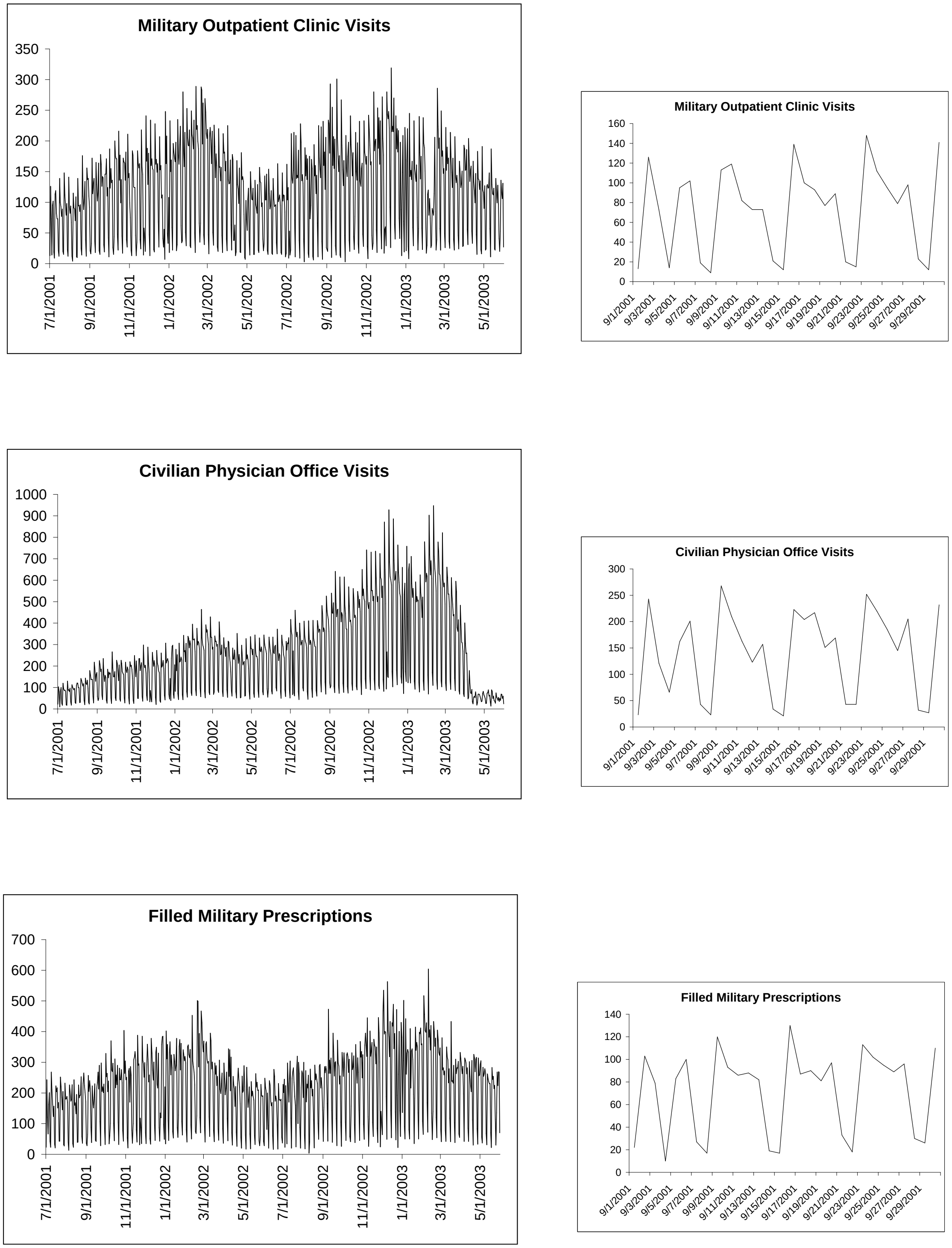


Figure 5: Daily respiratory-related counts in a certain US city from three data sources: Visits to military outpatient clinics (top), civilian physician visits (middle) , and filled military prescriptions (bottom). Entire 700-day period (left) and single month (right). These are part of a larger dataset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

setting is different than the industrial setting in terms of the nature of the collected data, the underlying "normal" process, the nature of an outbreak, how performance is evaluated, and the requirements and uses of a biosurveillance system. We discuss each of these next.

