PUBLISHED BY THE INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Advances in Disease Surveillance

VOLUME 3
NUMBER 1
DATE 2007

## ARTICLES

# Directionally Sensitive Multivariate Statistical Process Control Procedures with Application to Syndromic Surveillance

**Ronald D. Fricker, Jr.**

Operations Research Department, Naval Postgraduate School, 1411 Cunningham Road, Monterey, California.

*Received for publication March 20, 2006; accepted for publication December 6, 2006.*

Current syndromic surveillance systems run multiple simultaneous univariate procedures, each focused on detecting an outbreak in a single data stream. Multivariate procedures have the potential to better detect some types of outbreaks, but most of the existing methods are directionally invariant and are thus less relevant to the problem of syndromic surveillance. This article develops two directionally sensitive multivariate procedures and compares the performance of these procedures both with the original directionally invariant procedures and with the application of multiple univariate procedures using both simulated and real syndromic surveillance data. The performance comparison is conducted using metrics and terminology from the statistical process control (SPC) literature with the intention of helping to bridge the SPC and syndromic surveillance literatures. This article also introduces a new metric, the average overlapping run length (AORL), developed to compare the performance of various procedures on limited actual syndromic surveillance data. Among the procedures compared, in the simulations the directionally sensitive multivariate cumulative sum (MCUSUM) procedure was preferred, whereas in the real data the multiple univariate CUSUMs and the MCUSUM performed similarly. This article concludes with a brief discussion of the choice of performance metrics used herein versus the metrics more commonly used in the syndromic surveillance literature (sensitivity, specificity, and timeliness), as well as some recommendations for future research.

syndromic surveillance, biosurveillance, terrorism, disease, detection, statistical process control, CUSUM

Abbreviations: AORL, average overlapping run length; ARL, average run length; CDC, Centers of Disease Control and Prevention; CUSUM, cumulative sum; MCUSUM, multivariate cumulative sum; SPC, statistical process control.

## INTRODUCTION

Many existing multivariate statistical process control (SPC) procedures are directionally invariant, meaning that they are designed to detect changes in a mean vector in all directions. Examples of such procedures are Hotelling's $\chi^2$ (1), Crosier's multivariate cumulative sum (MCUSUM) (2), and more recently the nonparametric method of Qui and Hawkins (3). See Lowry and Montgomery (4) for a more detailed discussion. The lack of directional sensitivity can be a limitation of these methods, particularly when practitioners are interested in detecting changes in some directions more than others.

For example, the Centers for Disease Control and Prevention (CDC) as well as many state and local health departments around the United States have started to develop and field syndromic surveillance systems (5). Making use of existing health care or other data, often already in

electronic form, these surveillance systems are intended to give early warnings of bioterrorist attacks or other emerging health conditions. With such syndromic surveillance systems, it is important to flag increases in the relevant measures quickly because, in terms of signaling either a naturally occurring disease outbreak or a terrorist event, decreases are generally irrelevant. See Fricker and Rolka (6) or Stoto et al. (7) for a more detailed discussion. For a review of the use of control charts in the broader context of health care and public health surveillance, see Woodall (8).

Current syndromic surveillance systems run multiple simultaneous univariate SPC procedures, each focused on detecting an increase in a single dimension. Woodall and Ncube (9) first proposed the application of simultaneous univariate CUSUMs in a multivariate application. Multiple simultaneous univariate procedures have the advantages of ease of implementation and interpretation, but they can be less sensitive than multivariate methods to some types of changes. However, unless the signal thresholds of the multiple simultaneous procedures are properly set, they can suffer from a higher than desired combined false alarm rate.

Rogerson and Yamada (10) evaluated multiple univariate CUSUMs versus a directionally invariant multivariate CUSUM for monitoring changes in spatial patterns of disease. Recent work on directional multivariate procedures includes Testik and Runger (11) and Stoto et al. (7). Testik and Runger (11), building on the work of Follmann (12), Perlman (13), and Kudô (14), develop a number of multivariate "one-sided" procedures. In particular, Testik and Runger develop and compare multivariate procedures that look for 1) a shift in the mean vector corresponding to an increase in one or more components of the mean vector, 2) an increase in a pre-specified subset of the components of a mean vector while allowing the remaining components to either increase or decrease, and 3) a shift of the mean vector in the direction of a specific vector.

This work differs from that of Testik and Runger (11) in that the relevant alternative in syndromic surveillance is that at least one of the components of the mean vector has increased. Unlike in case 1) above, which, for a zero mean vector looks for vector shifts to the positive orthant, here the goal is to look for all mean vector shifts *except* those to the negative orthant. And, unlike in case 2), it is not possible to pre-specify the subset of the components to test for increases.

This work also differs from that of Stoto et al. (7) in that it demonstrates how to evaluate the performance of two directional multivariate procedures using metrics and terminology common to the SPC literature (e.g., run length and average run length (ARL)). As such, in recognition of the fact that there is a great deal that the syndromic surveillance and SPC communities can learn from each other, it is intended to help bridge the two literatures. In addition, a new metric, the average overlapping run length (AORL), is introduced and developed to compare the performance of various procedures on limited actual syndromic surveillance data.

## Terminology, notation, and assumptions

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 complaint at a specific hospital or within a particular geographic region under normal conditions. Let $F_1$ denote the *out-of-control* distribution; under the standard SPC paradigm, this would be a particular distribution representing a condition or state that is important to detect quickly. Within the syndromic surveillance problem, $F_1$ might represent an elevated mean daily count resulting from the release of a bioterrorism pathogen, for example.

The ARL is the performance metric used throughout the SPC literature to compare procedures. Roughly speaking, the *run length* is the length of time until a signal. In the syndromic surveillance case where daily counts are observed, the run length would be measured in units of days. Under the assumption that the process is always in-control—that is, all observations come from $F_0$—the *in-control ARL* is the mean time between false alarms. Denoted $ARL_0$, a larger in-control ARL is to be preferred, all things being equal.

Given the distribution shifts from $F_0$ to $F_1$ at some point in time, the *delay* (15) is the length of time from when the shift to $F_1$ occurred until a procedure signals. Referred to in the SPC literature as the *out-of-control ARL*, the notation $ARL_1$ is used to denote the expected delay for a given $F_1$ distribution. Hence $ARL_1$ is the average time it takes a procedure to signal from the time the shift occurred.

In the SPC literature, procedures are compared in terms of ARL, where the $ARL_0$ is first set equally for two procedures and then the procedure with the smallest $ARL_1$, for a particular out-of-control distribution, is deemed better. All procedures have strengths and weaknesses, each performing better on certain types of out-of-control distributions. Procedures that perform better across a range of $F_1$ distributions expected to be encountered in practice in a particular application are generally to be preferred for that application.

It is typical in SPC to assume that sequential observations are independent and identically distributed, according to either $F_0$ or $F_1$. In particular, the time series of observations resulting from the $F_0$ distribution is generally assumed to be stationary, meaning that the in-control observations do not exhibit any trends (periodic or otherwise) over time. In addition, the out-of-control condition most frequently evaluated is a jump change in the mean, meaning that the change from $F_0$ to $F_1$ results from the in-control mean $\mu_0$ jumping to some out-of-control mean $\mu_1 = \mu_0 + \delta$ for some $\delta$.

In industrial SPC applications, the assumption of independence can be reasonably well met by taking observations sufficiently far apart in time. In a similar way, the in-control observations can be reasonably assumed to come from a stationary distribution as some control is exercised over the process that produces the observations. These assumptions are more dubiously made in the syndromic surveillance problem, where it is desired to take observations as frequently as possible and where there is little or no control over the health conditions that give rise to the data which are frequently

observed or assumed to be nonstationary. Furthermore, out-of-control conditions characterized by simple jumps in the mean seem overly simplified for this problem.

Despite this, it is important to recognize that many of the procedures currently in use as syndromic surveillance systems, as well as those evaluated herein, were designed under these assumptions. This may be more or less of a problem in an actual syndromic surveillance application, depending on the characteristics of the specific data, and it is explored in more detail in the "Discussion" section.

### Organization

In this article, I present and then evaluate modifications of two existing multivariate methods—Hotelling's $\chi^2$ and a multivariate CUSUM by Crosier (2)—to make them directionally sensitive. The modifications are motivated by the univariate counterparts of the procedures and how those counterparts achieve directionality.

- The univariate counterpart to Hotelling's $\chi^2$ is the Shewhart procedure (16), where directionality is achieved by signaling only when an observation falls far enough out in one particular tail of the distribution. For the "modified Hotelling's $\chi^2$," directionality is achieved by signaling only when an observation falls within a particular region of the "tail" of the multivariate distribution.
- In the univariate CUSUM, directionality occurs naturally because the CUSUM statistic is bounded by zero in either the positive or negative direction. For the "modified MCUSUM," directionality is achieved by bounding each component of a CUSUM vector by zero in the desired direction.

I focus on the Shewhart and CUSUM procedures because these procedures are implemented in surveillance systems.

Following this, I compare and contrast the various procedures' performance via simulation and then demonstrate their performance on an actual syndromic surveillance–related data set. This article concludes with a discussion and some recommendations for future research.

## METHODS

### Univariate procedures

This section briefly describes two of the most common univariate procedures, Shewhart's procedure and CUSUM. See *Introduction to Statistical Quality Control* by Montgomery (17) and the references therein for additional detail.

*Shewhart's procedure.* Shewhart's procedure (16) is probably the simplest and best known of all SPC procedures. The basic idea is to evaluate sequentially one observation (or statistic) at a time, signaling when an observation that is rare under $F_0$ occurs. The most common form of the procedure, often known as the $\overline{X}$ *chart*, signals when the absolute value of an observed sample mean exceeds a pre-specified *threshold h*, often defined as the in-control mean value plus some number of standard deviations of the sample mean.

More sophisticated versions of the Shewhart procedure exist that look for increases in variation and other types of out-of-control conditions. These versions are not considered here to keep the evaluations simple.

The procedure can be made directionally sensitive by signaling for deviations in only one direction. For example, in syndromic surveillance, only deviations in the positive direction that would indicate a potential outbreak are assumed to be important to detect. Thus, for a univariate random variable $X$ with a known in-control density $f_0$, the threshold $h$ is chosen to satisfy

$$\int_{x=h}^{\infty} f_0(x)\mathrm{d}x = p,$$

where $p$ is the probability of a false signal and is fixed at some suitably small value. The algorithm proceeds by observing values of $X$; it stops and concludes $X \sim F_1$ at the first time $X > h$. For a given out-of-control distribution $F_1$ and its associated density $f_1$, assuming independence between observations, the ARLs can be directly calculated as $\mathrm{ARL}_0 = 1/p$ and

$$\mathrm{ARL}_1 = \left[\int_{x=h}^{\infty} f_1(x)\mathrm{d}x\right]^{-1}$$

*CUSUM procedure.* The CUSUM of Page (18) and Lorden (19) is a sequential hypothesis test for a change from a known in-control density $f_0$ to a known alternative density $f_1$. The procedure 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 procedure is usually started at $S_0 = 0$; it stops and concludes that $X \sim F_1$ at the first time when $S_i > h$, for some pre-specified threshold $h$ that achieves a desired $\mathrm{ARL}_0$.

