2006

Evaluating Statistical Methods for Syndromic Surveillance

Stoto, Michael A.; Fricker, Ronald D. Jr.; Jain, Arvind; Diamond, Alexis; Davies-Cole, John O.; Glymph, Chevelle; Kidane, Gebreyesus; Lum, Garrett; Jones, LaVerne; Dehan, Kerda...

Student Research Briefings: Biosurveillance, Defense Threat Deduction Agency, November 2010.
https://hdl.handle.net/10945/38738

Evaluating Statistical Methods for Syndromic Surveillance

Michael A. Stoto[1], Ronald D. Fricker, Jr.[2], Arvind Jain[3], Alexis Diamond[4], John O. Davies-Cole[5], Chevelle Glymph[6], Gebreyesus Kidane[7], Garrett Lum[8], LaVerne Jones[9], Kerda Dehan[10], and Christine Yuan[11]

[1] RAND Statistics Group, mstoto@rand.org
[2] Department of Operations Research, Naval Postgraduate School, rdfricker@nps.edu
[3] RAND Statistics Group, arvind_jain@rand.org
[4] The Institute for Quantitative Social Science, Harvard University, adiamond@fas.harvard.edu
[5] Bureau of Epidemiology and Health Risk Assessment, District of Columbia Department of Health, john.davies-cole@dc.gov
[6] Bureau of Epidemiology and Health Risk Assessment, District of Columbia Department of Health, chevelle.glymph@dc.gov
[7] Bureau of Epidemiology and Health Risk Assessment, District of Columbia Department of Health, gebreyesus.kidane@dc.gov
[8] Bureau of Epidemiology and Health Risk Assessment, District of Columbia Department of Health, garret.lum@dc.gov
[9] Bureau of Epidemiology and Health Risk Assessment, District of Columbia Department of Health, laverne.jones@dc.gov
[10] Bureau of Epidemiology and Health Risk Assessment, District of Columbia Department of Health, kerda.dehan@dc.gov
[11] Bureau of Epidemiology and Health Risk Assessment, District of Columbia Department of Health, christine.yuan@dc.gov

Since the terrorist attacks on September 11, 2001, many state and local health departments around the United States have started to develop *syndromic surveillance* systems. Syndromic surveillance — a new concept in epidemiology — is the statistical analyses of data on individuals seeking care in emergency rooms (ER) or other health care settings with preidentified sets of symptoms thought to be related to the precursors of diseases. Making use of existing health care or other data, often already in electronic form, these systems are intended to give early warnings of bioterrorist attacks or other emerging health conditions. By focusing on symptoms rather than confirmed diagnoses, syndromic surveillance aims to detect bioevents earlier than would be possible with traditional surveillance systems. Because potential bioterrorist agents such as anthrax, plague, brucellosis, tularemia, Q-fever, glanders, smallpox, and viral hemorrhagic fevers initially exhibit symptoms ("present"

in medical terminology) of a flulike illness, data suggesting a sudden increase of individuals with fever, headache, muscle pain, and malaise might be the first indication of a bioterrorist attack or natural disease outbreak. Syndromic surveillance is also thought to be useful for early detection of natural disease outbreaks [Hen04].

Research groups based at universities, health departments, private firms, and other organizations have proposed and are developing and promoting a variety of surveillance systems purported to meet public health needs. These include methods for analysis of data from healthcare facilities, as well as reports to health departments of unusual cases. Many of these methods involve intensive, automated statistical analysis of large amounts of data and intensive use of informatics techniques to gather data for analysis and to communicate among physicians and public health officials [WTE01]. Some of these systems go beyond health care data to include nonhealth data such as over-the-counter (OTC) pharmaceutical sales and absenteeism that might indicate people with symptoms who have not sought health care [Hen04].

There are a number of technological, logistical, and legal constraints to obtaining appropriate data and effective operation of syndromic surveillance systems [Bue04]. However, even with access to the requisite data and perfect organizational coordination and cooperation, the statistical challenges in reliably and accurately detecting a bioevent are formidable. The object of these surveillance systems, of course, is to analyze a stream of data in realtime and determine whether there is an anomaly suggesting that an incident has occurred. All data streams, however, have some degree of natural variability. These include seasonal or weekly patterns, a flu season that appears at a different time each winter or perhaps not at all, differences in coding practices, sales promotions for OTC medications, and random fluctuations due to small numbers of individuals with particular symptoms. Furthermore, for some natural outbreaks or bioterrorist attacks the "signal" (the number of additional cases over baseline rates) may be small compared to the "noise" (the random or systematic variation in the data). As a result, even the most effective statistical detection algorithms face a trade-off among three factors: sensitivity, false positives, and timeliness.

The goals of this chapter are (1) to introduce the statistical issues in syndromic surveillance, (2) to describe and illustrate approaches to evaluating syndromic surveillance systems and characterizing their performance, and (3) to evaluate the performance of a couple of specific algorithms through both abstract simulations and simulations based on actual data. Section 1 of this chapter introduces and discusses the statistical concepts and issues in syndromic surveillance, illustrating them with data from an ER surveillance system from the District of Columbia. Section 2 presents methods from the statistical process control (SPC) literature, including variants on existing multivariate detection algorithms tailored to the syndromic surveillance problem, and compares and contrasts the performance of univariate and multivariate techniques via some abstract simulations. Section 3 then compares

the new multivariate detection algorithms with commonly used approaches and illustrates the simulation approach to evaluation using simulations based on actual data from seven Washington, DC, hospital ERs. We conclude with a discussion about the implications for public health practice.

1 Background

Immediately following September 11, 2001, the District of Columbia Department of Health (DC DOH) began a surveillance program based on hospital ER visits. ER logs from nine hospitals are faxed on a daily basis to the health department, where health department staff code them on the basis of chief complaint, that is, the primary symptom or reason that the patient sought care, recording the number of patients in each of the following syndromic categories: death, sepsis, rash, respiratory complaints, gastrointestinal complaints, unspecified infection, neurological, or other complaints. These data are analyzed daily using a variety of statistical detection algorithms, and when a syndromic category shows an unusually high occurrence, a patient chart review is initiated to determine if the irregularity is a real threat.

Simply displaying the daily number of ER visits for any given symptom group results in a figure in which day-to-day stochastic variation dominates any subtle changes in numbers of cases over time. To address this problem, the DC DOH employs a number of statistical detection algorithms to analyze data on a daily basis and raise an "alarm" when the count is significantly greater than expected, which may suggest a possible outbreak or attack. This type of analysis can help to identify the onset of the annual influenza season. The data also reveal indications of the "worried well" who sought care during the 2001 anthrax attacks and a previously undetected series of gastrointestinal illness outbreaks that occurred over a four-month period in different hospitals. No single symptom group or detection algorithm consistently signaled each of the events [SSM04].

1.1 Characterizing the Performance of Statistical Detection Algorithms

Although it is possible to have different levels of certainty for an alarm, syndromic surveillance algorithms typically operate in a binary fashion; on any given day they either alarm or they do not. Operating in this way, the performance of a detection algorithm in the context of a particular dataset can be characterized according to its sensitivity, false-positive rate, and timeliness. *Sensitivity*, sometimes called the true positive rate and similar to the power of a statistical hypothesis test, is the probability that an outbreak will be detected in a given period when there in fact is an outbreak. Clearly, a surveillance system should have as much sensitivity as possible. Lowering the

threshold at which an alarm is sounded can generally increase sensitivity, but only at the expense of false positives.

A *false positive* occurs when an algorithm alarms on a day when there is no actual outbreak. In medical or epidemiological terminology, *specificity* is 1 minus the probability of a false positive, or the probability that an alarm will not be raised on a day that there is no outbreak. Ideally, the probability of a false positive would be zero, but practically it is always positive. Intrinsic variability in the data means that every methodology can alarm when in fact there is no event.

It is usually possible to make the false-positive rate tolerably small. There are two difficulties, however. First, lowering the false-positive rate generally involves either decreasing sensitivity or lowering timeliness (or both). Second, even with a very low false-positive rate for a single algorithm or system, it is still possible — even likely — that in the aggregate the number of false positives may be unacceptably large.

For example, sometime in the near future it is possible that thousands of syndromic surveillance systems will be running simultaneously in towns, cities, counties, states, and other jurisdictions throughout the United States. Each of these jurisdictions might be looking at data in six to eight symptom categories, separately from every hospital in the area, and so on. Suppose every county in the United States had a detection algorithm in place that was used daily and that had a 0.1% false-positive rate. Because there are approximately 3,000 counties, nationwide three counties a day on average would have a false-positive alarm. While any particular county would only experience a false positive about once every three years, which may be an acceptable rate at the county level, is the nationwide false-positive rate acceptable? The impacts of excessive false alarms are both monetary, as resources must respond to phantom events, and operational, as too many false alarms desensitize responders to real events.

Because a rapid response to a bioterrorist attack or natural disease outbreak is essential to minimizing the health consequences, timeliness is an important characteristic of all surveillance systems. With its focus on symptoms that occur before formal diagnosis, syndromic surveillance is specifically designed to enhance timeliness. While timeliness does not have a well-established definition to parallel sensitivity and specificity, we think of it as the speed at which an algorithm alarms during an outbreak.

Stoto, Schonlau, and Mariano [SSM04] characterized the trade-off between sensitivity and timeliness in a simulation study. Using the daily number of admissions of patients with influenzalike illness (ILI) over a three-year period to the emergency department of a typical urban hospital, which averages three per day outside the winter flu season, they added a hypothetical number of extra cases spread over a number of days to mimic the pattern of a potential bioterror attack. A "fast" outbreak was defined as 18 additional cases over three days — 3 on the first day, 6 on the second, and 9 on the third. A simulated "slow" outbreak involved the same total number of cases, but they

were distributed over nine days as follows: 1, 1, 1, 2, 2, 2, 3, 3, 3. Each of these simulated outbreaks was added on each day in the database outside the winter flu season. Four different detection algorithms were examined. The first used ER admissions from a single day; the others used data from multiple days using various CuSum (cumulative sums) methods (such as those to be defined in Sect. 2.3), with the algorithms varying in the weight they gave to more recent data.

