STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:3407–3429
Published online 31 January 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3197

Comparing syndromic surveillance detection methods: EARS' *versus* a CUSUM-based methodology[‡]

Ronald D. Fricker Jr*,†, Benjamin L. Hegler and David A. Dunfee

Operations Research Department, Naval Postgraduate School, 1411 Cunningham Road, Monterey, CA 93943, U.S.A.

SUMMARY

This paper compares the performance of three detection methods, entitled C1, C2, and C3, that are implemented in the early aberration reporting system (EARS) and other syndromic surveillance systems *versus* the CUSUM applied to model-based prediction errors. The cumulative sum (CUSUM) performed significantly better than the EARS' methods across all of the scenarios we evaluated. These scenarios consisted of various combinations of large and small background disease incidence rates, seasonal cycles from large to small (as well as no cycle), daily effects, and various types and levels of random daily variation. This leads us to recommend replacing the C1, C2, and C3 methods in existing syndromic surveillance systems with an appropriately implemented CUSUM method. Published in 2008 by John Wiley & Sons, Ltd.

KEY WORDS: syndromic surveillance; biosurveillance; early aberration reporting system; CUSUM

1. INTRODUCTION

The Centers for Disease Control and Prevention (CDC) as well as many state and local health departments around the United States are developing and fielding *syndromic surveillance* systems [1]. Making use of existing health-related data, these health surveillance systems are intended to give early warnings of bioterrorist attacks or other emerging health conditions.

A *syndrome* is 'a set of symptoms or conditions that occur together and suggest the presence of a certain disease or an increased chance of developing the disease' [2]. In the context of syndromic surveillance, a syndrome is a set of non-specific pre-diagnosis medical and other information that may indicate the release of a bioterrorism agent or natural disease outbreak. (See, for example, Syndrome Definitions for Diseases Associated with Critical Bioterrorism-associated Agents [3].)

*Correspondence to: Ronald D. Fricker Jr, Operations Research Department, Naval Postgraduate School, 1411 Cunningham Road, Monterey, CA 93943, U.S.A.
†E-mail: rdfricker@nps.edu

Contract/grant sponsor: Office of Naval Research; contract/grant number: N0001407WR20172

Received 22 June 2007
Accepted 26 November 2007

The data in syndromic surveillance systems may be clinically well defined and linked to specific types of outbreaks, such as groupings of ICD-9 codes from emergency room 'chief complaint' data, or only vaguely defined and perhaps only weakly linked to specific types of outbreaks, such as over-the-counter sales of cough and cold medication or absenteeism rates.

Since its inception, syndromic surveillance has mainly focused on early event detection: gathering and analyzing data in advance of diagnostic case confirmation to give early warning of a possible outbreak. Such early event detection is not supposed to provide a definitive determination that an outbreak is occurring. Rather, it is supposed to signal that an outbreak *may* be occurring, indicating a need for further evidence or triggering an investigation by public health officials (i.e. the CDC or a local or state public health department).

More recently, the purpose of syndromic surveillance has been expanded to include using 'existing health data in real time to provide immediate analysis and feedback to those charged with investigation and follow-up of potential outbreaks' [4]. This broader focus on *electronic biosurveillance* includes both early event detection and situational awareness. See Fricker [5, 6], Fricker and Rolka [7], and Stoto *et al.* [8] for more detailed exposition and discussion.

This paper assesses the performance of three syndromic surveillance early event detection methods, entitled 'C1', 'C2', and 'C3', that are implemented in the early aberration reporting system (EARS). EARS (www.bt.cdc.gov/surveillance/ears) was designed as a drop in syndromic surveillance system. 'Drop in' denotes a system designed to provide enhanced surveillance for a short duration around a discrete event (e.g. the Olympic Games or a national political convention), generally for which little or no prior data exist [4]. Since 11 September 2001, the EARS has also been increasingly used as a standard surveillance system (CDC, [9]). The C1, C3, and a modified form of the C2 methods are also implemented in the CDC's BioSense system (www.cdc.gov/biosense). BioSense is a national program intended to improve federal, state, and local public health capabilities for conducting near real-time biosurveillance, including both 'health situational awareness' and 'event recognition and response' (CDC, [10]).