If $F_0$ and $F_1$ are normal distributions with means $\mu$ and $\mu + \delta$, respectively, and equal variances, then equation (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 are only approximately normally distributed.

Note that, as the univariate CUSUM is bounded below at zero, it is capable of looking for departures only 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 equation (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. When directional sensitivity is desired, say to detect only positive shifts in the mean, it is only necessary to use equation (2).

### Directionally invariant multivariate procedures

This section describes two existing procedures that are the multivariate counterparts to the Shewhart procedure and CUSUM discussed in the previous section. The next section then describes modifications to these methods to make them directionally sensitive.

*Hotelling's $^2$.* Hotelling (1) introduced the $\chi^2$ procedure, sometimes also called the $T^2$ procedure. In this article, it is referred to as the $\chi^2$ procedure when the covariance matrix is known and the $T^2$ procedure when the covariance matrix is estimated. For multivariate observations $\mathbf{X}_i \in \mathbb{R}^d$, $i = 1, 2, \ldots$, the procedure computes

$$\chi_i^2 = (\mathbf{X}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_i - \boldsymbol{\mu}),$$

where $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix and $\boldsymbol{\mu}$ is the mean vector from the in-control distribution $F_0$. The procedure stops at the first time when $\chi_i > h$, for some pre-specified threshold $h$ that achieves a desired $\text{ARL}_0$.

As in Crosier (2), the $\chi^2$ procedure is called directionally invariant because its ARL depends on $\mu$ and $\Sigma$ only though the noncentrality parameter

$$\nu = [(\boldsymbol{\mu} - \mathbf{x})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{x})]^{1/2}.$$

This means the procedure can detect any shift of statistical distance $\nu$ equally well, regardless of direction. Like the univariate Shewhart procedure, because it uses only the most recent observation to decide when to stop, the $\chi^2$ can react quickly to large departures from the in-control distribution but will also be relatively insensitive to small shifts.

*Crosier's MCUSUM.* The abbreviation MCUSUM, for multivariate CUSUM, is used herein to refer to the procedure proposed by Crosier (2) that at each time $i$ considers the statistic

$$S_i = (S_{i-1} + X_i - \boldsymbol{\mu})(1 - k/C_i), \quad \text{if } C_i > k, \qquad (3)$$

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

$$Y_i = (S_i' \boldsymbol{\Sigma}^{-1} S_i)^{1/2}.$$

It concludes that $X \sim F_1$ at the first time when $Y_i > h$, for some pre-specified threshold $h$ that achieves a desired $\text{ARL}_0$.

In terms of choosing $\mathbf{k}$, Crosier (2) states, "In the univariate [CUSUM] case, the quantity $S_{i-1} + (X_i - \mu)$ is shrunk towards 0 by $k$ standard deviations. If this is to hold for the multivariate case, $\mathbf{k}$ must satisfy $\mathbf{k}'\boldsymbol{\Sigma}^{-1}\mathbf{k} = k^2$—that is, $\mathbf{k}$ must be of length $k$, where the length is defined by using the covariance matrix $\boldsymbol{\Sigma}$."

Crosier proposed a number of other multivariate CUSUM-like algorithms but generally preferred this form after extensive simulation comparisons. Pignatiello and Runger (20) proposed other multivariate CUSUM-like algorithms as well, but found that they performed similarly to Crosier's.

It is worth noting that Crosier derived his procedure in an *ad hoc* manner, not from theory, but found it to work well in simulation comparisons. Healy (21) derived a sequential likelihood ratio test to detect a shift in a mean vector of a multivariate normal distribution. However, although Healy's procedure is more effective (has shorter ARLs) when the shift is to the precise mean vector of $F_1$, it is less effective than Crosier's for detecting other types of shifts, including mean shifts that were close to but not precisely the specific mean vector of $F_1$.

### Directionally sensitive multivariate procedures

For the syndromic surveillance problem, it is desirable to focus the multivariate procedures in the direction of increases in incident rates because, as was previously discussed, the goal is to detect natural disease outbreaks or bioterrorism events. This section describes two directionally sensitive procedures that result from modifications to the procedures presented in the previous section.

*Modified Hotelling's $^2$.* A simple way to focus the $\chi^2$ procedure is to modify the stopping rule so that two conditions must be met: 1) $\chi_i > h$ and 2) $\mathbf{X}_i \in \mathcal{S}$, where $\mathcal{S}$ is a subspace of $\mathbb{R}^d$ that corresponds to a particular region of interest. The idea is that the modified procedure will signal only when an observation is far enough out in the "tail" of the distribution *and* it falls within some region that would be more likely if, say, one or more components of the mean vector increased.

To formalize this idea, let

$$\mathcal{G}(\alpha) = \left\{ x_1, x_2, \ldots, x_d : \int_{\mathcal{G}} f_0(\mathrm{x}) \mathrm{dx} = 1 - \alpha \right\},$$

where if $f_0$ is a multivariate normal density, then $\mathcal{G}(\alpha)$ is an ellipse centered at the mean containing probability $1 - \alpha$ with $f_0(\mathbf{x})$ constant on the boundary of the ellipse. Let

$$\mathcal{S}(\beta) = \left\{ x_1, x_2, \ldots, x_d : \int_{x_1=s_1}^{\infty} \int_{x_2=s_2}^{\infty} \cdots \int_{x_d=s_d}^{\infty} f_0(\mathrm{x}) \mathrm{dx} = 1 - \beta \right\}, \qquad (4)$$

where, for a suitably small $\beta$, $\mathcal{S}$ is the "upper right" quadrant of $\mathbb{R}^d$ that contains most of the probability of $F_0$. Then, for a particular $F_0$, choose $\mathcal{S}$ so that $\beta$ is small, say $\beta \approx 0.01$, and then choose $\mathcal{G}$ so that

$$\int_{\mathcal{S}} f_0(\mathbf{x}) \mathrm{dx} - \int_{\mathcal{G} \cap \mathcal{S}} f_0(\mathbf{x}) \mathrm{dx} = p$$

for some small $p$ chosen to achieve a particular $\text{ARL}_0$. Note that if $\beta = 0$, then this reduces to the directionally invariant procedure. Also note that the definition of $\mathcal{S}$ above focuses the procedure in the general direction of componentwise increases in the mean, but equation (4) could certainly be modified to focus in other directions.

To illustrate in $\mathbb{R}^2$, consider an in-control distribution following a bivariate normal distribution with some positive correlation, so that the probability contours for the density of $F_0$ are concentric ellipses with their main axis along 45-degree line in the plane. As shown in figure 1a, you can then think about $\mathcal{S}$ as the upper right quadrant (which continues out to positive infinity in the positive $x_1$ and $x_2$

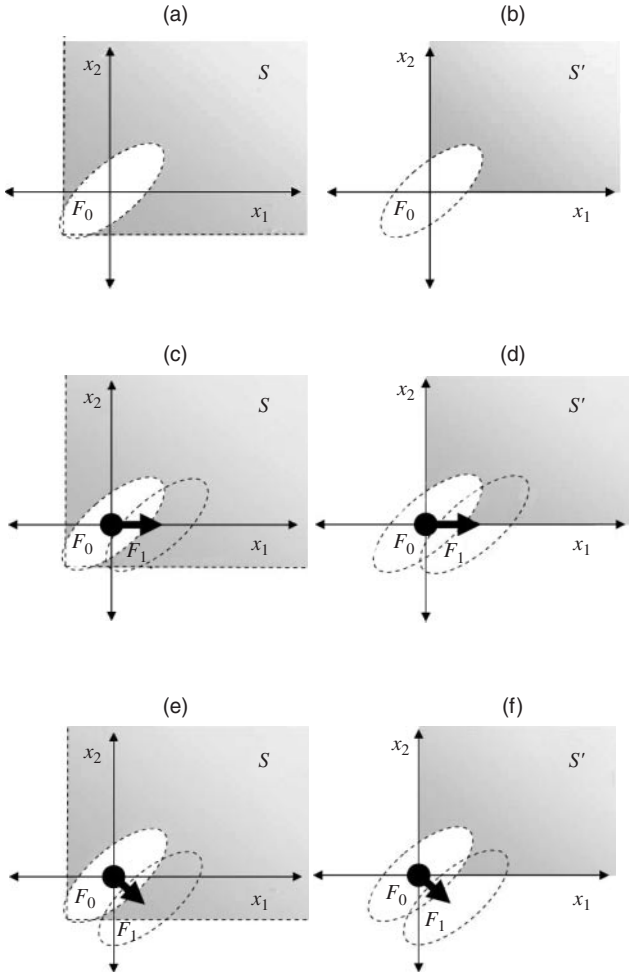


**FIGURE 1.** An illustration of $\mathcal{S}$ in (a) versus $\mathcal{S}'$ in (b), where $\mathcal{S}'$ is the positive orthant. Also, illustrative comparisons between $\mathcal{S}$ and $\mathcal{S}'$ for two types of mean vector shifts: (c) versus (d) and (e) versus (f). Except for small-to-moderate shifts in the mean, the $ARL_1$ under $\mathcal{S}$ will be smaller than the $ARL_1$ under $\mathcal{S}'$.

directions) that almost encompasses the $100 \times (1 - \alpha)$ percent probability ellipse. The stopping rule, then, is that an observation must fall in the shaded region in figure 1a, which, for $p = 0.01$, contains 1 percent of the $F_0$ probability.

There are two reasons to use this region for $\mathcal{S}$ as compared with, say, $\mathcal{S}'$ based on the positive orthant as illustrated in figure 1b. First, if $F_1$ represents a shift in the mean vector in any direction corresponding to an increase in one or more of the individual means, then a procedure based on $\mathcal{S}$ will generally have a higher probability of signaling than one based on $\mathcal{S}'$. Hence, the $ARL_1$s resulting from the use of $\mathcal{S}$ will generally be smaller than the associated $ARL_1$s for $\mathcal{S}'$.

For example, consider a case where only the $x_1$ component of the mean vector increases, as indicated in figure 1c and d by the dark arrow. In such a case, the probability associated with the portion of $F_1$ that falls in the shaded area is higher for $\mathcal{S}$ than for $\mathcal{S}'$ for all except small-to-moderate increases (in which case the probability associated with the portion of $F_1$ that falls in the shaded area using $\mathcal{S}'$ is slightly greater). Furthermore, even for the most extreme shifts, the $ARL_1$ using $\mathcal{S}'$ is bounded below by 2 whereas the $ARL_1$ for $\mathcal{S}$ can get close to 1 for small $\beta$s.

Second, in syndromic surveillance, it could be important to signal when one component of the mean vector increases even as perhaps one or more of the other components decrease. As illustrated in figure 1e and f, again the portion of $F_1$ that falls in the shaded area associated with $\mathcal{S}$ will generally be larger than that which falls in the shaded area resulting from $\mathcal{S}'$, and thus $\mathcal{S}$ will be more effective in detecting this type of mean change than $\mathcal{S}'$.