The simulation results suggest the minimum size and speed of outbreaks that are detectable. Even with an excess of 9 cases over two days, which is three times the daily average, there was only a 50% chance that the alarm would go off on the second day of an outbreak. Figure 1 indicates how this probability — the sensitivity of the algorithm — varies by day. In the slow outbreak, when 18 cases were spread over nine days (see Fig. 2), chances were no better than 50–50 that the alarm would sound by the ninth day.

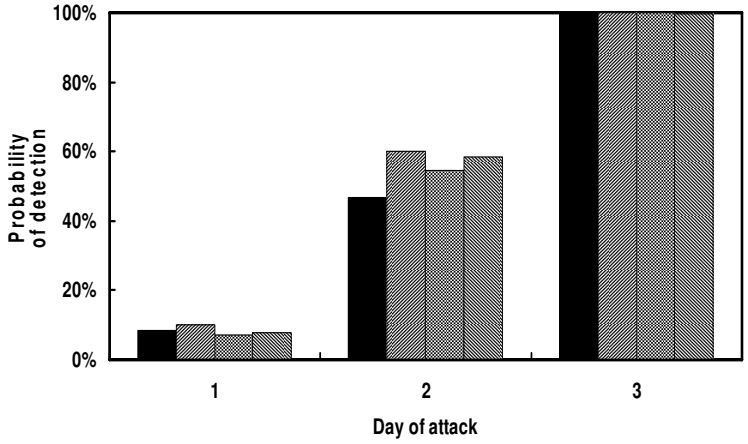

Fig. 1. Shaded bars correspond to four detection algorithms: the first using only one day's data, the other three combining data from multiple days. All four syndromic surveillance methods worked equally well for fast-spreading bioterrorist attacks, but had only about a 50–50 chance of detecting the outbreak by day two. See Stoto et al. [SSM04] for more information.

1.2 Evaluation of Syndromic Surveillance Systems

There are a number of ways to evaluate syndromic surveillance systems, formal and informal. For example, the Centers for Disease Control and Prevention's (CDC) "Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks" [CDC04a] offers a useful framework to guide

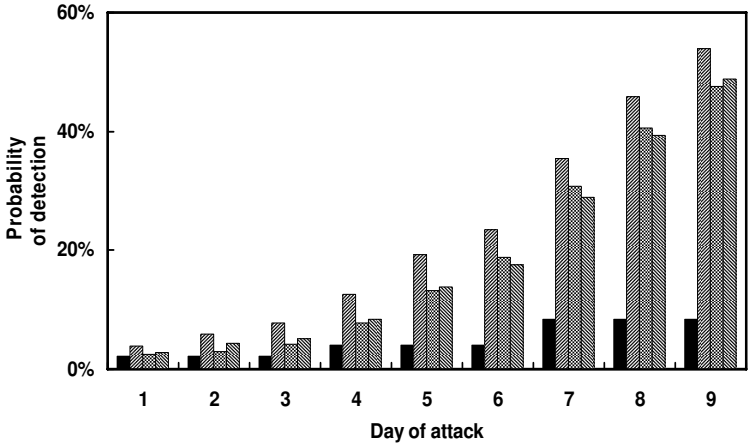

Fig. 2. Methods that combine data from multiple days (the hatched bars) were more effective at detecting slow-spreading attacks, but even the best method took until day nine to have a 50–50 chance of detecting a slow outbreak. See Stoto et al. [SSM04] for more information.

evaluation efforts. Other approaches focus on the completeness, timeliness, and quality of the data [BBM04], or on how syndromic surveillance efforts relate to public health practice [Rei03]. The annual national syndromic surveillance conference (see `http://www.syndromic.org`) offers many examples of such evaluations.

Formal approaches tend to focus on characterizing the statistical performance of detection algorithms applied to particular data streams. The Stoto, Schonlau, and Mariano [SSM04] analysis described above illustrates the simulation approach, and Sect. 3 of this paper presents a more detailed example. Both of these examples use real data as a baseline and add a simple simulated outbreak. As a perhaps more realistic alternative, Stacey [Sta04] has described an approach in which real data are used to model simulated outbreaks for testing purposes.

The retrospective analysis of known natural outbreaks is an alternative approach to evaluation. Siegrist and Pavlin [SP04], for instance, report on an exercise in which four leading biosurveillance research teams compared the sensitivity, specificity, and timeliness of their detection algorithms in two steps. First, an outbreak detection team identified actual natural disease outbreaks — eight involving respiratory illness and seven involving gastrointestinal illness — in data from five metropolitan areas over a 23-month period but did not reveal them to the research teams. Second, each research team applied its own detection algorithms to the same data, to determine whether and how quickly each event could be detected. When the false-alarm rate was set at

one every 2 to 6 weeks, the best algorithms from each research team were able to detect all of the respiratory outbreaks; for two of the four teams detection typically occurred on the first day that the outbreak detection team determined as the start of the outbreak; for the other two teams, detection occurred approximately three days later. For gastrointestinal illness, the teams typically were able to detect six of seven outbreaks, one to three days after onset. (Of course, as previously discussed, such detection times are partially a function of the false-alarm rate — decreasing the false-alarm rate will increase the detection time.)

One can also look at the epidemiological characteristics of various pathogens to clarify the implications for syndromic surveillance [Bue04]. For instance, Fig. 3 gives two examples that differentiate between attacks in which many people are exposed at the same time, and those in which a contagious agent might cause large numbers of cases in multiple generations. Example A (the line with the triangles) illustrates what might be found if 90 people were exposed to a noncontagious agent (such as anthrax) and symptoms first appeared eight days on average after exposure. Example B (the line with the squares) illustrates the impact of a smaller number of people (24) exposed to a contagious agent (such as smallpox) with an average incubation period of 10 days. Two waves of cases appear, the second larger and 10 days after the first. Because the two epidemic curves are similar on days one through three, it is difficult to know what can be expected, but if the agent were contagious (Example B), early intervention could save some or all of the second generation of cases. In Example A, however, everyone would already have been exposed by the time that the outbreak was detected.

1.3 Improving the Performance of Syndromic Surveillance

Faced with results like those in Figs. 1 and 2, one naturally asks whether more effective systems can be developed. There are a number of alternatives that could be considered and actually are the subject of current research.

Most detection algorithms can be characterized in three respects: (1) what they assume as the background level and pattern of diseases or symptoms, (2) the type of departures from normal that they are tuned to detect (an exponential increase in the number of cases, a geographic cluster of cases, and so on), and (3) the statistical algorithm they use to determine when the data indicate a departure from normal (i.e., an "anomaly"). Each presents opportunities to improve the performance of detection algorithms. Ultimately, however, there really is no free lunch. As is the case in other areas of statistics, there is an inherent trade-off between sensitivity and specificity, and the special need for timeliness makes it even more difficult in this application. Every approach to increasing sensitivity to one type of attack is likely to cause a detection algorithm to be less sensitive to some other scenario. To circumvent this trade-off, we would have to have some knowledge about how a terrorist may attack.

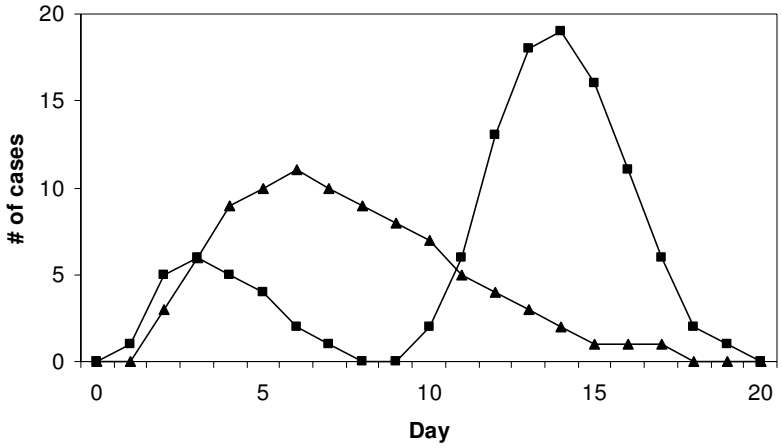

Fig. 3. Two epidemic curves that are similar on days one through three, but then diverge. The line with the triangles results from a gamma distribution with $\mu = 8$ and $\sigma = 4$. The line with the squares simulates an infectious agent with a mean latency of 10 days. It is made up of the sum of observations from two gamma distributions, one with $\mu = 4$ and $\sigma = 2$, and the second with $\mu = 14$ and $\sigma = 2$.

Siegrist's retrospective analysis [Sie04] summarizes the details of some of the leading syndromic surveillance systems, illustrating each of the approaches described below.

Background Level and Pattern

Models to adjust for background patterns can be simple or complex. At one extreme, a method may assume a constant mean number of cases and standard deviation over the entire year for each data series monitored. In other models, the expected number of cases varies seasonally, in a constant weekly pattern (reflecting availability of health services on weekends, for instance), or as represented in an autoregressive process.

Syndromic surveillance systems typically compare current cases to the number in the previous day or week, the number in the same day in the previous year, or some average of past values. More sophisticated approaches use statistical models to "filter" or reduce the noise in the background data to try to make the signal more obvious so that an outbreak would be easier to detect. For instance, if a hospital ER typically sees more ILI patients on weekend days (when other facilities are not open), a statistical model can be developed to account for this effect. With a long enough data series, annual effects can also be incorporated. Some patterns are not so easy to adjust for, however. Winter flu outbreaks, for instance, appear most years but vary in size and timing.