3.1 Determining and Modeling "Normal" Behavior

Determining whether an abnormality is present in the data requires defining normal behavior. One complication arises from the intended dual-use of biosurveillance systems for detecting natural and bioterror-related disease outbreaks, where "normal" has different meanings. In the bioterror-outbreak case all data are considered devoid of outbreaks and "normal" behavior includes natural outbreaks. In contrast, for natural outbreaks "normal" behavior is not straightforward to define or to model. It is nearly impossible to obtain exact dates of local natural disease outbreaks. 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 going to a doctor), the delay times from infection to these behaviors and delays between these behaviors are hard to quantify. Finally, the population being monitored is very dynamic. Changes in population, data reporting, hospital policies, and other factors lead to a non-stationary, constantly evolving "normal" behavior that is not easy to model using standard techniques. All this leads 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 we can design the most effective filter to detect it. In biosurveillance there exists knowledge about the progression of different diseases in the population using theoretical disease epicurve models (e.g., Burkom et al., 2005b) or by estimating it from historic data such as the accidental anthrax release in Sverdlovsk, Russia in 1979 (Meselson et al., 1994; Goldenberg et al., 2002; Brookmeyer et al., 2005).

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 bio-terrorist related disease and that from a natural disease. However, there has been very little discussion of the expected signature in non-clinical data and especially nontraditional data. For example, the manifestation of an anthrax attack in ambulance dispatches or in sales of cough remedies is yet unknown. This next step requires inputs from medical and public health experts as well as domain experts such as marketers. Such an approach is described in 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 non-specific 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 non-infectious.

Another major challenge arises from the difficulty in attaining properly labeled data with exact outbreak periods. This challenge arises also in traditional disease surveillance, where the onset of a local outbreak is hard to pinpoint. For example, for influenza the 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.

National outbreaks are then determined using this baseline (see Figure 6). However, 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 on different years; And lastly, 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 by The Defense Advanced Research Projects Agency (DARPA) was aimed at evaluating different algorithms

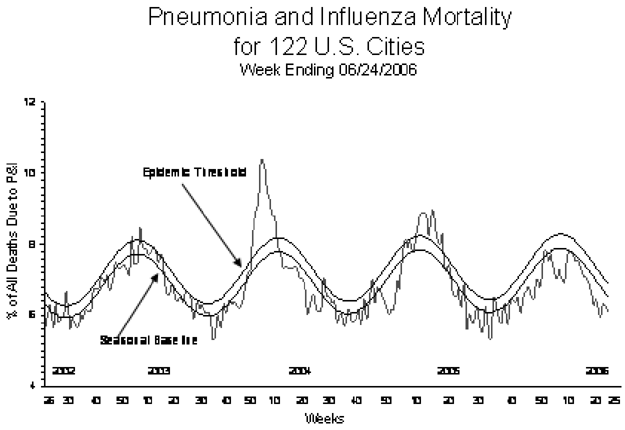


Figure 6: CDC's national baseline model for determining influenza outbreaks.

using a common set of syndromic data in multiple cities in the United States. To determine natural outbreaks in the data, a team of epidemiologists and medical specialists were assigned 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 of how outbreaks and their dates are determined, of the circularity of using statistical guides to help determine outbreaks, and of the determination of outbreak-free intervals. A different attempt to label outbreak periods is to use diagnostic information such as actual hospital admissions (e.g., Ivanov et al., 2003). However, the coding routine for diagnostic data is different than pre-diagnostic coding because it is aimed at billing.

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 such as 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 to assess the method's performance, thereby leaving opportunity for scientific confounding (Rolka, 2006). But more fundamentally, the unlabeled nature of the data causes serious problems in using standard evaluation methods.

The most widely used evaluation metrics in biosurveillance have been *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 in order 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 are also 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 vs. the false positive rate for a range of threshold values, and the area under the curve measures model accuracy (the larger the area, the better the model). However, because in practice only part of the ROC curve is of interest, there have been suggestions to use 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, which displays the aggregate performance, might mask inadequate performance, especially when the conditional performance is in opposite directions. 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 used in fraud detection, incorporate the timeliness aspect by plotting a timeliness score vs. 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 that is 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 in particular the average run length (ARL), which are the main evaluation metrics in statistical quality control, are rare in biosurveillance literature and practice and are found in the few statistically-oriented papers (e.g., Stoto et al., 2006). Another set of statistical predictive measures that appear in more statistically-oriented papers are the Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) (Reis and Mandl, 2003; Burkom et al., 2007). These do not directly measure algorithm performance in terms of alarms, but rather provide a measure for assessing the fit of the time series models.