In this paper we compare the C1, C2, and C3 methods with a cumulative sum (CUSUM) method applied to the prediction errors of a model based on the adaptive regression methodology proposed by Burkom *et al.* [11]. The need for a model-based approach arises because syndromic surveillance data generally contain uncontrollable trends, periodicities, and day-of-the-week and other effects. See Shmueli [12] for additional discussion. For a review of the use of control charts in the broader context of health care and public health surveillance, see Woodall [13].

The paper is organized as follows. In Section 2 the C1, C2, and C3 methods are described, as is the CUSUM and how the CUSUM was applied to residuals from an adaptive regression-based model. Section 3 describes how we generated synthetic background disease incident counts and outbreaks, and Section 4 describes the comparison methodology, including how we determined the form of the adaptive regressions used and how we selected various parameter values for the methods. Section 5 presents the results of the comparisons and the paper concludes in Section 6 with a discussion of the implications of our findings and some recommendations.

2. DESCRIPTION OF THE METHODS

2.1. EARS' C1, C2, and C3

The C1, C2, and C3 methods were intended to be CUSUM-like methods [6, 14] and, in fact, at least one paper [15] explicitly refers to them as CUSUMs. However, the C1 and C2 are actually

Shewhart variants that use a moving sample average and sample standard deviation to standardize each observation. (See Shewhart [16] or Montgomery [17] for more detail on the Shewhart method.) The C1 uses the 7 days prior to the current observation to calculate the sample average and sample standard deviation. The C2 is similar to the C1 but uses the 7 days prior to a 2-day lag. The C3 combines information from C2 statistics as described below.

Let $Y(t)$ be the observed count for period t representing, for example, the number of individuals arriving at a particular hospital emergency room with a specific syndrome on day t. The C1 calculates the statistic $C_1(t)$ as

$$C_1(t)=\frac{Y(t)-\bar{Y}_1(t)}{S_1(t)} \tag{1}$$

where $\bar{Y}_1(t)$ and $S_1(t)$ are the moving sample mean and standard deviation, respectively,

$$\bar{Y}_1(t)=\frac{1}{7}\sum_{i=t-1}^{t-7} Y(i) \quad \text{and} \quad S_1^2(t)=\frac{1}{6}\sum_{i=t-1}^{t-7}[Y(i)-\bar{Y}_1(i)]^2$$

As implemented in the EARS system, the C1 signals at time t when the C_1 statistic exceeds a threshold h, which is fixed at three sample standard deviations above the sample mean: $C_1(t)>3$.

The C2 is similar to the C1 but incorporates a 2-day lag in the mean and standard deviation calculations. Specifically, it calculates

$$C_2(t)=\frac{Y(t)-\bar{Y}_3(t)}{S_3(t)} \tag{2}$$

where

$$\bar{Y}_3(t)=\frac{1}{7}\sum_{i=t-3}^{t-9} Y(i) \quad \text{and} \quad S_3^2(t)=\frac{1}{6}\sum_{i=t-3}^{t-9}[Y(i)-\bar{Y}_3(i)]^2$$

and in EARS it signals when $C_2(t)>3$.

The C3 uses the C2 statistics from day t and the previous two days, calculating the statistic $C_3(t)$ as

$$C_3(t)=\sum_{i=t}^{t-2}\max[0, C_2(i)-1] \tag{3}$$

In EARS it signals when $C_3(t)>2$.

In our comparisons between the EARS methods and the CUSUM, we do not use the threshold values given above but adjust them to achieve an equal average time between false signals (ATFS) when no outbreak is present. The ATFS is analogous to the 'in-control average run length' in traditional statistical process control (SPC). How we calculated the ATFS and our reasons for preferring it are discussed in Section 4.