Departures from Routine Conditions

Better performance might also be obtained by carefully "tuning" the detection algorithm to detect specific types of outbreaks or perhaps one might choose to analyze a syndrome that is less common than ILI. Stoto, Schonlau, and Mariano [SSM04] used the same methods, for instance, to analyze the data on the number of patients with "Viral NOS" (NOS=not otherwise specified) symptoms, which averaged 1 per day. Outside of the flu season, they were able to detect a fast outbreak on day two 50% to 60% of the time, only a small improvement over ILI. With a slow outbreak, however, integrated methods had a 50% chance of detecting outbreaks on day 5 to 7, compared to day 9 for the same chance for ILI.

This improved performance, however, has a cost — it is only sensitive to symptoms that ER physicians would classify as Viral NOS. The combination of fever and rash is rare and suggests the early stages of smallpox. A syndromic surveillance system set up to look at this combination would likely be more effective than the results above suggest, but would only be sensitive to smallpox and not terrorist agents that have other symptoms.

Data also can be analyzed geographically, tuning detection algorithms to outbreaks that are focused in a small geographic area. For instance, if there were an extra 18 cases of ILI in a city, and all lived in the same neighborhood, that would surely be more informative than 18 cases scattered throughout the city — it would suggest a biological agent released at night in that area. This is only effective, however, for such a geographically focused attack. It would not work if terrorists chose to expose people in an office building during the workday or at an airport but the data were analyzed by home address.

Detection Algorithms

Finally, more sophisticated detection algorithms could lead to better performance. The simplest detection algorithms focus on the number of excess cases on a given day (the actual number minus some baseline value). If this is more than some number of standard deviations, an alarm is sounded.

Within this simple statement, however, are many choices, each of which affects the detection algorithm's sensitivity, false-positive rate, and timeliness. First, the normal background level and standard deviation must be determined. As indicated above, many choices — simple to complex — are possible for these variables. Second, the observation period must be chosen. Syndromic surveillance systems typically choose one day as the period for reasons of timeliness; any longer period would require waiting for data before the detection algorithm could be run. However, day-to-day variability in syndromic data due to small numbers sometimes means that adequate sensitivity can only be obtained at the cost of a high, false-positive rate. An alternative, therefore, would be to aggregate data over the period of one week, or to use a running average for the daily value. Both of these solutions are obviously less timely.

Current syndromic surveillance systems are typically set up to monitor eight or more separate sets of symptoms, perhaps in different geographical areas and from different hospital ERs. Doing so increases sensitivity simply because more conditions are monitored. If each set of symptoms has a 1% false-positive rate, however, increasing the number monitored will also increase the number of false positives.

One possibility is to pool data over multiple ERs, perhaps all hospitals in a metropolitan area or state, and indeed that is what cities such as Boston and New York are currently doing. If this results in both the signal and the background increasing proportionally, it will result in a more effective system. If, for instance, nine hospitals in the Washington area report daily, each with a daily average of 3 ILI cases, and outbreaks were nine times as large in the example above, the performance of detection algorithms would be substantially improved. If, however, there were 18 extra cases of ILI in the city and they all appeared in one hospital, this signal would be lost in the noise of the entire city's cases.

An alternative is to search for patterns in the set of symptoms; fever up but rash down, for instance, might lead to better performing detection algorithms. Statistical algorithms to determine whether a departure is sufficient to signal an alarm range from simple to sophisticated. The sophisticated What's Strange About Recent Events (WSARE) system developed at the Real Outbreak and Disease Surveillance (RODS) lab, for instance, is based on Bayesian belief networks [WMC03].

2 Statistical Process Control (SPC)

Quick detection of a change in a probability distribution is the fundamental problem of *statistical process control*. The problem arises in any monitoring situation, and lies at the foundation of the theory and practice of quality control. SPC methods use data to evaluate whether distributional parameters, such as the mean rate of a particular syndrome, have increased to an unacceptable level.

The simplest and best understood version of the problem specifies a one-parameter family of univariate distributions — the most studied family being the normal distribution with unknown mean — and aims to detect a change in the parameter from one value to another as quickly as possible after the change occurs. A number of popular and successful algorithms have been developed for this sort of problem, and a substantial body of theoretical and experimental research has accumulated.

Our interest here is in extending these methods to the problem of syndromic surveillance and, in particular, to the Washington, DC, ER data. That SPC is appropriate for syndromic surveillance is not immediately obvious, particularly since *a priori* one would expect a successful methodology would have

to account for seasonal and perhaps other cycles in the data, and that methods specifically designed to detect monotonic changes in incident rates would outperform conventional SPC methods.

We address these and other issues below. In so doing, we introduce some modified multivariate algorithms that may be applied to health-related data for syndromic surveillance and then compare their performance to univariate SPC methods, both using simulated and actual syndromic surveillance data.

2.1 SPC Background and Literature

Walter A. Shewhart [She31] developed the concept of the *control chart*, a graphical statistical tool to help control the behavior of manufacturing processes, and in so doing became one of the founders of the quality control movement. Shewhart's methodology defined a scientific, statistical framework upon which to base decision-making and hence allow objective decisions to be made about how to manage systems. The field of SPC has since grown from Shewhart's seminal work. An excellent introductory text to quality control and SPC is *Introduction to Statistical Quality Control* by Montgomery [Mon85].

In addition to Shewhart's methodology, the classical approaches to SPC have generally been parametric and univariate. These include the CuSum ("cumulative sum") procedure of Page [Pag54] and Lorden [Lor71], the Bayesian procedure of Shiryayev [Shi63, Shi73] and Roberts [Rob66], and the EWMA ("exponentially weighted moving average") procedure of Roberts [Rob59].

The most basic SPC problem is that of monitoring a sequence of random variables over time with the goal of raising an alarm as soon as possible after the mean becomes too large. The CuSum has optimality properties if the mean experiences a one-time jump increase from one known level to another. However, syndromic surveillance is probably not realistically described by this type of change. Rather, a disease outbreak or bioterrorism attack is likely to be characterized by monotonically increasing numbers of people presenting to an ER as the pathogen spreads or the fraction of those who were exposed who develop symptoms increases (as illustrated in Fig. 3).

This difference would seem to cast doubt on the applicability of SPC to the problem of syndromic surveillance. However, Chang and Fricker [CF99] compared the performance of CuSum and EWMA versus a repeated generalized likelihood ratio (GLR) test designed specifically for the monotone problem. They found that the CuSum and EWMA, appropriately applied, performed surprisingly well in comparison to the GLR test, usually outperforming it, and concluded that the CuSum was probably the best overall choice. This result provides some evidence that the simple SPC methods may perform well in the syndromic surveillance problem.

Multivariate CuSum research has centered around detecting changes in either the normal mean vector or the covariance matrix. Seminal work was by

Hotelling [Hot47] in the manufacture of bomb sights in World War II who developed a Shewhart-like methodology for multivariate observations. More recent research includes Pignatiello and Runger [PR90] and Healy [Hea87]. Pignatiello and Runger [PR90] and Crosier [Cro88], as well as other researchers, have looked at the application of CuSum-like recursions to the product of the observation vector and an assumed known covariance matrix. Others have dealt with multivariate data by applying a number of individual univariate algorithms, one to each marginal distribution [WN85], for example. More detailed background information about multivariate SPC can be found in [Alt85].

2.2 Some Notation and Terminology

In the simple case of detecting a shift from one specific distribution to another, let f_0 denote the *in-control* distribution, which is the desired or preferred state of the system. For syndromic surveillance, for example, this could be the distribution of the daily counts of individuals diagnosed with a particular chief complaint at a specific hospital or within a particular geographic region under normal conditions. Let f_1 denote the *out-of-control* distribution where, under the standard SPC paradigm, this would be a particular distribution representing a condition or state that is important to detect. Within the syndromic surveillance problem, f_1 might be a specific, elevated mean daily count resulting from the release of a bioterrorism pathogen for example.

Let τ be the actual (unknown) time when the process shifts from f_0 to f_1 and let T be the length of time from τ to when an algorithm alarms (which we call the *delay*). We use the notation $\mathrm{E}_\tau(T|T \geq 0)$ to indicate the expected delay, which is the average time it takes an algorithm to alarm *once the shift has occurred*. We also use the notation $\mathrm{E}_\infty(T)$ to indicate the expected time to a false alarm, meaning that $\tau = \infty$ and the process never shifts to the out-of-control distribution.

In the SPC literature, algorithms are compared in terms of the expected time to alarm, where $\mathrm{E}_\infty(T)$ is first set equally for two algorithms and then the algorithm with the smallest $\mathrm{E}_\tau(T|T \geq 0)$, for a particular f_1, is deemed better. Often when conducting simulation comparisons, τ is set to be 0, so the conditioning in the expectation is automatic.

The term *average run length* (ARL) is frequently used for the expected time to alarm, where it is understood that when $\tau = \infty$ the ARL denotes the expected time to false alarm. Similarly, in simulation experiments, the performance of various algorithms is compared by setting the expected time to false alarms to be equal and then comparing ARLs when $\tau = 0$, where it is then understood that the ARL is the mean delay time. In general terms, an algorithm with a smaller ARL has a higher sensitivity for detecting anomalies, though this comes at the expense of an increased false-alarm rate.

For syndromic surveillance, the out-of-control situation can be more than a jump change from f_0 to f_1. For example, if μ_0 is the mean of f_0, then one

possible out-of-control situation might be a monotonic increase in the mean so that for each time $i > \tau$, $\mu_1(i) = \mu_0 + (i - \tau)\delta$, for some positive δ. Yet, even for this type of out-of-control condition, algorithms can still be compared using $\mathrm{E}_\tau(T|T \geq 0)$.

Note that the specific value of τ is generally irrelevant to the analysis. What is important is how long an algorithm takes to alarm after time τ. However, setting $\tau = 0$ means that the algorithm is guaranteed to be in its initial condition when the shift to f_1 occurs (or starts to occur, in the case of something other than a jump change), which may be a help or hindrance to a particular algorithm.