*Modified MCUSUM.* Unlike some other multivariate CUSUMs, Crosier's MCUSUM formulation is easy to modify to look only for positive increases. As was described in the Introduction, the motivation for this modification is the univariate CUSUM equation (2), where directionality is achieved because the CUSUM statistic is bounded below by zero. In the modified MCUSUM, directionality is similarly achieved by bounding each component of the MCUSUM vector by zero.

In particular, for detecting positive increases, such as in the syndromic surveillance problem, when $C_i > k$ limit $\mathbf{S}_i$ to be positive in each dimension by replacing equation (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 $j = 1, 2, \ldots, d$.

## RESULTS

### Performance comparisons via simulation

In this section, the various procedures are evaluated by simulation using independent observations generated according to either an in-control distribution $F_0$ or an out-of-control distribution $F_1$. (The next section examines the performance of the procedures on real syndromic surveillance data that are autocorrelated.) Performance is assessed by ARL, first determining the thresholds ($h$) to achieve equal $ARL_0$s and then comparing the ARL performance under numerous out-of-control distributions resulting from various shifts in the mean vector at time 0.

For the simulations, $F_0$ is a six-dimensional multivariate normal with a zero mean vector, $\mu_0 = \{0, 0, 0, 0, 0, 0\}$, and a covariance matrix $\boldsymbol{\Sigma}$ consisting of unit variances on the diagonal and constant covariance $\rho$ on the off-diagonals. (In practice, assuming stationarity and that sufficient historical data are available, this can be achieved via standardization.) The $F_1$s have the same covariance structure but with the mean vector shifted by some distance $d$,

$$d = |\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1| = \left[\sum_{i=1}^{6}\left(\mu_1(i)\right)^2\right]^{1/2},$$

where the shift occurs in some number of dimensions $n$, $1 \leq n \leq 6$. For those dimensions with a shift, the shifts were made equally: $\mu_1(1) = \cdots = \mu_1(n) = \sqrt{d^2}/n$.

The simulations were conducted in Mathematica 5.0 (22), where the random observations were generated using the *MultinormalDistribution* function. For the univariate CUSUMs, I set $k = 0.5$ in equation (2). This is equivalent to saying that it is important to be able quickly to detect a one standard deviation increase in the mean. For the MCUSUM and modified MCUSUM, I fixed $\mathbf{k} = \{0.2, 0.2, 0.2, 0.2, 0.2, 0.2\}$ in equation (3) so that for $\rho = 0$ the univariate and multivariate procedures' $k$s were equal: $k = \{\mathbf{k}'\boldsymbol{\Sigma}^{-1}\mathbf{k}\} \approx 0.5$.

The in-control ARLs were set to 100 by empirically determining the threshold $h$ for each procedure. For the multivariate procedures, this involved determining a single threshold for each value of $\rho$ (except for Hotelling's $\chi^2$ procedure, for which one threshold applies to all values of $\rho$). For the simultaneous univariate procedures, which require a separate threshold for each individual procedure, 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 dimensions and so that the resulting expected time to false alarm for the combined set of univariate procedures was equal to the expected time to false alarm of the multivariate procedure.

In general, it is quite simple to estimate the ARLs empirically via simulation. For a particular $F_0$, choose an $h$ and run the algorithm $r$ times, recording for each run the time $t$ when the first $X > h$ (where each $X$ is a random draw from $F_0$, of course). Estimate the in-control ARL as $\widehat{\text{ARL}}_0 = \sum_{i=1}^{r} t_i/r$, adjusting $h$ and re-running as necessary to achieve the desired in-control ARL, where $r$ is set large enough to make the standard error of $\widehat{\text{ARL}}_0$ acceptably small. Having established the threshold $h$ for that $F_0$ with sufficient precision, then for each $F_1$ of interest re-run the algorithm $s$ times (where $s$ is often smaller than $r$), drawing the $X$s from $F_1$ starting at time 1. As before, take the average of the $t$s to estimate $\text{ARL}_1$.

Although I determined the necessary thresholds empirically via simulation, note that precise solutions and approximations have been derived for selecting thresholds to achieve desired ARLs for some procedures under some conditions. The easiest is the univariate Shewhart procedure, where for a given threshold, assuming independence between observations, the precise ARL can be calculated as described in the "Shewhart's procedure" section. For the CUSUM, computationally simple ARL approximations (assuming independent normal observations) include those of Reynolds (23) and Siegmund (24). Approximations have also been derived for the univariate exponentially weighted moving average (EWMA) procedure, which we do not consider here, but which are being evaluated elsewhere in the context of public health surveillance (e.g., Joner et al. (25)). Less work has been conducted for multivariate procedures, though Runger and Prabhu (26) have derived approximations for the multivariate EWMA, and Fricker (27) has derived an approximation for the nonparametric repeated two-sample rank procedure.

That all said, all of these methods require at least independence between observations, an assumption that is unlikely to apply to actual syndromic surveillance data. Furthermore, with the speed of today's computers, it is a simple matter to write a program to automate the empirical estimation, as described in the previous paragraph, that runs in a few seconds or minutes. Hence, my empirical approach to determining thresholds, an approach that is just as applicable to autocorrelated data as to data that are independent.

In the simulations to follow, the modified multivariate procedures' performances are first compared with those of their counterpart unmodified procedures. This quantifies the directional sensitivity and effectiveness of the modified procedures. This is followed by comparisons of the modified multivariate procedures with the application of simultaneous univariate procedures. The simultaneous univariate procedures are implemented to be directionally sensitive in the same direction as the modified multivariate procedures. Finally, the best procedures from the previous comparisons are compared in an effort to determine whether a single procedure is generally best.

*Modified procedures versus original procedures.* Figure 2 shows the improved performance of the modified $\chi^2$ procedure and the modified MCUSUM for almost all types of

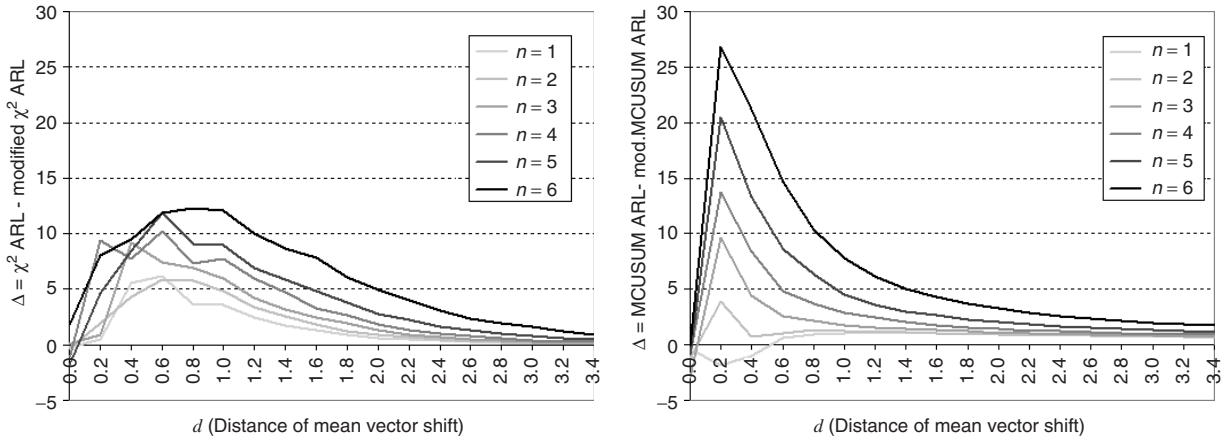


**FIGURE 2.** Performance comparison of the $\chi^2$ and MCUSUM procedures compared with that of their modified counterpart procedures for $\rho = 0.3$. A positive value of $\Delta$ indicates that the ARL for the modified procedure is shorter than the ARL of its unmodified counterpart.

mean vector shifts where, as just described, the component-wise shifts are in the positive direction. This is not surprising given that the modified procedures were designed to look for positive mean shifts.

In figure 2, the various lines correspond to the number of dimensions in $\mu_1$ that shifted, and the horizontal axis is the distance of the mean shift. For example, the $n = 1$ line shows the results for $\mu_1 = \{d, 0, 0, 0, 0, 0\}$, where the ARL was evaluated at $d = 0.0, 0.2, 0.4, \ldots, 3.4$. The $n = 2$ line shows the results for $\boldsymbol{\mu}_1 = \{\sqrt{d^2/2}, \sqrt{d^2/2}, 0, 0, 0, 0\}$. And so on.

The vertical axis is the difference $\Delta$ between the ARL for the original procedure and the modified procedure for a given mean vector shift. Positive values indicate the modified procedure had a smaller ARL, so that for a particular out-of-control condition the modified procedure had a shorter time to signal. In the figure, $\Delta = 0$ at $d = 0.0$ indicates that the in-control ARLs were set equally for each procedure before comparing the expected time to signal for various $\boldsymbol{\mu}_1$s (within the bounds of experimental error, where a sufficient number of simulation runs were conducted to achieve a standard error of $\Delta$ of approximately 2 percent of the estimated in-control ARLs).

As previously mentioned, the modified procedures generally outperform the original procedures in detecting positive shifts. Figure 2 shows this for the case of $\rho = 0.3$. Though not shown, the results for other values of $\rho$, from $\rho = 0$ to $\rho = 0.9$, are very similar.

In particular, the modified $\chi^2$ outperforms Hotelling's $\chi^2$ for all combinations of $1 \leq n \leq 6$, $0.0 < d \leq 3.4$, and $0 \leq \rho \leq 0.9$. The modified MCUSUM outperforms Crosier's MCUSUM except for larger values of $\rho$ with small $d$ and small $n$. For example, in figure 2, Crosier's MCUSUM slightly outperforms the modified MCUSUM for $n = 1$ with $0 < d < 0.6$ or so. For $\rho = 0.6$, Crosier's MCUSUM outperforms the modified MCUSUM on the order of $-6 < \Delta < 0$ or so for $n = 1, 2$ with $0 < d < 0.6$. And, for $\rho = 0.9$, Crosier's MCUSUM outperforms the modified MCUSUM on the order of $-9 < \Delta < 0$ or so for $n = 1, \ldots, 5$ with $0 < d < 1$.

For this work, moderate values of $\rho$ are of interest, as the syndromic surveillance data in the section "An application to syndromic surveillance" exhibit only moderate correlations, roughly on the order of $0 < r < 0.5$. In addition, the signals of interest—that is, shifts in the mean vector—are those consisting of small changes in multiple dimensions. (Indeed, if the expected shift is in only a small number of dimensions and/or the covariance $\rho$ is large, then it is likely that univariate methods would be more appropriate anyway.) With this in mind, what is most notable in figure 2 is that as $n$ increases, the modified procedures do considerably better than their counterparts, particularly for moderate $d$s.

*Modified procedures versus univariate procedures.* Given that the modified $\chi^2$ performs better than the original $\chi^2$ for this problem, figure 3 focuses on comparing the performance of the modified $\chi^2$ to six one-sided Shewhart procedures operating simultaneously. The left-side graph of figure 3, constructed just like figure 2, shows that six simultaneous univariate Shewharts are more effective (have shorter ARLs) than the modified $\chi^2$ for $\rho = 0.3$. At best, for large shifts, the ARL of the modified $\chi^2$ approaches the performance of the multiple univariate Shewharts, and for small-to-moderate shifts the multiple univariate Shewharts are clearly better.