BioSense originally implemented the C1, C2, and C3 methods but has since modified the C2. The modified C2 method, called 'W2', calculates the mean and standard deviation separately for weekdays and weekends using the relevant prior 7 days of data with a 2-day lag. That is, the sample mean for weekdays equals the mean of the previous 7 weekdays prior to a 2-day lag, and the sample mean for weekends equals the mean of the previous 7 weekend days prior to a two-day lag. BioSense also allows the user to set the threshold for the W2 rather than use the threshold associated with the C2 (CDC, [18]).

2.2. The CUSUM

The CUSUM method of Page [19] and Lorden [20] is a well-known SPC methodology. Montgomery [17] is an excellent introduction to the CUSUM method in an industrial SPC setting and Hawkins and Olwell's method [21] is a comprehensive treatment of the CUSUM.

Formally, 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 method monitors the statistic $S(t)$, which satisfies the recursion

$$S(t)=\max[0, S(t-1)+L(t)] \tag{4}$$

where the increment $L(t)$ is the log-likelihood ratio:

$$L(t)=\log\frac{f_1[Y(t)]}{f_0[Y(t)]}$$

The method is usually started at $S(0)=0$; it stops and concludes that $Y\sim F_1$ at the first time when $S(t)>h$, for some threshold h that achieves a desired ATFS when $Y\sim F_0$ (i.e. when no outbreak is present).

If f_0 and f_1 are normal densities with means μ and $\mu+\delta$, respectively, and unit variances, then equation (4) reduces to

$$S(t)=\max[0, S(t-1)+Y(t)-\mu-k] \tag{5}$$

with $k=\delta/2$, where k is commonly referred to as the *reference interval*. If Y's are independent and identically distributed according to f_0 before some unknown change point and according to f_1 after the change point, then the CUSUM has certain optimality properties; see Moustakides [22] and Ritov [23].

Equation (5) is the form routinely used, even when the underlying assumptions are only approximately met. However, in those and more general situations, the choice for the reference interval value is less well defined. In our work, for reasons we discuss in Section 4.2, we chose to still set $k=\delta/2$.

Also, equation (5) is a one-sided CUSUM, meaning that it will only detect increases in the mean. If it is important to detect both increases and decreases in the mean, a second CUSUM must be used to detect decreases. In syndromic surveillance, since it is important only to quickly detect increases in disease incidence, a second CUSUM is generally unnecessary.

In industrial settings, the CUSUM is applied directly to the observations because some control is exhibited over the process such that it is reasonable to assume F_0 is stationary. In syndromic surveillance, this is generally not the case as the data often have uncontrollable systematic trends, such as seasonal cycles and day-of-the-week effects. One solution is to model the systematic component of the data, use the model to forecast the next day's observation, and then apply the CUSUM to the forecast errors [17].

Examples in the literature of model-based methods include Burkom *et al.* [11], who compared a log-linear regression model, an 'adaptive regression model with sliding baseline,' and a Holt–Winters method for generalized exponential smoothing; Brillman *et al.* [24], who applied the CUSUM to prediction errors; the CDC's cyclical regression models discussed in Hutwagner *et al.* [14]; log-linear regression models in Farrington *et al.* [25]; and time series models in Reis and Mandl [26]. See Shmueli [12] for additional discussion of the use of regression and time series methods

for syndromic surveillance. Also see Lotze *et al.* [27] for a detailed discussion of preconditioning applied to syndromic surveillance data.