Also, note that comparisons using the expected value are characterizing the distribution of the delay via a single number. This has the advantage of allowing many comparisons to be easily graphically summarized (as we will show), but comes with all the inherent limitations of such summaries. Hence, here we used both the ARL in our initial simulation investigations and then subsequently used the distribution of the delay in the final simulations with actual data.

2.3 Applying SPC to Syndromic Surveillance

This section presents two standard univariate algorithms (the Shewhart and the CuSum) and two multivariate extensions of these two algorithms (Hotelling's T^2 and one of Crosier's multivariate CuSums). Here we also discuss how to apply the univariate algorithms to multivariate syndromic surveillance data and describe how we modified the multivariate algorithms to best apply to the syndromic surveillance problem. We focus on the Shewhart and CuSum algorithms, and not the EWMA, because the EWMA can be made to perform very similarly to either of the Shewhart or CuSum through the appropriate selection of the EWMA's weighting parameter.

Furthermore, we chose to use Shewhart and CuSum SPC methods due to the nature of our data. Specifically, for these particular data:

- The mean rates for each of the syndromic groups were quite constant, and
- The logarithmically transformed counts (not shown here) were quite normally distributed.

It is important to note that most SPC procedures, including those described here, have been developed under the assumption that the observations are independent. In industrial applications, this can often be reasonably well achieved by taking observations sufficiently far apart in time. For syndromic surveillance data that exhibit characteristics such as seasonal cycles or other trends, which we were frankly surprised not to find in our data, other methods such as the EWMA or those proposed by Nomikos and MacGregor [NM95] might be more appropriate and effective.

Univariate Shewhart Algorithm

Shewhart's algorithm [She31] is probably the simplest and best known of all SPC methods and is widely applied in industry. The basic idea is to sequentially evaluate one observation (or period) at a time, alarming when an observation that is rare under f_0 occurs. The most common form of the algorithm, often known as the $\bar{X}$ *chart*, alarms when the absolute value of an observed sample mean exceeds a prespecified *threshold* h, often defined as the mean value plus some number of standard deviations of the mean. There are variants on the algorithm for monitoring the variability of processes and the algorithm can be defined to only alarm for deviations in one direction.

For application to the syndromic surveillance problem, we assume that only deviations in the positive direction that would indicate a potential outbreak are important to detect. For a univariate random variable X, and for some desired probability p, the threshold h is chosen to satisfy

$$\int_{\{x>h\}} f_0(x)\,\mathrm{d}x = p.$$

The algorithm proceeds by observing values of X_i; it stops and concludes $X_i \sim f_1$ at time $\hat{\tau} = \inf\{i : X_i > h\}$.

If the change to be detected is a one-time jump in the mean and the probability of an observation exceeding the threshold is known, then simulation is not required as the delay is geometrically distributed and exact calculations for the average run lengths can be directly calculated as $\mathrm{E}_\infty(T) = 1/p$ and

$$\mathrm{E}_\tau(T|T \geq 0) = \mathrm{E}_0(T) = \left[\int_{\{x>h\}} f_1(x)\,\mathrm{d}x\right]^{-1}.$$

Generally, however, it is quite simple to empirically estimate the ARLs via simulation. For a particular f_0, choose an h and run the algorithm m times, recording for each run the time t when the first $X_i > h$ (where each X_i is a random draw from f_0, of course). Estimate the in-control ARL as

$$\widehat{\mathrm{E}_\infty(T)} = \sum t/m,$$

adjusting h and rerunning as necessary to achieve the desired in-control ARL, where m is made large enough to make the standard error of $\widehat{\mathrm{E}_\infty(T)}$ acceptably small. Having established the threshold h for that f_0 with sufficient precision, then for each f_1 of interest rerun the algorithm n times (where n is often smaller than m), drawing the X_is from f_1 starting at time 1. As before, take the average of $t_1, \ldots, t_n$ to estimate the expected delay.

For the multivariate syndromic surveillance problem, multiple univariate algorithms are applied, one to each data stream. When comparing the performance of simultaneous univariate algorithms applied to multivariate data to a multivariate algorithm it is important to ensure that the expected times to

false alarm are set equally. For multiple univariate algorithms running simultaneously, say j, one must choose how to set the j thresholds. If there is some reason to make the combined algorithms more sensitive to changes in some of the data streams, those thresholds can be set such that the probability of exceeding the threshold(s) is greater in those data streams than in the others. For the purposes of the simulations that follow in this chapter, there was no reason to favor one data stream over another, so all the thresholds were set such that the probability of false alarm was equal for all data streams.

Univariate CuSum Algorithm

The CuSum is a sequential hypothesis test for a change from a known in-control density f_0 to a known alternative density f_1. The algorithm monitors the statistic S_i, which satisfies the recursion

$$S_i = \max(0, S_{i-1} + L_i), \tag{1}$$

where the increment L_i is the log likelihood ratio

$$L_i = \log \frac{f_1(X_i)}{f_0(X_i)}.$$

The algorithm stops and concludes that $X_i \sim f_1$ at time $\hat{\tau} = \inf\{i : S_i > h\}$ for some prespecified threshold h that achieves a desired ARL under the given in-control distribution.

If f_0 and f_1 are normal distributions with means μ and $\mu + \delta$, respectively, and unit variances, then (1) reduces to

$$S_i = \max(0, S_{i-1} + (X_i - \mu) - k), \tag{2}$$

where $k = \delta/2$. This is the form commonly used, even when the underlying data is only approximately normally distributed. For the DC hospital data we examined, the log transformed data was generally very close to normally distributed, so we applied (2) to $\log(X_i)$. Note that k may be set to values other than $\delta/2$ and frequently users specify a value for k rather than the mean of f_1. What is relevant to the performance of the CuSum is that when the process shifts to a state where $\mathrm{E}(X_i) > \mu + k$, then the expected value of the increment$X_i - \mu - k$ is positive and the CuSum S_i tends to increase and subsequently exceed h relatively quickly.

Note that, since the univariate CuSum is "reflected" at zero, it is only capable of looking for departures in one direction. If it is necessary to guard against both positive and negative changes in the mean, then one must simultaneously run two CuSums, one of the form in (2) to look for changes in the positive direction, and one of the form

$$S_i = \max(0, S_{i-1} - (X_i - \mu) - k),$$

to look for changes in the negative direction. For the syndromic surveillance problem, we are only interested in looking for increases in rates, so we only use (2).

As with the univariate Shewhart, multiple univariate CuSum algorithms must be applied, one to each data stream, for the multivariate syndromic surveillance problem. As with the univariate Shewhart algorithms, for the purposes of our simulations, there was no reason to favor one data stream over another, so all the thresholds were set such that the probability of false alarm was equal in all data streams and so that the resulting expected time to false alarm for the combined set of univariate algorithms was equal to the expected time to false alarm of the multivariate algorithm.

Multivariate Shewhart Algorithm (Modified Hotelling's T^2)

Hotelling [Hot47] introduced the T^2 (sometimes referred to as the χ^2) algorithm. For multivariate observations $\mathbf{X}_i \in \mathbb{R}^d$, $i = 1, 2, \ldots$, compute

$$T_i^2 = \mathbf{X}_i' \Sigma^{-1} \mathbf{X}_i,$$

where Σ^{-1} is the inverse of the covariance matrix. The algorithm stops at time $\hat{\tau} = \inf\{i : T_i > h\}$ for some prespecified threshold h.

We refer to this as a multivariate Shewhart algorithm since it only looks at data from one period at a time. Like the original univariate Shewhart $\bar{X}$ algorithm, because it only uses the most recent observation to decide when to stop, it can react quickly to large departures from the in-control distribution, but will also be relatively insensitive to small shifts. Of course, it also requires that the covariance matrix is known or well-estimated.

For the syndromic surveillance problem, it is desirable to focus the T^2 algorithm on the detection of increases in incident rates. We accomplish that by modifying the stopping rule for the T^2 so that it meets two conditions: (1) $T_i > h$ and (2) $\mathbf{X}_i \in \mathcal{S}$, where $\mathcal{S}$ is a particular subspace of $\mathbb{R}^d$ that corresponds to disease outbreaks, for example an increase in one or more data streams.

For the purposes of the syndromic surveillance simulations, we defined $\mathcal{S}$ as follows. Choose values $s_1, s_2, \ldots, s_d$ such that

$$\int_{x_1=s_1}^{\infty} \int_{x_2=s_2}^{\infty} \cdots \int_{x_d=s_d}^{\infty} f_o(\mathbf{x}) \mathrm{d}\mathbf{x} \approx 0.99,$$

and then define $\mathcal{S} = \{x_1 > s_1, x_2 > s_2, \ldots, x_d > s_d\}$.

For example, consider an in-control distribution following a bivariate normal distribution with some positive correlation, so that the probability contour for the density of f_0 is an ellipse with its main axis along a 45-degree line in the plane. Then you can think about $\mathcal{S}$ as the upper right quadrant that almost encompasses the 99% probability ellipse.

The idea of using this region for $\mathcal{S}$ is that if f_1 represents a shift in the mean vector in any direction corresponding to an increase in one or more of the data streams, then the modified T^2 algorithm will have an increased probability of alarming, which should result in a decreased expected time to alarm. On the other hand, if f_1 represents a condition where the mean vector corresponds to a decrease in one or more of the data streams, then the probability of alarming will decrease and the algorithm will have less of a chance of producing an alarm.