The graph on the right side of figure 3 shows the performance comparison for $n = 3$ and for various values of $\rho$ (0.0, 0.3, 0.6, and 0.9). Here we see that the better procedure depends on $\rho$, where the modified $\chi^2$ is better for values of $\rho$ near 0.0 or 0.9 whereas the simultaneous univariate Shewharts are better for moderate values of $\rho$. Interestingly, the modified $\chi^2$ significantly outperforms the simultaneous univariate Shewharts when there is no correlation ($\rho = 0$) and when the shift is only in three of the six dimensions.

The results for the modified MCUSUM versus simultaneous univariate CUSUMs are presented in figure 4. These results differ from those for the Shewhart-type procedures in figure 3 in that the modified MCUSUM is generally better than the simultaneous univariate CUSUMs

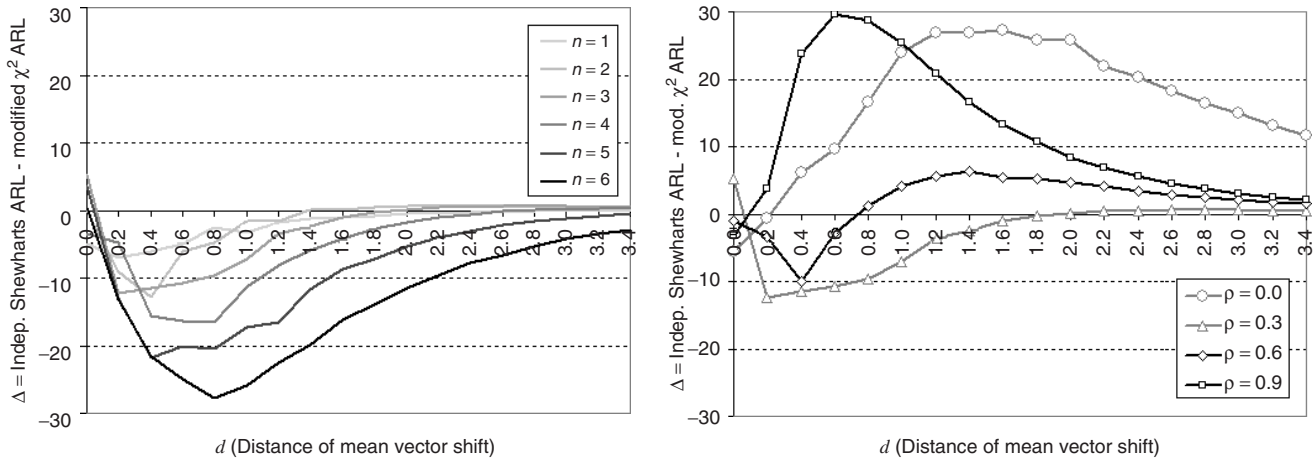


**FIGURE 3.** Performance comparison of the modified $\chi^2$ procedure and multiple simultaneous univariate Shewhart procedures. The figure on the left shows that for $\rho = 0.3$ the multiple simultaneous Shewhart procedures give smaller ARLs for all $n$. However, the figure on the right with $n = 3$ shows that either procedure can be significantly better than the other depending on the value of $\rho$.

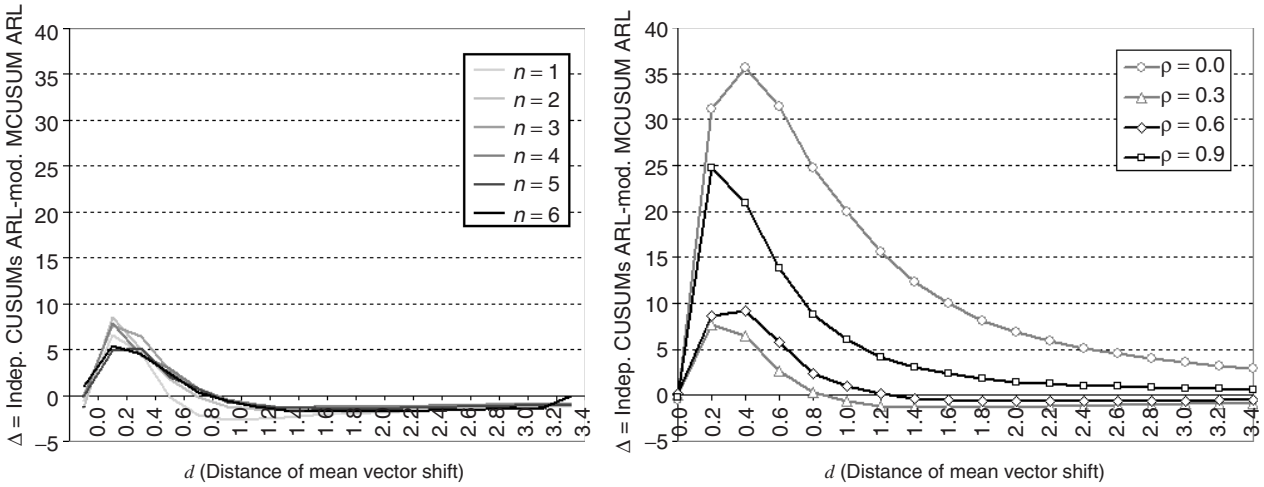


**FIGURE 4.** Performance comparison of the modified MCUSUM procedure and multiple simultaneous univariate CUSUM procedures. The figure on the left shows that for $\rho = 0.3$ the MCUSUM does better for small values of $d$ and marginally worse for large $d$. However, unlike the modified $\chi^2$ in figure 3, the figure on the right for the modified MCUSUM with $n = 3$ is generally better than simultaneous univariate CUSUMs for all values of $\rho$.

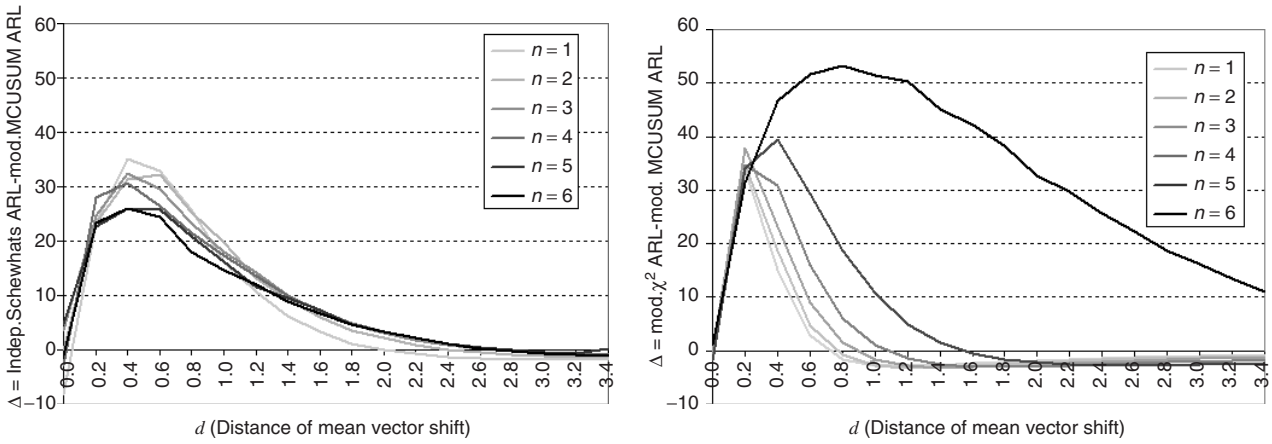


**FIGURE 5.** Performance comparison of the modified MCUSUM procedure to multiple simultaneous univariate Shewhart procedures (left graph) for $\rho = 0.3$ and to the modified $\chi^2$ procedure (right graph) for $\rho = 0.9$. In both cases, the modified MCUSUM procedure performs generally better than the preferred Shewhart-type procedure.

regardless of the value of $\rho$. In particular, in the left graph of figure 4 the modified MCUSUM performance when $\rho = 0.3$ is somewhat better for small shifts (roughly $0.0 > d > 0.6$ or so), and slightly worse than multiple univariate CUSUMs for moderate-to-large shifts. Yet, in the figure at the right we see that the modified MCUSUM is better for small shifts for all values of $\rho$ and performs only slightly worse for moderate values of $\rho$ combined with moderate-to-large values of $d$.

*Modified MCUSUM versus best other procedures.* What the previous simulations have shown is that the modified MCUSUM is generally better than the simultaneous univariate CUSUMs. However, whether the modified $\chi^2$ is better than simultaneous univariate Shewhart procedures depends on $\rho$. So, here the modified MCUSUM is compared with the better of either the simultaneous univariate Shewhart procedures or the modified $\chi^2$ under the conditions that favor each: the simultaneous univariate Shewhart procedures for moderate covariance ($\rho = 0.3$) and the modified $\chi^2$ for high covariance ($\rho = 0.9$). The results are shown in figure 5. In both comparisons, the modified MCUSUM procedure's performance is better. The obvious 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 positive covariance.

Now, all the figures up to this point have shown differences in ARL performance between two procedures. Figure 6 shows the ARLs for the modified MCUSUM for $n = 3$. Results for $n = 1, 2, 4, 5, 6$ were similar; although the individual $\rho$ curves moved around, they largely stayed within the same band/region. For example, for $n = 1$, the lowest ARLs were achieved for $\rho = 0.9$ whereas for $n = 6$ the lowest ARLs were achieved with $\rho = 0.0$.

## An application to syndromic surveillance

In this section, to demonstrate how the procedures perform under real-world conditions, the CUSUM-based

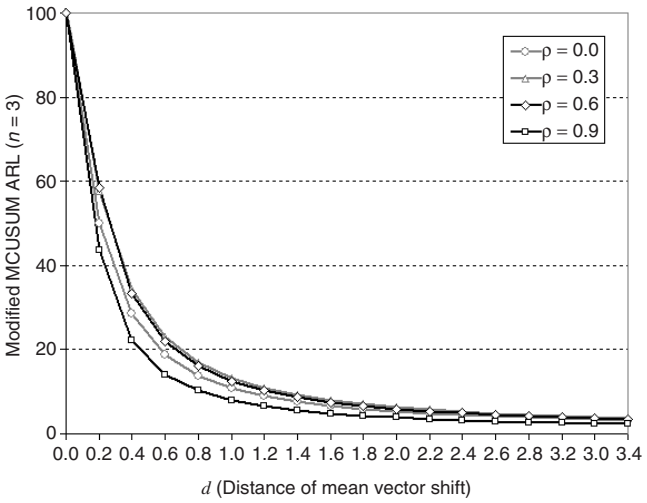


**FIGURE 6.** Modified MCUSUM ARLs for $n = 3$.

procedures—which performed better in the simulations—are applied to data from five hospitals located in a large metropolitan area. The data consist of respiratory "chief complaint" counts by hospital for two-and-a-half years from October 1, 2001, to March 31, 2004.

Chief complaints are broad categories—for example, respiratory, gastrointestinal, unspecified infection, neurological—into which patients are grouped *before diagnosis*. They are intended to capture the primary symptom or reason a patient sought care. Applying the procedures to these data, which capture naturally occurring incident rates and variation within hospitals as well as the covariation between hospitals, provides some insight into real-world performance.