2.3. Applying the CUSUM to adaptive regression residuals

We used the 'adaptive regression model with sliding baseline' of Burkom *et al.* [11] to model the systematic component of the syndromic surveillance data. The basic idea is as follows. Let $Y(i)$ be an observation, say chief complaint count on day i. Regress the observations for the past n days on time relative to the current period. Then use the model to predict today's observation and apply the CUSUM to the difference between the predicted value and today's observed value. Repeat this process each day, always using the most recent n observations as the sliding baseline in the regression to calculate the forecast error. For $t>n$ and assuming a linear formulation with day-of-the-week effects, the model is

$$Y(i)=\beta_0+\beta_1\times(i-t+n+1)+\beta_2 I_{\text{Mon}}+\beta_3 I_{\text{Tues}}+\beta_4 I_{\text{Wed}}+\beta_5 I_{\text{Thurs}}+\beta_6 I_{\text{Fri}}+\beta_7 I_{\text{Sat}}+\varepsilon \quad (6)$$

for $i=t-1,\ldots,t-n$. The I's are indicators, where $I=1$ on the relevant day of the week and $I=0$ otherwise, and ε is the error term that is assumed to follow a symmetric distribution with mean 0 and standard deviation σ_ε. Of course, as appropriate, the model can also be adapted to allow for nonlinearities by adding a quadratic term into equation (6).

Burkom *et al.* [11] used an 8-week sliding baseline ($n=56$). We compared the performance for a variety of n values and between a linear and quadratic form of the model. Section 4.1 describes how we determined the form for the adaptive regression and the n values.

The model is fit using ordinary least squares, regressing $Y(t-1),\ldots,Y(t-n)$ on $n,\ldots,1$. Having fit the model, the forecast error when day t is a Sunday is

$$r(t)=Y(t)-[\hat{\beta}_0+\hat{\beta}_1\times(n+1)]$$

where $\hat{\beta}_0$ is the estimated slope and $\hat{\beta}_1$ is the estimated intercept. For any other day of the week, the forecast error is

$$r(t)=Y(t)-[\hat{\beta}_0+\hat{\beta}_1\times(n+1)+\hat{\beta}_j]$$

where $\hat{\beta}_2$ is the estimated day-of-the-week effect for Monday, $\hat{\beta}_3$ is for Tuesday, etc.

Standardizing $r(t)$ on σ_ε, we have $x(t)=r(t)/\sigma_\varepsilon$ and the CUSUM is thus

$$S(t)=\max[0,\,S(t-1)+x(t)-k] \quad (7)$$

where we assume that the expected value of the residuals is zero. (If σ_ε is not known, it can be estimated in the usual way from the residuals.) It now remains to determine k.

As shown in the Appendix A, we can estimate the standard deviation of the forecast error for a simple linear adaptive regression as

$$\sigma_{\text{p.e.}}=\sigma_\varepsilon\sqrt{\frac{(n+2)(n+1)}{n(n-1)}} \quad (8)$$

Assuming that it is important to detect an increase in the mean disease incidence of one standard deviation of the prediction error, so that $\delta = \sigma_{\text{p.e.}}$, we then set

$$k = \frac{1}{2}\sqrt{\frac{(n+2)(n+1)}{n(n-1)}}$$

where σ_ε (or $\hat{\sigma}_\varepsilon$) does not appear in the expression because the CUSUM in equation (7) uses the standardized residuals. See Section 4.2 for further discussion and the Appendix A for derivation of equation (8) as well as the equivalent expression for a multiple regression incorporating day-of-the-week indicator variables.

3. SIMULATING SYNDROMIC SURVEILLANCE DATA

In order to compare the methods, we simulated the occurrence of a background disease incidence and then overlaid various types of simulated bioterrorism attacks/natural disease outbreaks (which we will refer to herein simply as 'outbreaks'). The simulations were conducted in MatLab 7.1.0.246 using the *randn* function to generate random normal variates and *lognrnd* to generate lognormal random variates. The simulations of both background disease incidence and outbreaks are purposely idealized depictions designed to capture the main features of syndromic surveillance data. The use of simulation and the idealization of the data features was done for the following reasons:

- so that we could definitively compare and contrast the relative performance of the various methods under known conditions, and
- so that we could clearly distinguish how the various features of the data did or did not affect each method's performance.