To evaluate true positive rates in the absence of real bioterror-related outbreaks, the popular approach has been to seed real syndromic data with simulated effects of artificial outbreaks (e.g., Burkom et al., 2007; Goldenberg et al., 2002; Reis and Mandl, 2003; Stoto et al., 2006, and 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., a detection after seven 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 et al., 2005b). Usually, different magnitudes of this shape are injected to evaluate the sensitivity of the algorithm to the outbreak magnitude. An alternative approach is simulating syndromic data and simulating outbreaks (see 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 upon 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 that are triggered by the system. In theory, in the event of an alarm they should trace its cause and determine whether indeed there is an outbreak taking place. A special attempt to do this was 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 they reported weekly false alarms, and for each alarm experts went out and examined the situation to evaluate the chance of a real outbreak (Fienberg and Shmueli, 2005). The procedure employed by the NYCDOHMH for investigating alarms included calling the hospitals which generated alarms, consulting with them whether today's behavior at the hospital is indicative of an outbreak and, if necessary, the emergency department staff was alerted (Paladini, 2006). However, these investigations were reported to be difficult and the authors proposed a new investigation protocol.

The current reality is that users see alarms on nearly every given day, most of which turn out to be false alarms. This leads many users to ignore alarms altogether 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". There are many reasons for the extremely high false alarm rate, some related to data quality but others are actually statistical. One statistical reason is inadequate handling of the non-stationary raw data, e.g., 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 sub-series (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 very visible.

Several methods for handling multiple testing exist in the statistics literature, including family-wise Bonferroni-type corrections, the 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 (Bonferroni is considered over-conservative, FDR corrections depend on the number of hypotheses and are problematic with too few hypotheses, and Bayesian methods are sensitive to the choice of prior and it is unclear how to choose a prior). But the major problem is in convincing users that these methods are not masking real outbreaks at a univariate level.

Burkom et al. (2005a) 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 takes place when creating source-specific hypotheses that are tested separately (or even when stratifying, e.g., by age-group), whereas consensus monitoring occurs when multiple series describing the same population are monitored. They proposed using family-wise error rates for parallel monitoring, and multiple univariate corrections (Edgington, 1972) for consensus monitoring. However, the exact definition of when each of these applies within a biosurveillance context is unclear. A challenge is therefore 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 as a better solution, it carries challenges such as lags in reporting (which misalign the different series), large differences in data structure across data sources, the absence of knowledge about the exact relationship between 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 non-stationarity. In the context of bioterror-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 in order to construct performance metrics that are meaningful and accurate within the biosurveillance context.

3.5 Evaluating Overall System Performance

Although there are many empirical studies of the performance of specific algorithms (e.g., the description of algorithms employed 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. When considering the system as a whole it includes multiple algorithms being performed on multiple data series. Therefore reported performance levels in published empirical studies appear optimistic compared to 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 in the DARPA BioALIRT program). But in practice the situation is much worse, with false alarms occurring nearly daily. As mentioned earlier, this results from inadequate monitoring methods being applied to the raw count data, but also from severe multiple testing that results from the multiple syndromic series being monitored by multiple algorithms on a daily basis. A challenge is therefore to move to system-level performance, thereby determining alarm rates and timeliness of the entire system rather than specific algorithms.