This analysis focuses on respiratory chief complaint data. Respiratory chief complaints tend to include those patients who come to emergency rooms with the flu and flu-like symptoms. Given that such symptoms could also be leading indicators of certain bioterrorism agents, monitoring respiratory chief complaints is thought to be useful in syndromic surveillance for bioterrorism (28).

Figure 7 shows the smoothed respiratory chief complaint counts by hospital. Each point is a 4-week moving average using the data for 2 weeks before and 2 weeks after the date. (Note that this centered moving average was used only for illustration purposes in figure 7. The section on "Implementation" describes how the actual data were used in the application of the procedures.) A number of features of the data are clear from the figure, including the following points:

- The hospital moving averages do not exhibit an increasing or decreasing trend, indicating the long-term incidence rate for respiratory chief complaints is constant.
- Yet, there are visible "events" in the data that persist for periods of time. For example, there are peaks across most or all of the hospitals in January–February 2002, March–June 2003, and December 2003–January 2004 that likely correspond to flu outbreaks.
- These events are consistent with the CDC's aggregate data on "percentage of visits for influenza-line illness reported by sentinel physicians" (29) for the South Atlantic region of the United States (where the city is located):
  - — The 2001–2002 flu season was characterized as "mild to moderate in the United States." The percentage of visits in the South Atlantic region peaked in February–March 2002. However "influenza activity as reported by sentinel physicians in the Mid-Atlantic and South Atlantic regions peaked during mid-to-late January."
  - — The 2002–2003 flu season was characterized as "mild," with the percentage of visits in the South Atlantic region peaking in February–March 2003. "Sporadic activity" was also reported in April and May 2003.
  - — The 2003–2004 flu season "began earlier than most seasons and was moderately severe." The percentage of visits in the South Atlantic region peaked in December 2003.
- The hospital counts are positively correlated. Indeed, using the first 6 months of the data, the correlations between all pairs of hospitals lie in the interval $0.0 \leq r \leq 0.49$.

In addition, there are significant differences in mean counts between hospitals, indicating that some hospitals either serve larger populations or serve populations with higher respiratory illness rates (or both), as well as significant variation in the raw counts around the smoothed mean.

*Implementation.* To implement the procedures, I first divided the data up into a "historical" set of data, consisting of the first 6 months (10/1/01-3/31/02), and a "future" set of data—the remaining two years (4/1/02-3/31/04). As one would do in practice, the idea was to use the "historical" data to estimate various quantities necessary for the procedures and then to calculate each procedure's performance using the "future" data.

In particular, using the first 6 months of data, I 1) determined that a square root transformation would make the data approximately normally distributed, and 2) estimated means and standard deviations for the transformed respiratory counts for each hospital, as well as the variance-covariance matrix for the joint distribution. I then used the sample means and standard deviations to standardize the (square root–transformed) data for each hospital. Given the differences in the raw rates, standardization is important to ensure that equal weight is given to each hospital.

For the independent CUSUMs, I set $k = 1$ and used $h = 2.125$ (for each individual procedure) to achieve a combined estimated in-control ARL of approximately 100. In a similar way, for the modified MCUSUM, I set $\mathbf{k} = \{0.55, 0.55, 0.55, 0.55, 0.55\}$, so that $k = \{\mathbf{k}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{k}\}^{1/2} = 1.0$, and used $h = 4.3$ to achieve an estimated in-control ARL of approximately 100.

*Detecting flu outbreaks.* Let us begin by illustrating how well the procedures perform on the real data. In particular, let us compare and contrast the modified MCUSUM and individual CUSUM signals and consider whether those

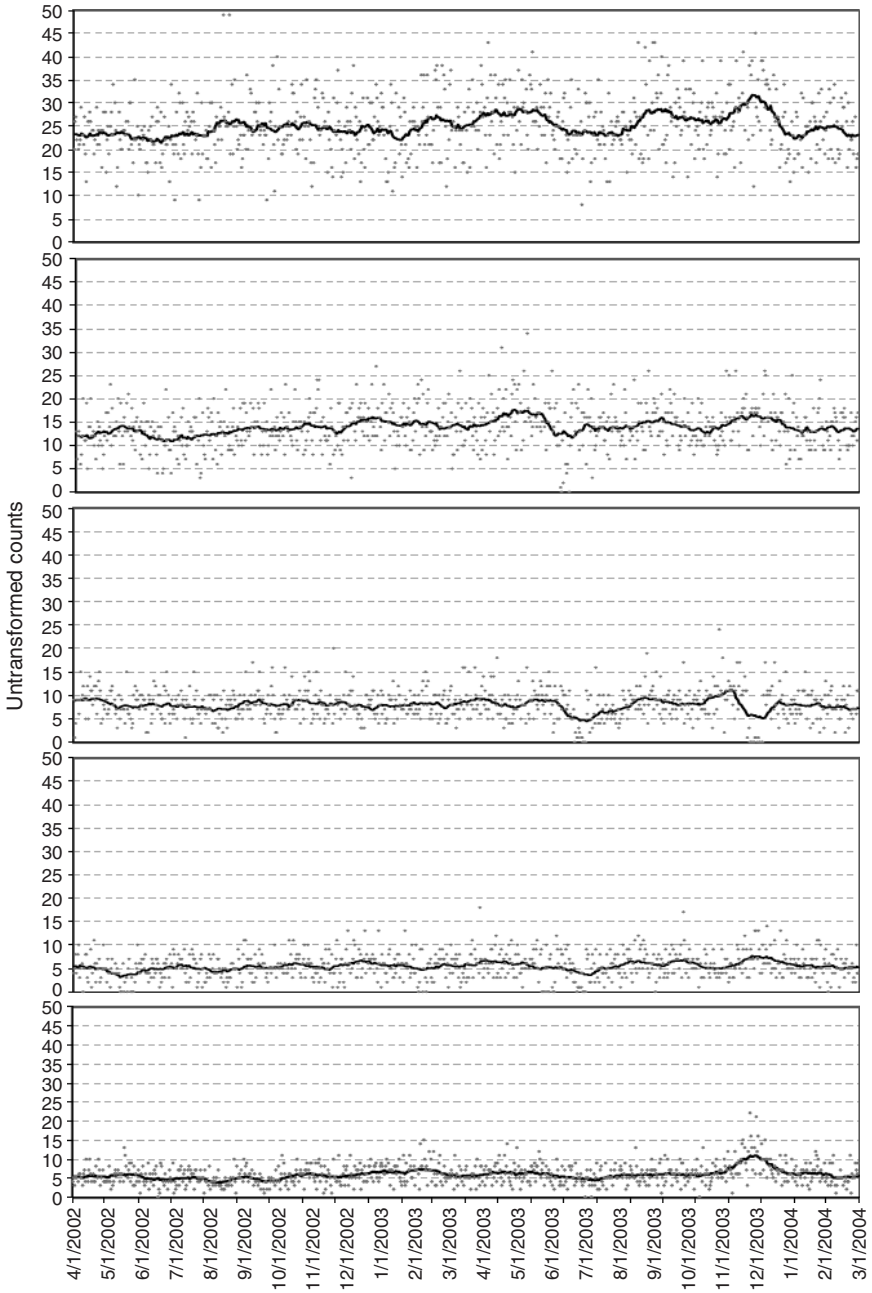


**FIGURE 7.** The data consisting of "chief complaint" respiratory counts by hospital with a smoothed mean line superimposed. The smoothed mean was calculated using a 4-week moving average from 2 weeks before to 2 weeks after each day.

signals are consistent with the CDC's information about influenza-line illness as reported by the sentinel physicians.

Figure 8 displays the signal times for the various procedures when they are run on the respiratory data. The figure shows the smoothed means (of figure 7) and first signal times overlaid. ("First signal time" means that repeated signals within 30 days of the first signal are suppressed for plot clarity.) The signal times for the modified MCUSUM are indicated by the dark vertical lines with the specific dates at the top. The signal times for the individual CUSUMs are indicated by the diamonds plotted on the relevant smoothed mean. For example, figure 8 shows that the first signal for the modified MCUSUM occurred on November 15, 2002, and that one of the individual CUSUMs also signaled on the same day.

What this figure generally shows is that the modified MCUSUM and the simultaneous individual CUSUMs performed very similarly:

- As discussed, on November 15, 2002, both schemes signaled on the same day.
- On May 16, 2003, the modified MCUSUM signaled, 1 day after an individual CUSUM signaled (May 15) and after which one CUSUM signaled on May 27.

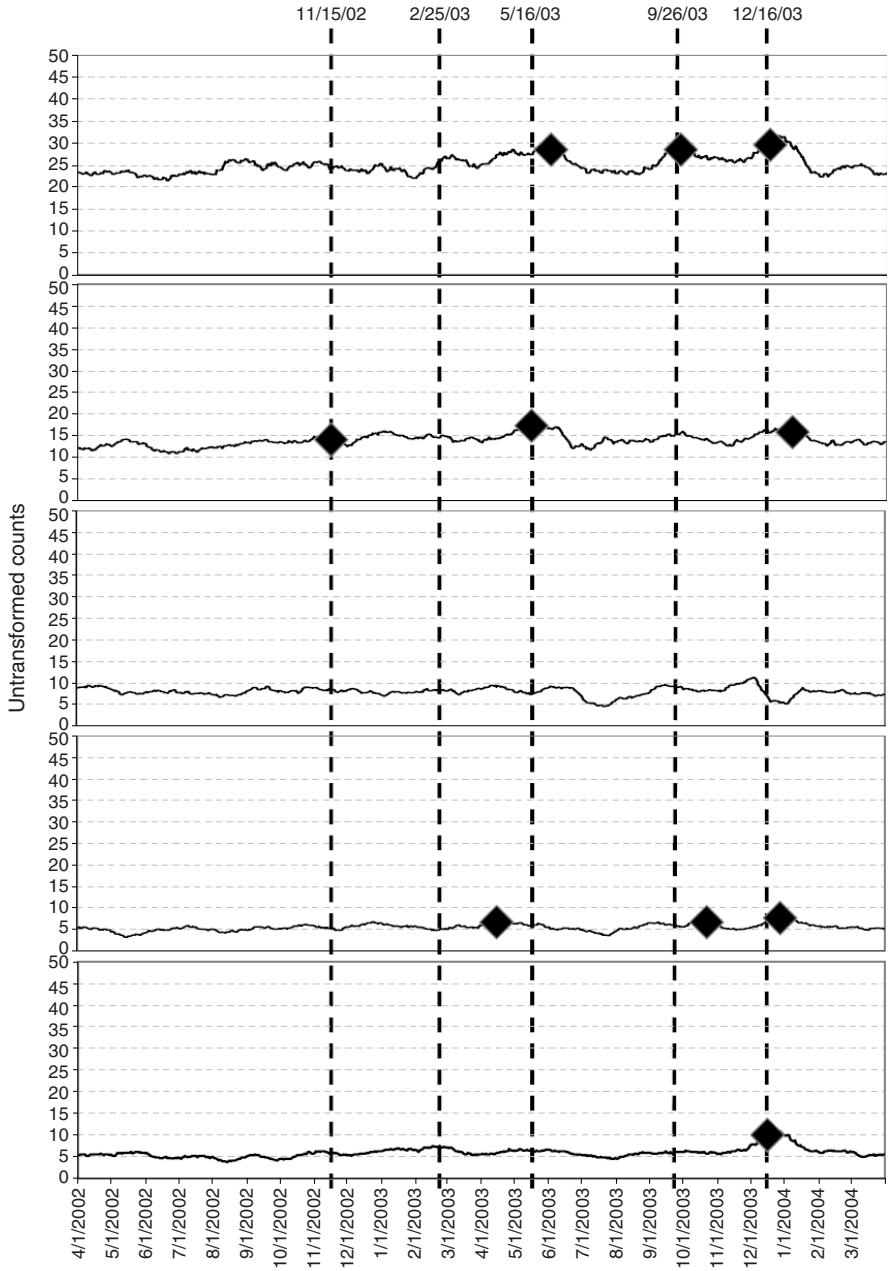


**FIGURE 8.** This plot shows when the modified MCUSUM and individual CUSUMs first signaled when run on the future data. The vertical lines are the signal times for the MCUSUM. The diamonds indicate the signal times for each individual CUSUM. "First signaled" means that repeated signals within 30 days of the first signal are suppressed for plot clarity.