The background disease incidence data were simulated as the sum of a mean disease incidence, a seasonal sinusoidal cycle, a systematic day-of-the-week effect, and a random fluctuation. Outbreaks, when they occurred, were incorporated as another additive term. That is, a daily observation $Y(t)$ was simulated as

$$Y(t) = \max(0, \lceil c + s(t) + d(t) + Z_c(t) + o(t) \rceil), \quad t = 1, 2, 3, \ldots \tag{9}$$

where

- c is a constant level of disease incidence;
- s is the seasonal deviation;
- d is the day-of-the-week effect;
- $Z_c(t)$ is the random noise around the systematic component, $c + s(t) + d(t)$;
- $o(t)$ is the mean outbreak level that, when an outbreak is occurring, increases the disease incidence level as described below; and
- $\lceil x \rceil$ is the ceiling function, which rounds x up to the next largest integer.

The seasonal effect is calculated as $s(t) = A[\sin(2\pi t/365)]$, where A is the maximum deviation from c with $t = 1$ corresponding to October 1st on a 365 day per year calendar. For the random component, we assumed $Z \sim \text{N}(\mu, \sigma^2)$ when c is large and $Z \sim L\text{N}(\mu, \sigma^2)$ when c is small. The day-of-the-week effect is the systematic deviation from $c + s(t)$, where $d(t) = d(t+7)$ for all t.

It is defined in terms of σ, a parameter of Z. In particular, we assumed $d=-0.5\sigma$ on Sunday, $d=0.1\sigma$ on Monday, $d=0.2\sigma$ on Tuesday, $d=0.3\sigma$ on Wednesday, $d=0.4\sigma$ on Thursday, $d=0$ on Friday, and $d=-0.3\sigma$ on Saturday.

Table I specifies the parameter values for equation (9), which define 12 'scenarios' designed to span a range of possible underlying disease incidence patterns. Scenarios 1–6 are large-count scenarios and scenarios 7–12 are low-count scenarios. The parameters were selected to generate synthetic data that mimic disease incidence patterns similar to selected data sets at the CDC's EARS simulation data sets (www.bt.cdc.gov/surveillance/ears/datasets.asp). In particular, $c=90$, $A=80$, $\mu=0$, and $\sigma=30$ or 10 in equation (9) result in disease incidence patterns similar to EARS data set S08. Setting $c=90$, $A=20$, $\mu=0$, and $\sigma=10$ results in disease incidence patterns similar to the S01 data set, as well as other patterns that are intermediate between S01 and S08. For scenarios 7–12, combinations of the values in Table I result in disease incidence patterns similar to S03, S04, S15, and S34.

The specific EARS data sets that we mimicked were chosen in consultation with a CDC expert. For the low-count scenarios, data set S04 is characteristic of hospital-level respiratory or influenza-like illness (ILI) chief complaint counts, S03 of hospital-level rash chief complaint counts, and S34 of hospital-level neurological chief complaint counts. For the high-count scenarios, S08 is characteristic of state-level aggregate respiratory or ILI chief complaint counts, S45 of state-level aggregate gastrointestinal chief complaint counts, and S15 of state-level neurological chief complaint counts [28].

Although these choices may seem either arbitrary or too restrictive, meaning that they do not characterize some particular pattern that occurs in a particular syndromic surveillance setting, we chose them because they capture a wide range of data patterns. Furthermore, as we will show in Section 5, the adaptive regression turns out to be remarkably good at removing the systematic trends in the data so that the specific choices made above are actually of little import and have little impact on the final result.

Figure 1 shows two sets of simulated data. The plot on the left shows one year of scenario 1 data simulated via equation (9): $c=90$, $A=80$, $\mu=0$, and $\sigma=30$ (so that $Z\sim \mathrm{N}(0,30^2)$). The plot on the right shows one year of scenario 7 data: $c=0$, $A=6$, and $Z\sim L\mathrm{N}(1,0.7^2)$.

Table I. Parameters for equation (9) for scenarios 1–12.