Another issue that is directly related to system performance is costs. An important piece of information that is hard to obtain in syndromic monitoring is resources and costs. In order to design a practical system that yields "reasonable" true and false alarm rates, it is essential to know what is considered reasonable. However, such estimates are hard to elicit. We therefore 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, in order 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 (including data quality issues, lagtimes in data reporting, etc.), 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 these aspects (and perhaps others) the real value of the system will remain ambiguous. This issue stresses further 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 in developing and quickly deploying modern biosurveillance systems has resulted in the use of a small set of monitoring tools and in growing research efforts to improve the 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 continual implementation even in this new environment. In the following we describe some of the main monitoring tools that are currently deployed 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 ones are 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 describe briefly an earlier national system called Early Aberration Reporting System (EARS, http://www.bt.cdc.gov/surveillance/ears/), which has evolved into several of the new biosurveillance systems. Developed by CDC (Hutwagner et al., 2003), EARS consists of three control charts: two Shewhart I-charts (called C1 and C2) applied directly to counts and a truncated CuSum chart that only sums up the last three days (called C3). The two differences of the EARS charts from standard control charts are aimed at accounting for non-stationarity 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), which is a window of several days preceding the most recent count that provides a guardband between the data used for parameter estimation (phase-I) and that 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. However, since EARS is applied to raw count data it is more suitable for traditional biosurveillance, where data arrive at weekly intervals and therefore suffer less from autocorrelation and cyclic behavior compared to syndromic data. However, some of these tools have permeated into modern surveillance systems such as ESSENCE and BioSense (where the CuSum-like method is in fact the EARS C3 chart). Most recently, BioSense introduced W2, a modification of the C2 chart that monitors data separately for weekdays and weekends/holidays (see BioSense bulletin, Sept 30, 2006 http://www.cdc.gov/biosense/).

Modern biosurveillance systems also use other monitoring tools. 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 post-holiday indicators. To determine which of the two 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 setting (28-day sliding-phase-I window and 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 User Manual http://rods.health.pitt.edu); The second tool is a nonstandard combination of CuSum and EWMA: an EWMA is used to predict next-day counts, and a CuSum monitors the residuals from these predictions with an empirically-chosen threshold of $h = 4.08$. In general, many current algorithms use thresholds that are determined from empirical experimentation rather than theoretical design.

The third monitoring tool in RODS is a Recursive Least Squares (RLS) algorithm, which fits an autoregressive model to the counts and updates estimates continuously by minimizing prediction error. A Shewhart I-chart is then applied to the residuals, using a threshold of 4 standard deviations. This is similar in notion to the regression-tool in ESSENCE. A similar approach is used in CDC's BioSense (called SMART scores, developed by Kleinman et al. (2004)), using a Poisson regression of daily counts on the following predictors (1) A secular (long term) linear trend over time (2) Sine and cosine effects for seasonality (3) Month indicators (11 dummies) for non-trigonometric effects of season (4) Day-of-week indicators (6 dummies) for day-to-day variability, and (5) 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-the-week coefficients, and a couple of years to minimally account for seasonality or monthly effects. The regression model is then used to predict next-day counts, and a SMART score is generated by transforming the calculated p-value that compares the prediction with the actual count (www.cdc.gov/phin/component-initiatives/biosense/FAQ_BioSense_App.pdf).

The only tool that is not regression-based or classic control charts 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, it is a nonparametric de-trending method. The residuals (de-trended actual counts) are then monitored using an ordinary Shewhart I-chart with a threshold of 4 standard deviations.

4.1.2 Methods in the Literature

Aside from the algorithms used within the large systems, there has been a growing literature on new proposed monitoring algorithms, with empirical studies showing their performance when applied to syndromic data. Since there has not been a dedicated journal to this field prior to the recent new journal Advances in Disease Surveillance, the literature is dispersed across journals from multiple fields (e.g., medicine, epidemiology, bioinformatics, quality control, and public health). A good resource for recent publications is the International Society for Disease Surveillance's website (www.syndromic.org). In the following we brie fly survey some of the methods proposed and the context in which they were tested. The goal of the survey is to show the relative infancy of modern biosurveillance and to attract further statistical involvement.

Model-Based Approaches

There have been several efforts to directly model explainable effects (e.g., day-of-week, seasonality, and 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 non-stationary nature of data and its diversity across series. Such fitting requires customized treatment for each series, where the process of pre-processing, identification, and estimation requires expertise, time, sufficient history, and computational power. An example is Reis and Mandl (2003) who used ARIMA models for a single series of eight years of daily visits to a pediatric hospital. Even if applied prospectively, there exists the danger of incorporating gradual outbreaks into the model, thereby masking the outbreak (Reis and Mandl, 2003). For these reasons ARIMA models are more likely to serve in retrospective analyses rather than for 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 showed 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 is usually unavailable or unrepresentative of current behavior due to changes in treatments, coding, population behavior, evolving informatics, faster data rates, changes in reporting practices, etc.