Multivariate CuSum Algorithm (Modified Crosier's MCuSum)

The abbreviation MCuSum, for multivariate CuSum, is used here to refer to the algorithm proposed by Crosier [Cro88] that at each time i considers the statistic

$$\mathbf{S}_i = (\mathbf{S}_{i-1} + \mathbf{X}_i - \boldsymbol{\mu})(1 - k/C_i), \text{ if } C_i > k, \tag{3}$$

where k is a statistical distance based on a predetermined vector $\mathbf{k}$, $k = \{\mathbf{k}'\Sigma^{-1}\mathbf{k}\}^{1/2}$ and $C_i = \{(\mathbf{S}_{i-1} + \mathbf{X}_n - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{S}_{i-1} + \mathbf{X}_i - \boldsymbol{\mu})\}^{1/2}$. If $C_i \leq k$, then reset $\mathbf{S}_i = \mathbf{0}$. The algorithm starts with $\mathbf{S}_0 = \mathbf{0}$ and sequentially calculates

$$Y_i = (\mathbf{S}_i'\Sigma^{-1}\mathbf{S}_i)^{1/2}.$$

It concludes that $X_i \sim f_1$ at time $\hat{\tau} = \inf\{i : Y_i > h\}$ for some threshold $h > 0$.

Crosier proposed a number of other multivariate CuSum-like algorithms but generally preferred (3) after extensive simulation comparisons. Pignatiello and Runger [PR90] proposed other multivariate CuSum-like algorithms as well, but found that they performed similarly to (3).

It is worth noting that Crosier derived his algorithm in an ad hoc manner, not from theory, but found it to work well in simulation comparisons. Healy [Hea87] derived a sequential likelihood ratio test to detect a shift in a mean vector of a multivariate normal distribution that is a true multivariate CuSum. However, while we found Healy's algorithm to be more effective (had shorter ARLs) when the shift was to the precise f_1 mean vector, it was less effective than Crosier's for detecting other types of shifts, including mean shifts that were close to but not precisely the specific f_1 mean vector.

In this application we prefer Crosier's algorithm to Healy's since it seems to be more effective at detecting a variety of departures from the in-control mean vector and the types of shifts for the syndromic surveillance problem are not well-defined. That is, if we knew the type of departure to look for, we could design a detection algorithm that would have more power to detect that specific signal. However, given that the types of signals will vary, we have opted for Crosier's method because it is robust at detecting many types of departures well.

We also prefer Crosier's formulation for the syndromic surveillance problem as it is easy to modify to look only for positive increases. In particular, in

our simulations, when $C_i > k$ we bound $\mathbf{S}_i$ to be positive in each data stream by replacing (3) with $\mathbf{S}_i = (S_{i,1}, \ldots, S_{i,d})$ where

$$S_{i,j} = \max[0, (S_{i-1,j} + X_{i,j} - \mu_j)(1 - k/C_i)],$$

for $j = 1, 2, \ldots, d$.

2.4 Performance Comparisons via Abstract Simulations

Before evaluating the performance of the methods using actual data, we compared their performance using simulated data from normal and multivariate normal distributions. The purpose of these simulations was to:

1. Compare and contrast the performance of the methods under known, ideal conditions;
2. Gain some insight into how they performed as the dimensionality of the data changed; and
3. Reach some preliminary conclusions about how best to implement the algorithms for the real data.

In these simulations, we compared the performance by average run length, first setting the ARL under the in-control distribution (i.e., $\mathrm{E}_\infty(T)$, the expected time to false alarm) equally, and then comparing the ARL performance under numerous out-of-control distributions resulting from various shifts in the mean vector at time 0 (i.e., $\mathrm{E}_0(T)$).

For example, Fig. 4 illustrates the improved performance of the modified T^2 algorithm and the modified MCuSum regardless of dimensionality and size of (a positive) mean shift. Here (and in the other figures in this section) the in-control distribution is a six-dimensional multivariate normal centered at the zero vector with unit variance in all the dimensions and covariance $\varrho = 0.3$ between all the dimensions; that is, the in-control distribution is

$$f_0 = N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{pmatrix}\right).$$

The out-of-control distributions are the same as the in-control distributions but with components of the mean vector shifted as indicated on the horizontal axis for the number of dimensions shown in the key. So, for example, the darkest line is for a mean vector that was shifted in all six dimensions from 0.0 — no shift — on the left to 3.4 on the right.

The vertical axis in Fig. 4 is the difference (Δ) between the ARL for the unmodified algorithm and the modified algorithm for a given mean vector shift (measured only at the values indicated on the horizontal axis). Positive values

indicate the modified algorithm had a smaller ARL and so performed better, so that for a particular out-of-control condition the modified algorithm had a shorter time to alarm. A difference of 0 at mean shift = 0.0 indicates that the false-alarm rates (equivalently, the in-control ARLs) were set equally for each algorithm before comparing the expected time to alarm for various out-of-control mean vector shifts (within the bounds of experimental error, where a sufficient number of simulation runs were conducted to achieve a standard error of approximately 2.5 on the estimated in-control ARLs).

Figure 4 shows, as expected, that the modified algorithms perform better than the original algorithms at detecting positive shifts regardless of whether the shift occurs in one dimension, in all the dimensions, or in some number of dimensions in-between, and for all magnitudes of shift. As the number of dimensions experiencing a shift of a given size increases, the modified algorithms do considerably better. However, for the largest shifts, the performance of the original algorithms approaches that of the modified algorithms.

Not shown here, the results for other low-to-moderate values of ϱ, from $\varrho = 0$ to $\varrho = 0.9$, are very similar. Only for large ϱ and small shifts in a low number of dimensions does the original MCuSum algorithm best the modified algorithm. However, in our actual data the covariances between chief complaints, both within and between hospitals, whether aggregated or not, tended to be quite low, generally less than 0.1 and never greater than 0.3.

A further benefit of the modified algorithms, at least in terms of syndromic surveillance, is that they will not alarm if incidence rate(s) decrease. While a decrease in rates might be interesting to detect for some purposes, for the purpose of syndromic surveillance such detection would constitute a false alarm. In addition, because these multivariate algorithms only look for positive shifts, they can be directly compared to multiple one-sided univariate algorithms operating simultaneously.

Given that the modified T^2 performs better than the original T^2 for this problem, Fig. 5 focuses the performance of the modified T^2 as compared to six one-sided Shewhart algorithms operating simultaneously. The comparison is shown in two different ways: in terms of the distance of a shift measured in the direction of one or more of the axes ("on axes" in the left graph), or in terms of the distance of a shift "off axes" (right graph). At issue is that the univariate algorithms are direction specific, meaning they are designed to look for shifts along the axes. The multivariate algorithms are direction invariant, meaning they are just as effective at detecting a shift of distance x whether the shift occurs in the direction of one or more axes ("on axes") or in some other direction ("off axes").

The left-side graph of Fig. 5, constructed just like Fig. 4, shows that six simultaneous univariate Shewharts are more effective (have shorter ARLs) than the modified T^2 when the shift occurs on axes. At best, for large shifts, the ARL of the modified T^2 is equivalent to the multiple univariate Shewharts, and for smaller shifts (roughly > 0.0 to 1.0) the multiple univariate Shewharts are clearly better.

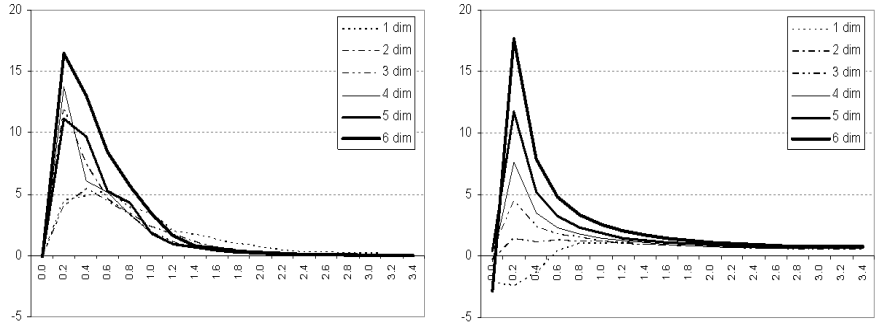

Fig. 4. Performance comparison of the T^2 and MCuSum algorithms versus their modified counterpart algorithms. The modified algorithms (T^2 on the left and MCuSum on the right) perform better than the original algorithms at detecting positive shifts regardless of whether the shift occurs in one dimension, in all the dimensions, or in some number of dimensions in-between, and for all magnitudes of shift.

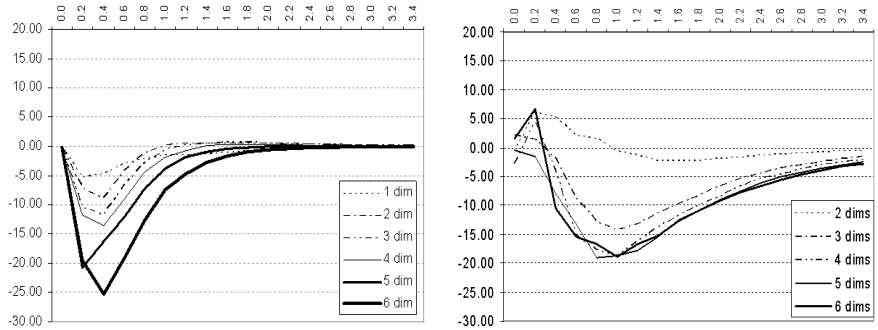

Fig. 5. Performance comparison of the modified T^2 algorithm versus multiple simultaneous univariate Shewhart algorithms for $\varrho = 0.3$. The multiple simultaneous Shewhart algorithms generally have smaller ARLs except for small "off axis" shifts.

The graph on the right side of Fig. 5 is constructed differently. The horizontal axis of this graph shows the *distance* of the shift, where the shift is in the number of dimensions indicated in the key, and was constructed so that the projection of the shift onto the axes for those dimensions was equal. That is, for a shift of distance l in n dimensions, the mean vector component for each of the affected dimensions shifted from 0 under f_0 to $l/\sqrt{n}$ under f_1 (and where in the other 6–n dimensions, the mean vector components remain unchanged at 0).