Scenario	c	A	μ	σ
1	90	80	0	30
2	90	80	0	10
3	90	20	0	30
4	90	20	0	10
5	90	0	0	30
6	90	0	0	10
7	0	6	1.0	0.7
8	0	6	1.0	0.5
9	0	2	1.0	0.7
10	0	2	1.0	0.5
11	0	0	1.0	0.7
12	0	0	1.0	0.5

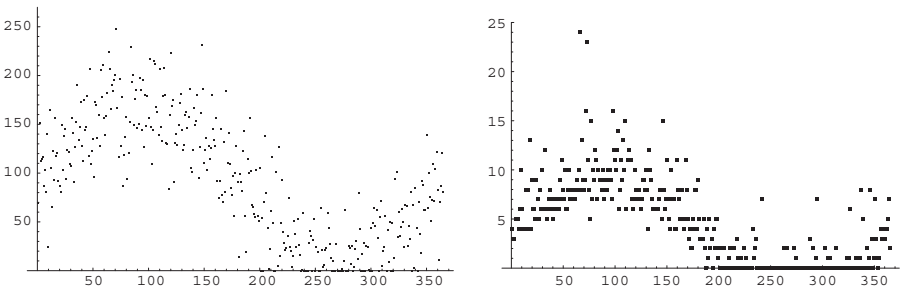

Figure 1. Examples of simulated data. The left plot shows one year of scenario 1 data and the right plot shows one year of scenario 7 data.

As described above, outbreaks were incorporated into equation (9) as an additive term $o(t)$ representing the mean outbreak level. As with the simulated data itself, we employed an idealized outbreak form that could be parameterized simply, in terms of a peak magnitude M, a duration D, and a random start day τ, where outbreaks increased linearly up to M and then linearly back down to zero:

$$o(t)=\begin{cases} M[2(t-\tau+1)/(D+1)], & \tau\leqslant t\leqslant \tau+D/2-\frac{1}{2} \\ M[1-(2(t-\tau)-D+1)/(D+1)], & \tau+D/2-1/2<t\leqslant \tau+D-1 \\ 0 & \text{otherwise} \end{cases}$$

We evaluated the methods' performance for outbreaks of various magnitudes and durations. For scenarios 1–6, we used three magnitudes—small, medium, and large—defined as a fraction of the constant disease incidence c: $M=0.1c$, $0.25c$, and $0.5c$, respectively (where $c=90$ from Table I). For scenarios 7–12, we used four magnitudes—very small, small, medium, and large—defined as a fraction of the mean plus three standard deviations of the lognormally distributed random variable Z from equation (9). In particular, for a large outbreak we defined $M=E(Z)+3\sigma_Z$, medium $M=0.5[E(Z)+3\sigma_Z]$, small $M=0.25[E(Z)+3\sigma_Z]$, and very small $M=0.1[E(Z)+3\sigma_Z]$, where

$$E(Z)=\exp(\mu+\sigma^2/2)$$

and

$$\sigma_Z^2=(\mathrm{e}^{\sigma^2}-1)\mathrm{e}^{2\mu+\sigma^2}$$

where μ and σ are specified in Table I for the various scenarios. For all the scenarios we considered, durations and ranged from short to long: $D=3,5,\ldots,15$ days.

As we previously mentioned, the characterization of disease incidence in equation (9) is purposely idealized in order to facilitate comparison of the relative performance of the methods under various scenarios. The idea is to mimic the most salient and important features of syndromic surveillance data in a simulation environment where we can know precisely when outbreaks occur so that we can clearly assess and evaluate performance. However, it is important to note that the methods do not exploit the idealized features of the data and can be readily adapted to account

for those features of real data that are not included in equation (9). For example:

1. *Regular seasonal cycles*: Equation (9) contains a regular, perfectly idealized cycle that could be exploited to make artificially accurate predictions. However, the C1, C2, and C3 methods do not make use of this information. Similarly, as will be discussed in Section 4.1, the length of time over which the adaptive regressions are run is too short to model the sinusoid and thus the CUSUM applied to the residuals also does not use the information in the idealized cycle.
2. *No linear trends*: Growing or shrinking populations, or changes in health conditions, could result in linear (or other) trends in the disease incidence. A trend term is not included in equation (9) since, if the methods can appropriately adjust for the seasonal component, it can also adjust for a linear trend.
3. *No holidays or other such effects*: Holiday and other types of systematic effects are often present in real syndromic surveillance data. Although these effects are not included in equation (9), it would be relatively straightforward to extend these methods to account for holiday effects.
4. *Constant year length*: Years as parameterized in equation (9) are always 365 days long. However, extending this to account for leap years is an unnecessary complication that does not affect the results or conclusions.