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 to multiple univariate charts. Similarly, Stoto et al. (2006) used 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 being multivariate normal and independent over time. Second, because the interest is in discovering increases in disease incidences, charts must be modified to be directionally sensitive (as discussed by Hawkins (1991, 1993); Rogerson and Yamada (2004) and implemented in Burkom et al. (2004); Fricker (2006); Joner et al. (2007)). Third, the cross-covariance structure is assumed to remain constant although empirical evidence indicates a time-varying structure.

A different multivariate formulation was used by Najmi and Magruder (2005) 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 the clinical data. Shmueli and Fienberg (2006) described several other multivariate schemes that are potentially more suitable for syndromic surveillance because they make less restricting assumptions about the underlying data and have proven useful in other fields where similar data and goals are encountered. An important factor, however, is a balance between simplicity and performance.

Data-Driven Approaches

Because of the difficulty to find a parametric "one model fits all" that is sufficiently flexible to accept a wide array of non-stationary input series in an automated fashion, an alternative approach employs data driven methods for removing explainable patterns to achieve iid normal residual series.

Transformations were suggested for achieving normality and accounting for the multiplicative day-of-week effects in several studies (Fricker, 2006; Brillman et al., 2005; Burkom et al., 2007; Stoto et al., 2006). Seven-day differencing was suggested by Muscatello (2004) for removing day-of-week effects.

Data smoothing has also been suggested for removing 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 2-7 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 et al., 2007). The study by Burkom et al. (2007) found that Holt-Winter's exponential smoothing, with some adaptations, outperformed ordinary regression and adaptive regression models yet is highly automatable.

There have been some attempts 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 (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. However, their use for process prediction or monitoring, and in general for prospective tasks, is much rarer. Goldenberg et al. (2002) used a redundant spline-based wavelet for decomposing series of OTC medication sales in order 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 de-trend data and to remove post-holiday dips, by subtracting the low-frequency scale from the original series (Zhang et al., 2003).

Shmueli (2005) proposed a modified scheme of that by Aradhye et al. (2003), where coefficients within each scale are thresholded using 3-sigma limits and then the original series is 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, with a comparison to regression-based methods. Finally, there exist some promising methods for multivariate wavelet-based monitoring in other fields, such as that by Bakshi (1998), which have not been explored in biosurveillance.

4.2 Spatial and Spatio-Temporal Monitoring

Although the focus of this paper is on temporal monitoring, we must mention the complementary area of spatio-temporal monitoring, where patient, customer, or clinic location information in data records is used for identifying localized case clusters. An issue of *Statistics in Medicine* (edited by Lawson et al., 2006) was devoted to disease cluster detection. Currently the most widely used method is the spatio-temporal 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 for obtaining p-values.

The main focuses of subsequent research have been (1) comparing the empirical performance of the spatio-temporal scan statistic with other algorithms (e.g. Kulldorff et al., 2003; Kedem and Wen, 2007);

(2) Improving the scan statistic in terms of computational time (e.g., Neill et al. (2006) who developed a faster Bayesian alternative that does not require randomization); and (3) 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 are the locations of care providers or clinics available in a dataset, 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 the 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 further work is needed to choose 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 the patient residence address. Data sets that provide a work address are rare, so that workplace-based clusters are unlikely to be found using space-based algorithms like scan statistics. The workflow scan statistic of Duczmal and Buckeridge (2006) gives an approach for using demographic data to find these clusters, and 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 an identified clusters are linked and worthy of investigation, and clearing this hurdle will require close cooperation of care providers, informaticists, and statisticians.

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

5 Conclusions and Future Directions

The goals of this paper are to introduce the important area of modern biosurveillance and to enumerate the challenges that it poses to traditional statistical monitoring. There are currently not many statisticians involved in the discipline, and there is a pressing need to develop improved biosurveillance systems. There are opportunities for developing statistical methodology 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 that requires 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, etc. 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 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 bioterror-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 that is hard to model if not specified as an outbreak. The current approach for evaluation has therefore been 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 which should be carefully addressed.

An important seemingly non-statistical challenge is the actual use of biosurveillance systems by public health officials. The current disconnect between the algorithm developers, implementers, and users has lead to systems with uncontrolled alert rates dependent on a variety of epidemiological and well as informatics issues. Such experiences foster distrust in statistical monitoring and in biosurveillance itself. It is our responsibility to use all statistical ammunition available to create adequate yet simple methods that will assist expert decision making rather than confuse the user.