- On September 26, 2003, the modified MCUSUM signaled, after which two of the individual CUSUMs signaled, the earliest on September 27, a difference of just 1 day.
- And, on December 16, 2003, the modified MCUSUM signaled, after which four of the individual CUSUMs signaled, the earliest of which was also on December 16.

There are only two times when the modified MCUSUM and the simultaneous individual CUSUM signals diverge:

- On February 25, 2002, the modified MCUSUM signaled, whereas none of the individual CUSUMs did so. Although there was a visible increase in the mean of the first series, it was not enough to signal the univariate CUSUM for that period of time. We can speculate that there also seem to be very slight increases in series 2 and 5 (reading figure 8 from top to bottom), so perhaps the combination of the three series was enough to cause the multivariate method to signal.
- On April 14, 2003, one of the individual CUSUMs (series 4) signaled, whereas the MCUSUM did not. This signal does seem to correspond to a slight increase in the mean for the particular series, but it was not enough to cause the modified MCUSUM to signal.

Whether either one or both of these signals were true or false is impossible to tell from just this data. Certainly, one false signal each for the length of time monitored is well within the false alarm rate.

How do these signals compare with the CDC's sentinel physician information? The MCUSUM signals on February 25, 2003, and May 16, 2003, along with the associated individual CUSUM signals, are consistent with the CDC's characterization of the 2002–2003 flu season, where the CDC reported a mild season that peaked in February–March 2003 with "sporadic activity" in April and May 2003. In a similar way, the MCUSUM signals on September 26, 2003, and December 16, 2003, along with the associated individual CUSUM signals, are consistent with the CDC's report of the 2003–2004 flu season, which the CDC said "began earlier than mostseasons and was moderately severe" and which peaked in December 2003.

Only the November 15, 2002, MCUSUM and CUSUM signals do not seem to correspond to any events described in the CDC's summary of influenza-like illnesses. That may be because the November 15 signal is false or it may be because a localized outbreak of flu was not sufficiently large to be detected/reported by the sentinel physician system. Although there is no way to reach a definitive conclusion from just this data, a visual inspection of the smoothed means in figure 8 around November 15 does not show any unusual increases in the respiratory chief complaint means, which would lead one to suspect this was a false alarm.

*Detecting bioterrorism.* Running the procedures on real data gives some idea of how the methods might perform in detecting natural disease outbreaks under real-world conditions. However, a bioterrorism event will occur as a man-made outbreak among, or perhaps on top of, natural disease outbreaks. To illustrate how well these procedures might perform under such conditions, similar to what has been done elsewhere in the syndromic surveillance literature (see, for example, Goldenberg et al. (30)), on the real (standardized, square root–transformed) data I superimposed an artificial increase in the mean to get some insight into how well these methods might detect such a bioterrorism event.

The metric I used to compare performance in this scenario is the AORL. Unlike the ARL, which averages independent run lengths, for the real data the run lengths that result from running each procedure starting at each day in the data are averaged to calculate the AORL. That is, starting on April 1, 2002, I ran each procedure until it signaled and recorded the resulting run length. I repeated this process starting on the next day, calculating the run length for April 2, 2002, and then cycled through all possible start dates in the data (where if during a run the last date (March 31, 2004) was reached, the procedure continued by cycling back through the data starting with April 1, 2002). The resulting 731 run lengths were then averaged to calculate the AORL (either for a particular outbreak condition or under the condition of no outbreak).

The rationale for using this metric, instead of the ARL, follows from the fact that a bioterrorism incident might occur at any time. That is, a bioterrorism attack could occur during a period of easy detection, say when natural disease incident rates are low and so the bioterrorism signal would more easily stand out, or perhaps during a period of difficult detection, say during a severe flu outbreak. Hence, arbitrarily choosing only 1 day or a small number of days on which to insert the outbreak might not reflect how the procedures would have performed had the incident occurred at some other time in the data.

Instead, it is relevant to measure how the procedures perform allowing for an outbreak to occur on any day in the data. The AORL is such a metric, where the average is calculated using all the run lengths resulting from allowing the outbreak to occur on each day in the data set. Not only does this metric provide an overall measure of performance, but in so doing it makes the most use of limited syndromic surveillance data sets. As it turns out, and as shown in the appendix, under some relatively mild conditions the AORL is also approximately proportional to the average squared run length.

I conducted two types of comparisons. In the first the mean linearly increases over time, and in the second it experiences a one-time jump change (and then remains constant). The results are shown in figure 9, where on the left side the mean vector experiences a one-time jump of distance $\Delta$ and on the right side the mean vector linearly increases by $\delta$ at each time period. In the top row, the change occurs for only one hospital; in the middle row, it occurs for three hospitals; and in the bottom row, it occurs for all five hospitals.

That is, denoting the actual (transformed and standardized) counts at time $i$ as $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4}, x_{i,5}\}$, then for the plot in the top left the out-of-control data, starting at time $t$, are

$$\tilde{\mathbf{x}}_{t+j} = \{x_{t+j,1} + \Delta, x_{t+j,2}, x_{t+j,3}, x_{t+j,4}, x_{tf+j,5}\} \quad \text{for } j = 0, 1, 2, \ldots.$$

For the plot in the middle left,

$$\tilde{\mathbf{x}}_{t+j} = \{x_{t+j,1} + \Delta/\sqrt{3}, x_{t+j,2} + \Delta/\sqrt{3}, x_{t+j,3} + \Delta/\sqrt{3}, x_{t+j,4}, x_{t+j,5}\} \quad \text{for } j = 0, 1, 2, \ldots.$$

For the plot in the top right, the out-of-control state is

$$\tilde{\mathbf{x}}_{t+j} = \{x_{t+j,1} + (j+1)\times\delta, x_{t+j,2}, x_{t+j,3}, x_{t+j,4}, x_{tf+j,5}\} \quad \text{for } j = 0, 1, 2, \ldots.$$

For the plot in the middle right,

$$\tilde{\mathbf{x}}_{t+j} = \{x_{t+j,1} + (j+1)\times\delta/\sqrt{3}, x_{t+j,2} + (j+1) \times \delta/\sqrt{3}, x_{t+j,3} + (j+1)\times\delta/\sqrt{3}, x_{t+j,4}, x_{t+j,5}\} \quad \text{for } j = 0, 1, 2, \ldots.$$

And so on.

As shown in figure 9, the results using real data are consistent with the simulations in the section "Performance

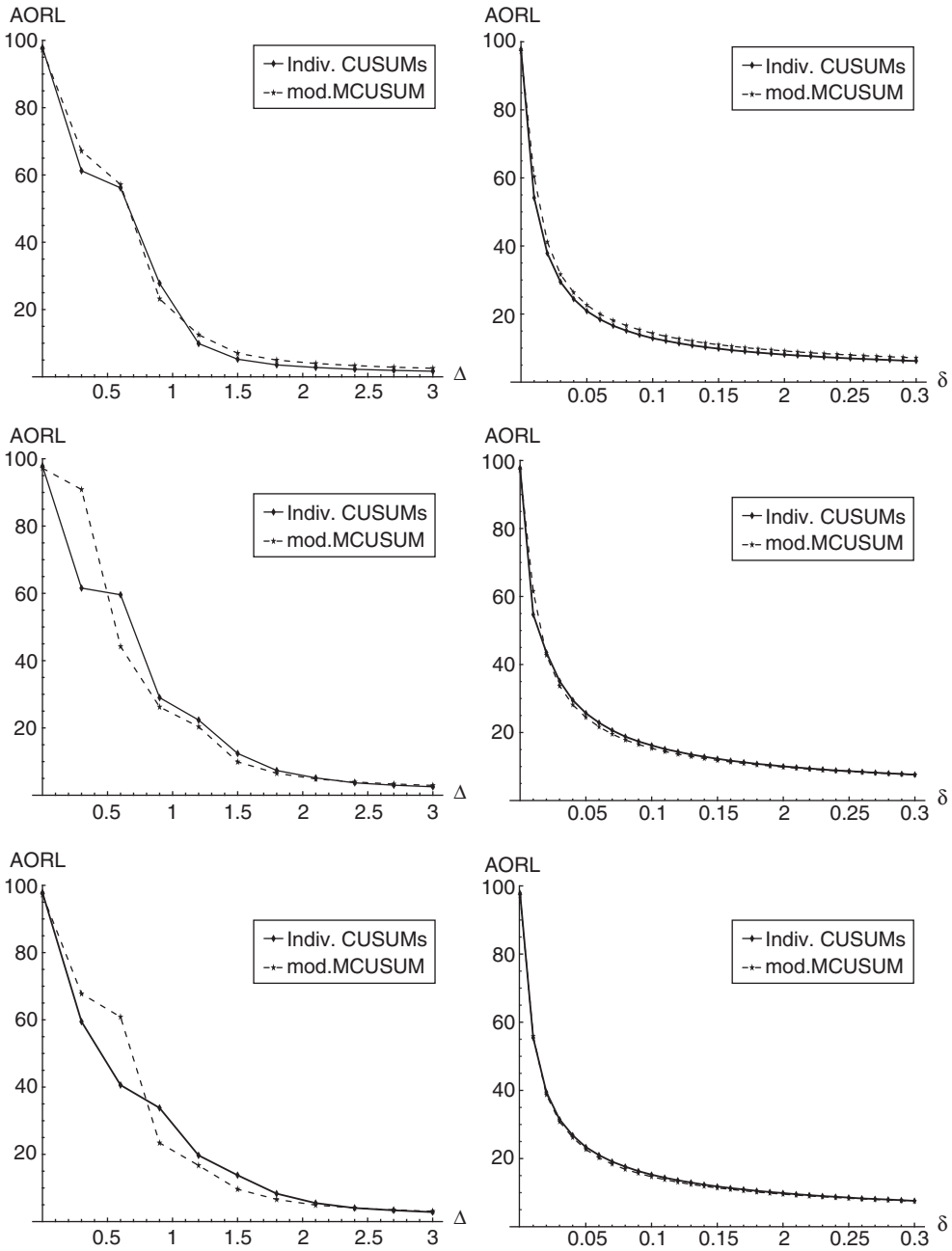


**FIGURE 9.** Performance comparison of the modified MCUSUM procedure and multiple simultaneous univariate CUSUM procedures using the real hospital data. On the left side the mean vector experiences a one-time jump of distance $\Delta$, and on the right side the mean vector linearly increases by $\delta$ at each time period. In the top row, the change occurs for only one hospital; in the middle row, it occurs for three hospitals; and in the bottom row, it occurs for all five hospitals.

comparisons via simulation". In particular, the individual CUSUMs tend to perform slightly better than the modified MCUSUM when the shift occurs in only one dimension (the top plots), whereas the modified MCUSUM tends to perform slightly better when the shift is in multiple dimensions (the middle and bottom plots). However, we also see that the difference in performance is smaller when the mean increases linearly (the right column of plots) than when there is a jump change in the mean vector (the left column of plots).