This type of shift is the most extreme off-axis type of shift (meaning for a given distance l, the maximum projection on the nonzero axes was the smallest) and here we see a result similar to Fig. 5, except that the modified T^2 does better than the simultaneous univariate Shewhart algorithms for very small shifts.

Why is the distinction between the two types of shifts (on-axes versus off-axes) relevant? Well, if each of the types of bioterrorism events to be detected will manifest itself in the data being monitored as a separate increase in one of the data streams, such as ER admit counts for a particular chief complaint, then thinking about and optimizing the detection algorithm to look specifically for shifts along the axes makes sense. On the other hand, for a bioterrorism event that will manifest itself as changes in a number of dimensions of the data being monitored, such as with less specific health data that in combination may increase, it makes sense to provide for an event that manifests itself more like a latent variable and hence appearing most strongly in some off-axes direction.

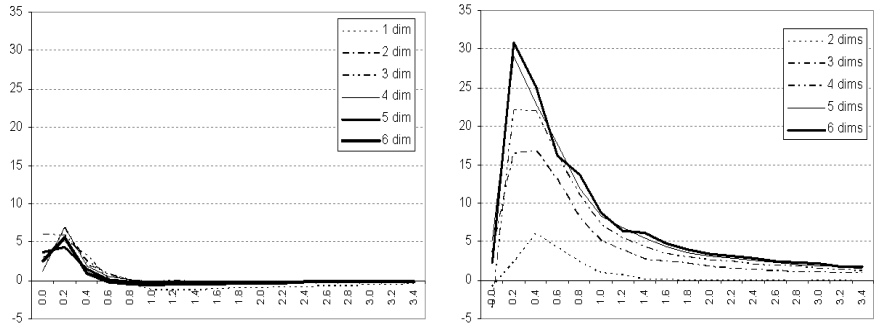

Fig. 6. Performance comparison of the modified MCuSum algorithm versus multiple simultaneous univariate CuSum algorithms. The modified MCuSum tends to have smaller ARLs whether the shift is along the axes (left) or whether the shift is off the axes (right).

Given that the goal is a robust methodology to guard against either possibility, the results for the simultaneous univariate Shewharts versus the modified T^2 are mixed. However, the results for the modified MCuSum versus simultaneous univariate CuSums presented in Fig. 6 differ in that the modified MCuSum is generally better than the simultaneous univariate CuSums regardless of whether the shift is on- or off-axis. In particular, in the left graph of Fig. 6 the modified MCuSum performance is substantially better for small shifts (roughly > 0.0 to 0.5 or so), equivalent for large shifts (roughly > 3.0), and only marginally degraded for other shifts, with an ARL difference of less

than 1. As expected, in the right graph of Fig. 6 the modified MCuSum performance is better than or, for very large shifts, equivalent to the simultaneous univariate CuSums.

Though not shown here, these results also hold for a range of low to moderate correlations, from $\varrho = 0$ to $\varrho = 0.6$. Hence, these results would tend to indicate that the modified MCuSum would be preferable to simultaneous univariate CuSums for detecting a variety of types of mean shifts. What remains, then, is a comparison of the modified MCuSum to either the multiple univariate Shewharts or the modified T^2 in those scenarios where each does better.

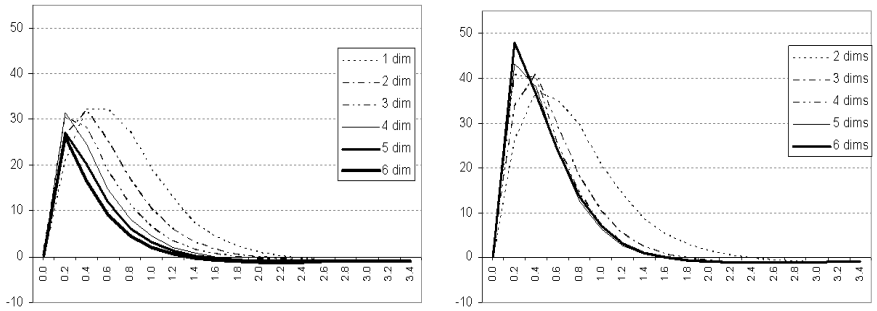

Fig. 7. Performance comparison of the modified MCuSum algorithm to multiple simultaneous univariate Shewhart algorithms (left) and to the modified T^2 algorithm (right). Whether the shift is on-axes or off-axes, the modified MCuSum algorithm performs better than the preferred Shewhart-type algorithm.

Figure 7 provides this comparison: to the simultaneous univariate Shewharts when the shift is on-axes (left graph) and to the modified T^2 when the shift is off-axes (right graph). In both cases, the modified MCuSum algorithms' performance is better. Our conclusion, then, is a preference for the modified MCuSum, at least in these simulations for a jump change in the mean vector of multivariate normal distributions with moderate covariance. In the next section, then, we further examine the performance of these methods using real data and more realistic shifts to evaluate the performance of the algorithms.

3 A Simulation Study Using DC ER Syndromic Surveillance Data

One would expect that a properly designed multivariate algorithm would be more effective — particularly, more sensitive and timely when the false-positive rate is controlled — than standard univariate methods. However, as the previous section demonstrated, some multivariate methods are better than others, and there are situations in which simultaneous univariate algorithms are preferable. Furthermore, since the evaluations in the SPC literature tend to focus on a jump change in the mean, as did the evaluations in the preceding section, it does not necessarily follow that those results will directly apply to the syndromic surveillance problem in which the mean will likely change in some monotonically increasing fashion. Hence, to evaluate the univariate and multivariate algorithms described in Sect. 2, we also conducted a simulation study on data from the DC Department of Health ER syndromic surveillance system and then evaluated how the algorithms performed under a series of outbreak scenarios.

3.1 Data and Methods

As baseline data for our simulation study we used data on the daily number of ER admissions for four syndromic group "chief complaints" (unspecified infection, rash, respiratory complaints, and gastrointestinal complaints) from seven Washington, DC, hospitals with relatively complete data. Of the eight syndromic groups available, these four were chosen because they are the most common and, in univariate analyses, are most effective at detecting disease outbreaks. The data on the resulting 28 data streams (4 syndromic groups x 7 hospitals) span the period of September 2001 through May 2004 (with missing data imputed as required [SJF04] to simplify the comparisons of the detection algorithms).

This data provides the naturally occurring incident rates and variation in the hospital ERs for the four syndromic groups. We then "seeded" these data in various ways, meaning we added extra cases to the data, to simulate a bioterrorism event. In the base case, Scenario A, we seeded the data adding 1 additional observation on day τ, 2 additional on day $\tau+1$, and so on up to 10 on day $\tau+9$ for each of the 28 data streams resulting in a total of 1,540 extra cases over 10 days. Scenario A is intended to represent a bioterrorism event that manifests itself in multiple ways across the entire population. Hence, all of the chief complaints increase in all the hospital ERs.

In contrast, we defined Scenario C to represent a situation in which the outbreak shows up in one syndromic group only, so we only seeded the "unspecified infection" syndromic group only for all seven hospitals adding 1 additional observation on day τ, 2 additional on day $\tau+1$, and so on up to 10 on day $\tau+9$ (for a total of 385 extra cases).

Since the total number of cases added in Scenario C is only one-quarter of that of the base scenario, we also constructed Scenario CA in which the seed was increased to 4 on day τ, 8 on day $\tau + 1$, and so on, resulting in 220 extra unspecified infection cases in each of seven hospitals, which is a total of 1,540 extra cases over 10 days. Hence, like Scenario C, Scenario CA represents an event that manifests itself in only one syndromic group but with the magnitude of Scenario A.

Scenarios D and DA repeat this with a focus on hospitals rather than syndromic groups. In Scenario D we seeded the data adding 1 additional observation on day τ, 2 additional on day $\tau + 1$, and so on up to 10 on day $\tau + 9$ for every syndromic group but in only one medium-sized hospital. In Scenario DA, the seed was increased to 7 on day τ, 14 on day $\tau + 1$, and so on, resulting in 385 extra cases in each syndromic group in only one hospital. So, Scenarios D and DA represent an outbreak in a smaller geographic region, with Scenario D being of a smaller magnitude and Scenario DA having the magnitude of Scenario A.

These five scenarios were chosen to represent the extremes of a range of ways in which a real bioevent might occur. (As the gap in the naming convention suggests, we investigated other scenarios as well, but do not present them here.) Some might regard Scenario C, in which the outbreak is concentrated in only one syndromic group, as the most likely of the scenarios. However, we expect that any real outbreak will look like some combination of these scenarios, so detection algorithms that work well across the test scenarios are likely to be effective in actual practice.

Given these scenarios, we then compared the performance of the algorithms described in Sect. 2.3 and a trend-adjusted CuSum (see Stoto et al. [SSM04] for additional detail) applied in two ways. First, we applied simultaneous univariate algorithms or one multivariate algorithm to the individual 28 data streams, setting the detection threshold empirically so that the probability of an alarm outside the flu season (i.e., the false-alarm rate) was 1%. Second, as an alternative to reduce the dimensionality of the problem, we first summed the total number of cases across all hospitals in each of the four syndromic groups and then applied either simultaneous univariate algorithms or a multivariate algorithm to the resulting four data streams (again setting the false-alarm rates equal at 1%).

To carry out the simulation we began by setting $\tau = 1$ and adding the appropriate seed on days 1 through 10 of the dataset. We repeated this setting $\tau = 2$ and adding the appropriate seed on days 2 through 11, and so on, until we had created 970 alternative datasets. We then applied the detection algorithms to each alternative dataset and calculated the proportion of times that each algorithm alarmed on day τ through $\tau + 9$, the first day of the simulated bioevent to the 9th day after the simulated bioevent, to estimate the sensitivity of the detection algorithm. Because we expect performance to differ by season, the results are calculated separately for the flu season (defined as December 1–April 30) and the rest of the year.