All these issues highlight the similarity between bioterror-related outbreak detection and other event detection tasks such as fraud detection in accounting (e.g., Bay et al., 2006) or 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 evaluation of 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 hard to determine. Also, 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 reasons mentioned above we believe that these tools are not always adequate for the purpose and requirements of biosurveillance. The biosurveillance literature, however, 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 are currently not applied directly to modern biosurveillance (e.g., Hidden Markov Models for influenza surveillance by Ozonoff and Sebastiani (2006); temporal scan statistic for monitoring weekly national reports of brucellosis by Wallenstein and Naus (2004); neural networks for multivariate health surveillance by Adams et al. (2006); or the moving-F chart that is not affected by reduction in variability by Riffenburgh and Cummins (2006)). The lack of application to modern biosurveillance raises attention to one of the main barriers to statistical involvement in this field, which is data access. Currently syndromic data are only available to researchers affiliated with a particular biosurveillance system or research group, for reasons of data confidentiality and non-disclosure agreements. This is a major obstacle in the way of scientific progress in both temporal and spatio-temporal biosurveillance, and hopefully some data will be made available to academic researchers. There have been various attempts at drawing statisticians to biosurveillance research, such as the workgroup on Anomaly Detection in National Defense and Homeland Security by the Statistical and Applied Mathematics Sciences Institute (SAMSI http://www.samsi.info/200506/ndhs/workinggroup/ad/), and a growing number of conference sessions in statistics conferences that are devoted to biosurveillance. A second barrier of entry into this research field is the dispersion of the relevant literature across journals in a variety of fields, and the lack of detailed descriptions of methods that are 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.

Although this paper concentrates on challenges in biosurveillance, we do have a vision of an improved future. As a field that is in its infancy, but which has grown fast to its current stage, it is now time for reflecting on the past and for planning into the future. With further involvement of statisticians and stronger collaborations between the different experts creating and using modern biosurveillance systems, many of the challenges described here are likely to be tackled and solutions created. Better implementation and evaluation of these systems will enable the construction and implementation of reliable and eventually trusted biosurveillance systems with the potential for earlier outbreak detection and the broader utility for corroboration and for public health threat characterization and tracking.

Acknowledgements

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 the Harvard Medical School and Harvard Pilgrim Health Care, Dr. Bill Woodall from Virginia Tech, Dr. Ron Fricker from the The Naval Postgraduate School, Dr. Karen Kafadar from the University of Colorado at Denver Health Sciences Center, and the AE 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 NIH grant RFA-PH-05-126.

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.

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.

Bakshi, B. R. (1998). Multiscale pca with application to multivariate statistical process monitoring. *AIChE Journal*, 44:1596–1610.

Bay, S., Kumaraswamy, K.and Anderle, M., Kumar, R., and Steier, D. (2006). Large scale detection of irregularities in accounting data. In *6th International Conference on Data Mining (ICDM), 75-86*.

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, Series B*, 57:289–300.

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.

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.

Brookmeyer, R., Johnson, E., and Barry, S. (2005). Modelling the incubation period of anthrax. *Statistics in Medicine*, 24(4):531–542.

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.

Burkom, H. S, a. M. S., Coberly, J., and Hurt-Mullen, K. (2005a). Public health monitoring tools for multiple data streams. *MMWR*, 54(Suppl):55–62.

Burkom, H. S., Elbert, Y., Feldman, A., and Lin, J. (2004). Role of data aggregation in biosurveillance detection strategies with applications from essence. *MMWR*, 53 (suppl):67–73.

Burkom, H. S., Hutwagner, L., and Rodriguez, R. (2005b). Using point-source epidemic curves to evaluate alerting algorithms for biosurveillance. In *2004 Proceedings of the American Statistical Association, Statistics in Government Section [CD-ROM], Toronto.*

Burkom, H. S., Murphy, S. P., and Shmueli, G. (2007). Automated time series forecasting for biosurveillance. *Statistics in Medicine*, 26(2):4202–4218.

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. *MMWR*, 53(Suppl):86–94.

Dash, D., Lane, T., Margineantu, D., and Wong, W.-K. (2006). Opening remarks. Workshop on Machine Learning Algorithms for Surveillance and Event Detection, 23rd Intl Conference on Machine Learning.

Duczmal, L. and Buckeridge, D. L. (2006). A workflow spatial scan statistic. *Statistics in Medicine*, 25(5):743 – 754.