## DISCUSSION

In this article, I have demonstrated how to modify two directionally invariant multivariate procedures to make them directionally sensitive. The results of these and other simulations not included here show, not unexpectedly, that the modified multivariate procedures work better than their original counterparts in the problem for which they were designed. It is not unexpected as the modified procedures specifically look for positive changes so that, when given

such changes, they should outperform their counterparts that are not so designed.

An interesting result was the simulation comparison of simultaneous univariate Shewhart procedures and the modified $\chi^2$, which was mixed, with the better procedure depending upon the covariance structure (i.e., $\rho$). In contrast, in the simulations, the modified MCUSUM generally performed better than the simultaneous univariate CUSUMs for all values of (positive) $\rho$. Furthermore, it also performed better than the simultaneous Shewharts and the modified $\chi^2$ except in those cases where the shift $d$ was moderate to very large (in which case a statistical detection procedure may not even be required).

These results thus suggest that when the covariance structure is well known or well estimated, the modified MCUSUM procedure is the preferred choice for monitoring multivariate processes for directional shifts. However, the use of the modified MCUSUM does come with some costs. First, unlike Hotelling's $\chi^2$ procedure, the choice of threshold, and hence the ARL performance of the procedure, depends on the covariance structure of the data. Second, practitioners are often less comfortable using multivariate procedures because they tend to feel such procedures do not provide sufficient information about the cause(s) of a signal. The modified MCUSUM is no different in this regard, though because it is directional the practitioner is at least assured that the signal is related to an out-of-control condition of interest.

In the comparison between the modified MCUSUM and the simultaneous individual CUSUMs using the real hospital data, the two schemes performed very similarly, particularly for a linear increase in the mean and when applied to just the raw data without an out-of-control condition superimposed (figure 8). It is not clear from this work whether the similarity in performance is the result of having to estimate the covariance matrix, the choices of $k$ and $\mathbf{k}$, the presence of autocorrelation in the real data, or simply an artifact of the specific data that were used. These results do suggest that the current practice of using multiple simultaneous CUSUMs may provide the same performance as an appropriate multivariate method, but additional research is warranted.

Although the modified MCUSUM and simultaneous individual CUSUMs exhibited similar performance, in general each seemed to demonstrate a separate specific strength: the modified MCUSUM is slightly better in detecting small shifts in many or all dimensions, whereas the simultaneous individual CUSUMs seem better in detecting a shift in only one dimension. This suggests a strategy of using both in combination, where in the public health arena, for example, individual hospitals might monitor their own trends using individual CUSUMs, whereas a city, county, or state public health department might monitor an area using the modified MCUSUM. Alternatively, a public health department might use both the individual CUSUMs and the modified MCUSUM simultaneously, but interpret their signals differently: an individual CUSUM signal indicates the possibility of a localized event, whereas a modified MCUSUM signal indicates the possibility of a larger, area-wide event.

That said, figures 6 and 9 should give public health practitioners pause. These figures show that a bioterrorist attack that manifests itself as a one standard deviation increase in the mean may take anywhere from 10 to 20 time periods for a procedure to signal (given a false alarm rate of 1 period out of 100). If the data are daily counts, then on average it is going to take more than a week and perhaps up to 3 weeks to detect such an event. This performance can be improved by allowing a greater false alarm rate, but that comes at the expense of investigating and adjudicating the additional resultant signals. Of course, these methods may still be considerably better than other alternatives, particularly when the event is manifested as a relatively small increase in rates. But it is important for practitioners to recognize that these methods will not and cannot produce signals instantaneously except for events that manifest themselves via very large shifts.

### Assessing performance

This research departs from the bulk of similar surveillance research in the metrics it uses to assess performance. Surveillance research typically assesses performance using the metrics of sensitivity, specificity, and timeliness. See, for example, the discussion in Stoto et al. (7) and Shmueli (31). Using these metrics is akin to thinking of the problem as a sequence of hypothesis tests. Doing so naturally leads to thinking of the problem as a series of ROC curves and then raises questions about how to combine the information across the series of ROC curves to judge performance. See, for example, Kleinman and Abrams (32).

In this article, I used the metric standard in the SPC literature, the ARL, which I find more naturally focuses on the important dimension of the problem: speed of detection. That is, whether the problem is industrial quality control or bioterrorism incident detection, what is relevant is how fast a procedure gives a true signal (for a fixed false signal rate and a particular outbreak manifestation).

Among the difficulties in using sensitivity, specificity, and timeliness is that they are a little redundant, meaning that if one interprets "timeliness" as speed of detection, then increases in sensitivity (for a set specificity) must almost surely result in improvements in timeliness and vice versa. Hence, for the purposes of measuring and comparing performance, it seems necessary to use only two of the three metrics. Thus, if we think of the problem as assessing performance in terms of timeliness for a fixed specificity, then this is very much like saying assess performance by $ARL_1$ for a fixed $ARL_0$.

Now, although I suggest that it is important to focus on speed of detection, this does not necessarily imply that the ARL is the most appropriate metric for the syndromic surveillance problem. Indeed, note that the ARL is but one possible summary statistic for the entire distribution of run lengths that fully characterizes a procedure's speed of detection performance. For syndromic surveillance, it might be more appropriate to use the median run length or perhaps some other percentile of the run length distribution. If it is important to ensure that there is a high probability of outbreak detection before some number of days ('$x$')

after the onset, then one could use the number of run lengths less than $x$ as the appropriate performance measure.

Furthermore, some care must be taken when applying the run length paradigm to the syndromic surveillance problem. In the typical industrial SPC application, one assumes that once a process goes out of control, it stays out of control. Hence, the delay, as described in the section "Terminology, notation, and assumption," is always well defined. In syndromic surveillance, an outbreak ("out-of-control") condition can be short lived, and thus the concepts of the delay, out-of-control run length, and $ARL_1$ are more difficult to apply. One approach to resolving this difficulty is to focus on those types of outbreaks that take some time to manifest and to focus further on the initial period of the outbreak—say the first $x$ days. Then a metric such as the probability of detecting the outbreak within those first $x$ days, defined as the fraction of run lengths less than or equal to $x$, is still well defined.

Also note that focusing on the run length distribution as the primary measure of performance does not imply that this has been reduced to a one-dimensional problem. Rather, in the simplest case, there are three relevant dimensions that must be explored to evaluate a procedure's performance. In general terms, they are the $F_0$ run length performance, the $F_1$ run length performance, and the out-of-control (outbreak) condition. In the specific parameterization of the section "Performance comparisons via simulation," these are $ARL_0$, $ARL_1$, and $d$, which for a fixed $ARL_0$ were shown as two-dimensional plots of $ARL_1$ versus $d$. This is the simplest case because the outbreak condition could be expressed as a distance $d$ between mean vectors. For more complex syndromic surveillance scenarios, an outbreak may have to be characterized by multiple parameters. However, although it is more complex to analyze, the approach remains the same: the procedure that exhibits smaller $ARL_1$s (or other run length summary statistics) over a relevant range of false alarm rates and outbreak conditions is to be preferred.

Though I used the ARL in the simulations in the section "Performance comparisons via simulation," I introduced a new measure, the AORL, in the section "An application to syndromic surveillance" to assess performance on real data. There are a number of reasons why I did not use the more traditional ARL metric. Chief among them is that the real data—in fact, the syndromic surveillance problem in general—depart from the traditional industrial setting in a number of important ways. First, the idea of a process being "in control" in the syndromic surveillance setting is nebulous at best. In reality, there is no control over the "process," and the fluctuating data simply characterize the background state of natural disease incidence. Second, that background state could—and does in the case of these data—contain natural disease outbreaks. So, it is also not clear what "out-of-control" means in the syndromic context. It could mean the occurrence of a natural disease outbreak or a bioterrorism attack (or both) depending on the purpose of the surveillance.

The result is that calculating an $ARL_0$ by recording and averaging sequential runs on the real data, while computationally possible, is not a particularly informative measure. Part of the problem is that when a fluctuation occurs in the data (whether because of a natural outbreak or perhaps just because of an elevated incidence rate) the procedure tends to signal repeatedly for the same outbreak or fluctuation. Hence, simply averaging the sequential runs is not a good measure of the average time to the signal of an outbreak or attack as what is desired is the average time between *initial outbreak signals*.

In addition, if the out-of-control condition one is interested in is a bioterrorism attack, such an attack will occur within the natural disease incidence background state—including naturally occurring outbreaks. If only one series of sequential runs is used to calculate an out-of-control ARL, it is not clear exactly how one would impose the "attack" on top of the limited real data. That is, it is possible for the attack to occur during a natural disease outbreak period, or perhaps during a period when the natural disease incidence is unusually low, or at any other time. Hence, the procedure's performance could thus vary significantly depending on the timing of the attack.

The AORL overcomes this particular problem by evaluating all possible times when the attack could occur in the data. Hence, in the context of a bioterrorism attack occurring on top of the pattern of natural disease incidence, it provides a measure of how a procedure performs over all possible times of attack. In addition, as shown in the appendix, the AORL is approximately proportional to the average of the squared run lengths. Hence, in a computational sense, it is closely related to the ARL.

### Future research

As was first discussed in the section "Terminology, notation, and assumptions," the procedures used in this research were derived from a number of assumptions, particularly independent and identically distributed (*i.i.d.*) observations from a stationary distribution. Lack of independence, if not accounted for, results in higher false alarm rates (17). In this work, I accounted for this by empirically determining the thresholds to achieve a specific in-control ARL. Nonetheless, the question remains as to whether and when these methods based on *i.i.d.* assumptions perform better than those that explicitly account for autocorrelation (as well as other features that commonly occur in syndromic surveillance data, such as seasonal periodicities, long-term trends, and day-of-the-week and holiday effects).

In addition, in the presence of autocorrelation arising from nonstationarity, say an $F_0$ in which the mean follows a periodic cycle, the methods evaluated in this article probably need to be modified. That is, the methods and their evaluation in this article were designed around the data in figure 7, which did not exhibit an obvious regular seasonal or linear trend. Had they done so, then using some sort of moving average for the estimation of the mean vector would likely have been more appropriate than an average based on a fixed historical period. Future work should compare against data with annual and other periodicities that reflect the more general syndromic surveillance and public health problems. See, for example, the data plots in Shmueli (31).