3.2 Results

Figure 8 compares the performance of the modified MCuSum ("MV" in the key) and simultaneous univariate CuSum methods ("Z" in the key) outside of the flu season for all five scenarios — just one summary result of the many simulations we ran. Unlike the graphs presented in the previous section, showing estimated ARLs for various changes in the mean, Fig. 8 plots the probability of detection (which can be interpreted as an estimated probability of alarm) for each algorithm under each scenario by day of the outbreak.

Note first that the probability of detection on day 0, that is, the day before the outbreak begins, is 1% for each detection algorithm, the false-alarm rate we set. Focusing first on Scenario A (in which the seed appears in all 28 data streams), the results show that in 18% of the sample datasets the simultaneous univariate CuSum algorithms (dashed line with open circles) alarm on day 2 of the outbreak, increasing to 67% on day 3 and 100% on day 4 and higher. In this scenario, the modified MCuSum (solid line with open circles) does slightly better. The probabilities of alarming on days 2, 3, and 4 are 36%, 93%, and 100%, respectively.

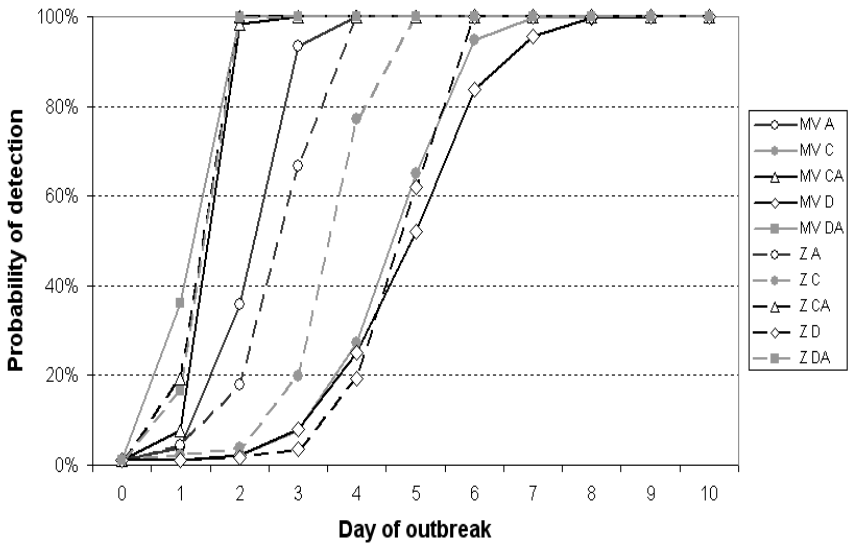

Fig. 8. Comparison of the modified MCuSum ("MV" in the key, solid lines in the graph) and the simultaneous univariate CuSums ("Z" in the key and dashed lines in the graphs) in terms of probability of outbreak detection by day of outbreak for Scenarios A, C, CA, D, and DA previously defined.

In Scenario C (closed circles), in which the simulated outbreak is concentrated in only one syndromic group and consequently involves only one-quarter the number of cases of Scenario A, both detection algorithms not surprisingly do less well. On day 4 of the simulated outbreak the univariate CuSums have only a 77% probability of alarming, and the modified MCuSum only a 27% probability of alarming. The results for Scenario CA (triangles), however, show that most of the reason for the poorer performance is that there are fewer excess cases. In this scenario the univariate and multivariate CuSum algorithms have 100% and 98% probabilities of alarming, respectively, by day 2. Note that in Scenario C by day 4 a total of 10 excess cases of unspecified infection have been seen in each hospital, and in Scenario CA there are 40 excess cases. The average daily number of such cases in the baseline data is less than 1 for two of the hospitals, between 3 and 6 for four hospitals, and over 30 for one hospital in the analysis.

Scenarios D and DA, in which the outbreak is concentrated in only one hospital, show similar results. In Scenario D, which involves only one-seventh the number of cases as Scenario A (open circles), the univariate CuSums alarm probability reaches 62% only on day 5, and the modified MCuSum only reaches 52% on that day. With the same number of cases as in Scenario A, the performance of both algorithms improves under Scenario DA (squares). Both reach a 100% alarm rate by day 2.

Comparing the performance of the two CuSum algorithms across these scenarios, it is difficult to conclude that one is better than the other. The modified MCuSum does noticeably better than the univariate CuSums in Scenario A (solid versus dashed lines with open circles, respectively), but worse in Scenario C (closed circles). In the other scenarios their performance is similar.

To summarize this type of comparison for all of the algorithms we tested, we calculated a performance index, defined as Σ_iProb(detection on day i) for $i = 1$ to 10. This is essentially the area under the curve in Fig. 8. In Scenario A, for instance, the performance index for the modified MCuSum is 8.34 and for the univariate CuSums is 7.90. In Scenario C, the situation is reversed: 5.99 versus 7.04. To get a sense of the range of the performance index, the best performance represented in Fig. 8 is the modified MCuSum in Scenario DA, with a performance index of 9.37, and the worst is the modified MCuSum in Scenario D, with a performance index of 5.68.

Table 1 displays the performance index for 12 detection algorithms for the five scenarios. "Z" results are univariate analyses operating on all 28 data streams, and "C" results sum the syndromic group data over seven hospitals (resulting in four data streams). In both cases we investigate simultaneous univariate Shewhart (1a and 3a), CuSum (1b and 3b), and trend-adjusted CuSum (1c and 3c) algorithms. "MV" results are multivariate algorithms, and "CMV" results are for multivariate algorithms but with the data summed across hospitals as in the C results: Hotelling's T^2 (5a, 6a), the modified T^2 (5b, 6b), and the modified MCuSum (7, 8).

Table 1. Comparison of performance of the univariate and multivariate algorithms for the five scenarios using a performance index of Σ_iProb(detection on day i) for $i = \tau$ to τ+9. The "Z" results (1a, 1b, and 1c) are simultaneous univariate algorithms operating simultaneously on all 28 series. The "C" results (3a, 3b, and 3c) are also simultaneous univariate algorithms operating on syndromic group data summed over the seven hospitals. "MV" results (5a, 5b, and 7) and "CMV" results (6a, 6b, and 8) are multivariate algorithms operating on the 28 data streams and the summed four data streams, respectively. These include Hotelling's T^2 (5a and 6a), the modified T^2 (5b, 6b), and the modified MCuSum (7 and 8)

	Scenario				
Performance Indices	A	C	CA	D	DA
"Z" Algorithms					
1a – Shewhart	6.34	5.76	9.18	3.30	9.34
1b – CuSum	7.90	7.04	9.20	5.89	9.17
1c – Trend-adjusted CuSum	1.64	0.68	9.34	0.24	9.34
"C" Algorithms					
3a – Shewhart	8.03	4.78	9.08	0.67	8.03
3b – CuSum	2.00	0.16	0.94	0.49	2.00
3c – Trend-adjusted CuSum	8.22	5.97	9.08	0.62	8.22
"MV" Algorithms					
5a – Hotelling's T^2	6.71	3.27	8.71	0.87	9.03
5b – Modified T^2	7.40	4.30	8.22	1.87	9.03
7 – Modified MCuSum	8.34	5.99	9.07	5.68	9.37
"CMV" Algorithms					
6a – Hotelling's T^2	7.63	0.92	7.68	0.29	7.63
6b – Modified T^2	8.00	1.35	7.76	0.52	8.00
8 – Modified MCuSum	8.43	0.58	6.42	1.75	8.43

As displayed in Table 1, these results suggest that no one or two detection algorithms clearly dominate the others across all five of the scenarios tested. However, the two best are the simultaneous univariate CuSums and modified MCuSum algorithms (Z-1b and MV-7), which are the focus of Fig. 8. Each has a performance index in the 8 to 10 range for scenarios A, CA, and DA, but in the 5 to 7 range for scenarios C and D.

Pooling data across hospitals is a common way to analyze multiple data streams, the rationale being that the signal is more likely to emerge above the random variability. Our results, however, suggest that at least for the scenarios we used, algorithms operating on the pooling data (the C and CMV results) were less effective than those same algorithms operating on the unpooled data (Z and MV results).

Among the unpooled data for the simultaneous univariate algorithms (the Z results), the standard CuSum algorithm (1b) performs at least as well

and usually better than the Shewhart algorithm (1a) and the trend-adjusted CuSum (1c). That the CuSum performs better than the Shewhart algorithm should be expected since the CuSum is better at detecting small changes and, in our scenarios, the outbreaks all begin with relatively small increases early on. However, in contrast, with the pooled data (C results) the standard CuSum (3b) performs substantially less well than the alternatives (3a and 3c).

Stoto et al. [SJF04] extend these results by investigating other detection algorithms and performance outside the flu season and perform various sensitivity analyses.

It should be noted that these results are potentially sensitive to many arbitrary choices that had to be made in the details of the detection algorithms tested and the design of the simulation. The performance of CuSum methods, for instance, depends on the choice of the parameter k, and may be better or worse for fast- or slow-growing outbreaks. The CuSum also depends on the estimated mean count $\hat{\mu}_0$ used as the baseline to calculate departures for each series. The trend-adjusted CuSum method depends on the weighting parameter λ in the exponentially weighted moving average.

In addition, we chose to set the false-alarm rate to 1% outside the flu season, which we arbitrarily defined as December 1–April 30; a different set of dates may have given different results. Our simulated outbreaks used seeds of the same size in every hospital, ignoring substantial variability in the background ER admission rates; again, a different and possibly more realistic choice might lead to different results. Finally we should note that the results also depend on the particular dataset used as the baseline for the simulation. The results are likely to apply to similar data in the future, but may be different for syndromic surveillance systems in cities other than Washington, DC.