Edgington, E. S. (1972). A normal curve method for combining probability values from independent experiments. *Journal of Psychology*, 82:85–89.

Fawcett, T. and Provost, F. (1999). Activity monitoring: Noticing interesting changes in behavior. In Chaudhuri and Madigan, editors, *5th ACM SIGKDD Intl Conference*, pages 53–62, San Diego, CA.

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.

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

Fricker, Jr., R. D. (2006). Directionally sensitive multivariate statistical process control methods with application to syndromic surveillance. *Advances in Disease Surveillance*, 3:1.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B*, 64:499–451.

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.

Hawkins, D. M. (1991). Multivariate quality control based on regression-adjusted variables. *Technometrics*, 33:61–75.

Hawkins, D. M. (1993). Regression adjustment for variables in multivariate quality control. *Journal of Quality Technology*, 25:170 –182.

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.

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 Annu Symp*, number 318-322.

Jin, J. and Shi, J. (1999). Feature-preserving data compression of stamping tonnage information using wavelets. *Technometrics*, 41(4):327–339.

Joner, M. D., J., Woodall, W. H., Reynolds, M. R., J., and Fricker, R. D. (2007). A one-sided mewma chart for health surveillance. *Quality and Reliability Engineering Intl (submitted)*.

Kedem, B. and Wen, S. (2007). Semi-parametric cluster detection. *Journal of Statistical Theory and Practice*, 1(1):49–72.

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.

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.

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. *Am J Epidemiol*, 159:217–224.

Kleinman, K. P. and Abrams, A. M. (2006). Assessing surveillance using sensitivity, specificity and timeliness. *Statistical Methods in Medical Research*, 15(5):445–464.

Kulldorff, M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A*, 164(1):61–72.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, 25(22):3929–3943.

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.

Kulldorff, M., Tango, T., and Park, P. J. (2003). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42(4):665 – 684.

Lawson, A., Gangnon, R., and Wartenberg, D. (2006). Special issue on developments in disease cluster detection. *Statistics in Medicine*, 25(5).

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 Intl Conference on Machine Learning.*

Lotze, T., Shmueli, G., and Yahav, I. (2007). Simulating multivariate syndromic time series and outbreak signatures. Technical report, Smith School of Business, University of Maryland.

Meselson, M., Guillemin, J., Hugh-Jones, M., a. L. A., Popova, I., Shelokov, A., and Yampolskaya, O. (1994). The sverdlovsk anthrax outbreak of 1979. *Science*, 266(5188):12021208.

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.

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.

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.

Neill, D. B., Moore, A. W., and Cooper, G. F. (2006). *Advances in Neural Information Processing Systems*, chapter A Bayesian spatial scan statistic, pages 1003–1010. MIT Press, Cambridge, MA.

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.

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.

Pavlin, J. A. (1999). Epidemiology of bioterrorism. *Emerging Infectious Diseases*, 5:528–565.

Reis, B. and Mandl, K. (2003). Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*, 3(2).

Riffenburgh, R. and Cummins, K. (2006). A simple and general change-point identifier. *Statistics in Medicine*, 25(6):1067–1077.

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.

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.

Rolka, H. (2006). *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, chapter Emerging Public Health Biosurveillance Directions, pages 101–107. Springer.

Shmueli, G. (2005). Wavelet-based monitoring for modern biosurveillance. Technical report, RHS-06-002, University of Marlyand, Robert H Smith School of Business.

Shmueli, G. and Fienberg, S. E. (2006). *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, chapter Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Bio-Surveillance, pages 109–140. Springer.

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.

Siegrist, D. and Pavlin, J. (2004). Bio-alirt biosurveillance detection algorithm evaluation. *Morbidity and Mortality Weekly Reports (MMWR)*, 53 (suppl):152–158.

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 ICML.*

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). *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, chapter Evaluating Statistical Methods for Syndromic Surveillance, pages 141–172. Springer.

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.

Wallenstein, S. and Naus, J. (2004). Scan statistics for temporal surveillance for biologic terrorism. *MMWR*, 53(Suppl):74–78.

Woodall, W. H. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, 38(2):89–104.

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.

Zhang, J., Tsui, F., Wagner, M., and Hogan, W. (2003). Detection of outbreaks from time series data using wavelet transform. In *AMIA Annual Symposium Proceedings*, pages 748–752.