Furthermore, even with our data, in the bioterrorism detection problem using a moving average to estimate the mean (and, more generally, using a moving window of data to estimate both the mean and covariance) might be more appropriate to account for natural disease outbreaks. However, the choices and trade-offs in how to construct a moving average such that it incorporates information from natural disease outbreaks in the estimation but *not* information from a bioterrorism attack are open and difficult questions.

Future work also should consider the effects of estimation on the performance of the procedures. In particular, the multivariate procedures require the estimation of the entire covariance matrix, whereas the simultaneous univariate procedures require only estimation of the diagonal elements. As discussed, whether and how this estimation affects the performance of the procedures is not well known. In addition, although the effects of changing $k$ in the CUSUM are known, the effects of changes in $\mathbf{k}$ in the modified MCUSUM are not and were not fully explored in this work.

I conclude by noting that the method that was applied to make the multivariate procedures directionally sensitive can be applied to other directionally invariant procedures, such as the nonparametric method of Qui and Hawkins (3). How the performance of those new methods compares with the performance of the modified MCUSUM requires further research.

## ACKNOWLEDGMENTS

I would like to acknowledge two anonymous reviewers, the associate editor, and one of the editors. This article was substantially improved as a result of their helpful suggestions. In addition, the section on assessing performance benefited from discussions with Lance Waller, Karen Kafadar, Mike Stoto, Dan Jeske, and Kathe Bjork. Though they may not necessarily agree with what I have written, my thoughts on the matter were greatly clarified as a result of our discussions.

## REFERENCES

1. Hotelling H. Multivariate quality control—illustrated by the air testing of sample bombsights. In: Eisenhart C, Hastay MW, Wallis WA, eds. Techniques of statistical analysis. New York: McGraw-Hill, 1947:409–12.
2. Crosier RB. Multivariate generalizations of cumulative sum quality control schemes. Technometrics 1988;30:291–303.
3. Qui P, Hawkins D. A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. The Statistician 2003;52:151–64.
4. Lowry CA, Montgomery DC. A review of multivariate control charts. IIE Transactions 1995;27:800–10.
5. Centers for Disease Control and Surveillance. Syndromic Surveillance: Reports from a National Conference, 2003. Morbidity and Mortality Weekly Report 2004;(Supplement), 53, September 24, 2004.
6. Fricker RD Jr., Rolka H. Protecting against biological terrorism: statistical issues in electronic biosurveillance. Chance 2006;19:4–13.
7. Stoto MA, Fricker RD Jr., Jain A, et al. Evaluating statistical methods for syndromic surveillance. In: Wilson A, Wilson G, Olwell D, eds. Statistical methods in counterterrorism: Game theory, modeling, syndromic surveillance, and biometric authentication. New York: Springer, 2006.
8. Woodall WH. The use of control charts in health-care and public-health surveillance. J Qual Tech 2006;38:1–16.
9. Woodall WH, Ncube MM. Multivariate CUSUM quality control procedures. Technometrics 1985;27:285–92.
10. Rogerson PA, Yamada I. Monitoring change in spatial pattens of disease: comparing univariate and multivariate cumulative sum approaches. Stat Med 2004;23:2195–214.
11. Testik MC, Runger GC. Multivariate one-sided control charts. IIE Trans 2006;30:635–45.
12. Follman D. A simple multivariate test for one-sided alternatives. J Am Stat Assoc 1996;91:854–61.
13. Perlman MD. One-sided testing problems in multivariate analysis. Ann Math Stat 1969;40:549–67.
14. Kudô A. A multivariate analogue of the one-sided test. Biometrika 1963;50:403–18.
15. Chang JT, Fricker RD Jr. Detecting when a monotonically increasing mean has crossed a threshold. J Qual Tech 1999;31:217–34.
16. Shewhart WA. Economic control of quality of manufactured product. Princeton, New Jersey: D. van Nostrand Company, Inc, 1931.
17. Montgomery DC. Introduction to statistical quality control. 4th edition. New York: John Wiley & Sons, 2001.
18. Page ES. Continuous inspection schemes. Biometrika 1954;41:100–15.
19. Lorden G. Procedures for reacting to a change in distribution. Ann Math Stat 1971;42:1897–908.
20. Pignatiello JJ Jr., Runger GC. Comparisons of multivariate CUSUM charts. J Qual Tech 1990;3:173–86.
21. Healy JD. A note on multivariate CUSUM procedures. Technometrics 1987;29:409–12.
22. Wolfram Research, Mathematica5 Doumentation. Available online at http://documents.wolfram.com/v5/ (accessed on October 12, 2006).
23. Reynolds MR Jr. Approximations to the average run length in cummulative sum control charts. Technometrics 1975;17:65–71.
24. Siegmund D. Sequential analysis tests and confidence intervals. New York: Springer-Verlag, 1985.
25. Joner MD Jr., Woodall WH, Reynolds MR Jr., Fricker RD Jr. The use of multivariate control charts in public health surveillance (draft) 2006.
26. Runger GC, Prabhu SS. A Markov chain model for the multivariate exponentially weighted moving averages control chart. J Am Stat Assoc 1996;91:1701–6.
27. Fricker RD Jr. Nonparametric control charts for multivariate data. Yale University, Ph.D. dissertation, 1997.
28. Espino JU, Wagner MM. The accuracy of ICD-9–coded chief complaints for detection of acute respiratory illness. AMIA Annu Symp Proc 2001;164–8.
29. Centers for Disease Control and Surveillance. Flu activity: reports & surveillance methods in the united states. Available online at www.cdc.gov/flu/weekly/fluactivity.htm (accessed December 14, 2005), 2005.
30. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE, Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proc Natl Acad Sci USA 2002;99:5237–40.
31. Shmueli G. Statistical challenges in modern biosurveillance. In submission to Technometrics (draft) 2006.

32. Kleinman K, Abrams AM. Metrics for assessing the performance of spatial surveillance. Statistical Methods in Medical Research 2006;15:445–64.

## APPENDIX

This appendix demonstrates that the AORL introduced in the section "Detecting bioterrorism" has a direct connection to the run lengths that would result from running a procedure sequentially through the data ($r_1$, $r_2$, $r_3$, …, $r_k$, defined below). In fact, I show that the AORL is proportional to the average of the squared sequential run lengths as the run lengths get arbitrarily large.

To start with, let us define how one would estimate the ARL for a series of data from $n$ time periods, $t = 1, 2, 3, \ldots, n$. Applying a procedure to the data series results in a set of sequential run lengths, $r_1, r_2, r_3, \ldots, r_k$, where the run of length $r_1$ starts at time $t = 1$, the run of length $r_2$ starts at time $t = r_1 + 1$, etc. By a run of length $r_1$ starting at time $t = 1$, I mean that the procedure that was started at time $t = 1$ did not signal for $r_1 - 1$ time periods and then signaled at the next time period. The final run of length $r_k$ was the last run to complete within the $n$ periods, so that $r_1 + r_2 + \cdots + r_k \leq n$. Note that each data point in the series was used in the computation of only one run length. Using these run lengths, then, the ARL would be estimated as

$$\widehat{\mathrm{ARL}} = \frac{1}{k}\sum_{i=1}^{k} r_i.$$

In contrast, in the section "Detecting bioterrorism," overlapping run lengths were calculated for runs starting at *each* time period. This results in a set of run lengths, $s_1, s_2, s_3, \ldots, s_n$, where the run of length $s_1$ started at time $t = 1$, the run of length $s_2$ started at time $t = 2$, the run of length $s_3$ started at time $t = 3$, etc. Here, most points in the data series were used in the computation of multiple run lengths. Using these overlapping runs, the AORL is calculated as

$$\mathrm{AORL} = \frac{1}{n}\sum_{i=1}^{n} s_i.$$

To demonstrate that the AORL is proportional to the average of the squared sequential run lengths, assume $r_1 + r_2 + \cdots + r_k = n$. That is, the final sequential run signaled precisely on the last period in the data series. We know $s_1 = r_1$ as the same procedure is being run on the same data starting at the same time period. In a similar way, we know $s_{r_1} + 1 = r_2$ and that every run length in $[r_1, r_2, r_3, \ldots, r_k]$ has a matching run length in $\{s_1, s_2, s_3, \ldots, s_n\}$. What this means is that the signal times for the sequential runs, $\{r_1, r_2, r_3, \ldots, r_k\}$, are also signal times in the larger set of runs, $\{s_1, s_2, s_3, \ldots, s_n\}$.

Now assume that the signal times for the sequential runs are the *only* signal times, so that, for example, for $i = 1, 2, 3, \ldots, r_1$ we have $s_i = s_1 - (i - 1)$. That is, the run lengths for each run after $s_1$ decrease by one for each time period up until the time $s_1$ signals. In a similar way for $i = r_1 + 1, r_1 + 2, r_1 + 3, \ldots, r_1 + r_2$, $s_i = s_2 - (i - 1)$. Etc.

Then, we can write

$$\begin{aligned}
\mathrm{AORL} &= \tfrac{1}{n}\sum_{i=1}^{n} s_i \\
&= \tfrac{1}{n}\left[\sum_{i=1}^{r_1} s_i + \sum_{i=r_1+1}^{r_1+r_2} s_i + \cdots + \sum_{i=r_1+\cdots+r_{k-1}+1}^{n} s_i\right] \\
&= \tfrac{1}{n}\left[r_1\left(\tfrac{s_1+1}{2}\right) + r_2\left(\tfrac{s_{r_1+1}+1}{2}\right)\right. \\
&\quad \left. + \cdots + r_k\left(\tfrac{s_{r_1+\cdots+r_{k-1}+1}+1}{2}\right)\right] \\
&= \tfrac{1}{n}\left[r_1\left(\tfrac{r_1+1}{2}\right) + r_2\left(\tfrac{r_2+1}{2}\right)\right. \\
&\quad \left. + \cdots + r_k\left(\tfrac{r_k+1}{2}\right)\right] \\
&= \tfrac{1}{2n}\sum_{i=1}^{k} r_i(r_i+1) \\
&\simeq \left(\tfrac{k}{2n}\right)\tfrac{1}{k}\sum_{i=1}^{k} r_i^2.
\end{aligned}$$

To walk through the expressions, the first equality is the definition of the AORL. The second simply separates the summation in the first expression into summations of those sets of sequential runs that decrease by 1 in each subsequent time period. The third equality takes each summation and replaces it by the number of runs in the summation times the mean run length, and the fourth substitutes the sequential run for the appropriate overlapping run. The fifth equality is just an algebraic simplification of the previous expression and the final follows if we assume $\sum r_i$ is negligible compared with $\sum r_i^2$.

The assumption that the overlapping runs signal only at the same times as the sequential runs signal is somewhat artificial, but not too far from what I observed empirically in the data presented in the section "An application to syndromic surveillance." That is, in general, subsequent overlapping runs almost always did decrease by one in the next time period. Violations of this occurred only when the run lengths became very small (in single digits), at which point the runs would frequently drop off to only one or two.