These results show roughly similar performance for the simultaneous univariate CuSum and modified MCuSum algorithms, with one better than the other or both having similar performance characteristics depending on the scenario. In contrast, the abstract simulations in Sect. 2.4 show that the modified MCuSum has a clear advantage when the shift to be detected is "off-axes" and seems to show some performance improvements over the simultaneous univariate CuSum algorithm even when the shifts are on-axes. Whether these differences are the result of the simulation choices (jump change in the mean versus gradual increase, for example) or some other factor or factors remains to be determined.

However, some conclusions are clear:

- CuSum and CuSum-like algorithms are preferable to Shewhart and Shewhart-like algorithms for syndromic surveillance applications.
- For multivariate algorithms, appropriately modifying the algorithms to look only for increases in rates, such as we did in Sect. 2.3, provides additional detection power in syndromic surveillance applications.

- When designing, implementing, and comparing syndromic surveillance algorithms it is critical to ensure the appropriate thresholds are chosen to achieve a common aggregate false-alarm rate.
- While the CuSum algorithms generally performed better than the others we evaluated, unless the bioevent is so large so as to be obvious, a syndromic surveillance system will take some time to detect the incident — likely on the order of 2 to 5 days, depending on the size of the incident, for a system using data similar to what we have evaluated here.

4 Discussion

Out of concern about the possibility of bioterrorist attacks, many health departments throughout the United States and elsewhere are energetically developing and implementing a variety of syndromic surveillance systems. Our analyses suggest that while these systems may be valuable, their effectiveness for this purpose has not yet been demonstrated, and health departments ought to be cautious in investing in this area and take the time and effort to evaluate the performance of proposed systems in their own setting.

The central problem is that syndromic surveillance has been sold on the basis that it is able to detect outbreaks hours after people begin to develop symptoms, but our analyses suggest that unless the number of people affected is exceptionally large, it is likely to be a matter of days before enough cases accumulate to trigger detection algorithms. Of course, if the number of people coming to emergency departments is exceptionally large, sophisticated detection systems are simply not needed — the incident will be obvious. Further, the window (in terms of number of excess cases and time) between what is reasonably detectable with a syndromic surveillance system, and what is obvious, may be small.

Although an increasing number of statistically sophisticated detection algorithms have been developed, there is a limit to their efficacy. More generally, detection algorithms can be tuned to particular types of outbreaks (e.g., those that are geographically focused), but are only effective if the terrorists choose a matching method of exposing people. Moreover, as Stoto, Schonlau, and Mariano [SSM04], Reingold [Rei03], and others have pointed out, the value of an alarm system is limited by what happens when the alarm goes off. Simply knowing that there are an excess number of people with flulike symptoms is not enough, in itself, to initiate or guide a public health response.

Syndromic surveillance systems, however, can serve other public health purposes. The information technology that has been developed in many cities and states is truly impressive, and many health departments have worked hard to build relationships with hospitals and other entities in their communities to get access to data. The resulting systems and relationships would have additional value for detecting food-borne disease and other outbreaks. For many public health issues, for instance, knowing what is happening in a matter

of days rather than weeks or months would indeed be a major advance for state and local health departments. During the cryptosporidium outbreak in Milwaukee in 1993, for instance, a syndromic surveillance system would have made health officials aware of the outbreak weeks/months before they actually were [MNG98].

Indeed, syndromic surveillance might prove to be most useful in determining the arrival of influenza in a community each year and in helping to determine whether pandemic flu has emerged. Nationally, influenza surveillance is based on a network of sentinel physicians who report weekly on the proportion of their patients with influenzalike symptoms, plus monitoring deaths attributed to influenza or pneumonia in 122 cities. Laboratory analysis to determine whether a case is truly the flu, or to identify the strain, is only rarely done [CDC04b]. Whether the flu has arrived in a particular state or local area, however, is largely a matter of case reports, which physicians often do not file. Pandemic influenza, in which an antigenic shift causes an outbreak that could be more contagious and/or more virulent, and to which few people are immune by virtue of previous exposure, is a growing concern [WW03]. Syndromic surveillance of flulike symptoms might trigger more laboratory analysis than is typically done and hasten the public health response.

References

[Alt85] Alt, F. B. 1985. "Multivariate quality control." In *Encyclopedia of statistical science*, edited by S. Kotz and N. L. Johnson, Volume 6. New York: John Wiley & Sons.

[BBM04] Buckeridge, D. L., H. Burkom, A. Moore, J. Pavlin, P. Cutchis, and W. Hogan. 2004. "Evaluation of syndromic surveillance systems — design of an epidemic simulation model." *Morbidity and Mortality Weekly Report* 53 (Supplement): 137–143.

[Bue04] Buehler, J. W. 2004. "Review of the 2003 National Syndromic Surveillance Conference – Lessons learned and questions to be answered." *Morbidity and Mortality Weekly Report* 53 (Supplement): 18–22.

[CDC04a] Centers for Disease Control and Prevention. 2004. "Framework for evaluating public health surveillance systems for early detection of outbreaks; recommendations from the CDC Working Group." *Morbidity and Mortality Weekly Report* 53 (RR-5): 1–13.

[CDC04b] Centers for Disease Control and Prevention. 2004. Overview of influenza surveillance in the United States fact sheet. http://www.cdc.gov/flu/weekly/pdf/flu-surveillance-overview.pdf accessed January 28, 2005.

[CF99] Chang, J. T., and R. D. Fricker, Jr. 1999. "Detecting when a monotonically increasing mean has crossed a threshold." *Journal of Quality Technology* 31:217–233.

[Cro88] Crosier, R. B. 1988. "Multivariate generalizations of cumulative sum quality control schemes." *Technometrics* 30:291–303.

[Hea87] Healy, J. D. 1987. "A note on multivariate CuSum procedures." *Technometrics* 29:409–412.

[Hen04] Henning, K. J. 2004. "What is syndromic surveillance?" *Morbidity and Mortality Weekly Report* 53 (Supplement): 7–11. Syndromic Surveillance: Reports from a National Conference, 2003.

[Hot47] Hotelling, H. 1947. "Multivariate quality control — Illustrated by the air testing of sample bombsights." In *Techniques of statistical analysis*, edited by C. Eisenhart, M. W. Hastay, and W. A. Wallis, 409–412. New York: McGraw-Hill.

[Lor71] Lorden, G. 1971. "Procedures for reacting to a change in distribution." *Annals of Mathematical Statistics* 42:1897–1908.

[Mon85] Montgomery, D. C. 1985. *Introduction to statistical quality control*, 2nd ed. New York: John Wiley & Sons.

[MNG98] Morris, R. D., E. N. Naumova, and J. K. Griffith. 1998. "Did Milwaukee experience waterborne cryptosporidiosis before the large documented outbreak in 1993?" *Epidemiology* 9:264–270.

[NM95] Nomikos, P., and J. F. MacGregor. 1995. "Multivariate SPC charts for monitoring batch processes." *Technometrics* 37:41–59.

[Pag54] Page, E. S. 1954. "Continuous inspection schemes." *Biometrika* 41:100–115.

[PR90] Pignatiello, J. J., Jr., and G. C. Runger. 1990. "Comparisons of multivariate CuSum Charts." *Journal of Quality Technology* 3:173–186.

[Rei03] Reingold, A. 2003. "If syndromic surveillance is the answer, what is the question?" *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 1 (2): 1–5.

[Rob59] Roberts, S. W. 1959. "Control chart tests based on geometric moving averages." *Technometrics* 1:239–250.

[Rob66] Roberts, S. W. 1966. "A comparison of some control chart procedures." *Technometrics* 8:411–430.

[She31] Shewhart, W. A. 1931. *Economic control of quality of manufactured product*. Princeton, NJ: D. van Nostrand Company.

[Shi63] Shiryayev, A. N. 1963. "On optimum methods in quickest detection problems." *Theory of Probability and its Applications* 8:22–46.

[Shi73] Shiryayev, A. N. 1973. *Statistical sequential analysis*. Providence, RI: American Mathematical Society.

[Sie04] Siegrist, D. 2004. "Evaluation of algorithms for outbreak detection using clinical data from five U. S. cities." Technical Report, DARPA Bio-ALERT Program.

[SP04] Siegrist, D., and J. Pavlin. 2004. "Bio-ALERT biosurveillance detection algorithm evaluation." *Morbidity and Mortality Weekly Report* 53 (Supplement): 152–157.

[Sta04] Stacey, D. 2004, November. Simulating pharmaceutical sales and disease outbreaks based on actual store sales and outbreak data. *2004 Syndromic Surveillance Conference, Boston, MA*.

[SJF04] Stoto, M. A., A. Jain, R. D. Fricker, Jr., J. O. Davies-Cole, S. C. Washington, G. Kidane, C. Glymph, G. Lum, and A. Adade. 2004, November. Multivariate methods for aberration detection: A simulation study using DC ER data. *2004 Syndromic Surveillance Conference, Boston, MA*.

[SSM04] Stoto, M. A., M. Schonlau, and L. T. Mariano. 2004. "Syndromic surveillance: Is it worth the effort?" *Chance* 17:19–24.

[WTE01] Wagner, M., F. C. Tsui, J. Espino, V. Dato, D. Sittig, R. Caruana, L. McGinnis, D. Deerfield, M. Druzdzel, and D. Fridsma. 2001. "The emerging science of very early detection of disease outbreaks." *Journal of Public Health Management and Practice* 7:50–58.

[WW03] Webby, R. J., and R. G. Webster. 2003. "Are we ready for pandemic influenza?" *Science* 302:1519–1522.

[WMC03] Wong, W. K., A. Moore, G. Cooper, and M. Wagner. 2003. "Bayesian network anomaly pattern detection for disease outbreaks." Edited by T. Fawcett and N. Mishra, *Proceedings of the Twentieth International Conference on Machine Learning*. Menlo Park, CA: AAAI Press, 808–815.

[WN85] Woodall, W. H., and M. M. Ncube. 1985. "Multivariate CuSum quality control procedures." *Technometrics* 27:285–292.