See Shmueli [12] for a detailed exposition on the features of syndromic surveillance data. See Fricker *et al.* [29] and Hu and Knitt [30] for comparisons of a multivariate CUSUM to a multivariate exponentially weighted moving average method using data simulated much as was done here. See Kleinman *et al.* [31] for an alternate methodology designed to simulate syndromic surveillance data in both space and time.

4. COMPARISON METHODOLOGY

The metrics used to compare performance between methods were (1) the fraction of times a method missed detecting an outbreak and (2) the average time to first outbreak signal (ATFOS). The former is a measure of detection capability, whereas the latter is a conditional measure of the timeliness of detection. The ATFOS is defined as the average time until the first signal among all simulations for which a signal occurred during the outbreak period. Clearly, performance in both dimensions must be considered since a desirable method must simultaneously have a short ATFOS and a low fraction of outbreaks missed. A method that is small in one dimension while being large in the other is not particularly useful.

This approach differs from most of the syndromic surveillance literature that attempts to evaluate performance simultaneously in three dimensions: 'sensitivity, specificity, and timeliness.' Although we do assess performance in terms of timeliness via ATFOS and fraction missed, a sensitivity-like measure, we use a fixed ATFS in the third dimension to simplify the analysis.

This is similar to the approach used in the SPC literature, where the ATFS is roughly equivalent to the 'in-control average run length' and the ATFOS is equivalent to the 'out-of-control average run length.' The average run length (ARL) is the average number of observations until a signal. In the SPC literature, it is the common and well-accepted practice to compare the performance of methods by first setting thresholds that achieve a specific in-control average run length and then compare out-of-control average run lengths under various conditions. The method that demonstrates

lower out-of-control average run lengths across a variety of conditions deemed important is judged to be the better method.

However, this approach differs from the SPC literature because we also use the fraction missed metric. In the SPC literature, once a process goes out of control, it is assumed to stay in that condition until a method signals and the cause is identified and corrected. Thus, once a method goes out of control, any signal is a true signal. This is not the case in syndromic surveillance where outbreaks are transient and after some period of time disappear. In this situation, it is possible for a method to fail to signal during an outbreak, after which a signal is a false signal.

Returning to the syndromic surveillance literature's 'specificity' metric, we prefer ATFS because the concept of specificity is not well defined in sequential testing problems. In classical hypothesis testing, specificity is the probability that the null hypothesis is not rejected when it is true. It is one minus the probability of a Type I error. In medicine, it is the probability that a medical test will correctly indicate that an individual does not have a particular condition. However, syndromic surveillance involves sequential testing where, in the absence of an outbreak, the repeated application of any method will eventually produce a false signal. In other words, in the absence of an outbreak, one minus the 'specificity' for a sequential test must approach 100 per cent as the number of tests is allowed to get arbitrarily large.

In the syndromic surveillance literature, specificity is often (re)defined as the fraction of times a method fails to signal divided by the number of times the method is applied to a stream of syndromic surveillance data without outbreaks (cf. Reis *et al.* [32]). If the data are independent and identically distributed from day to day, and if the method results in test statistics that are independent and identically distributed from day to day as well, then such a calculation is an appropriate estimate of the specificity of a test on a given day. However, syndromic surveillance data are generally autocorrelated and, even if they were not, any method that uses historical data in the calculation of a test statistic will produce autocorrelated statistics. Under these conditions, it is not clear what the quantity from the above calculation represents. It is certainly not specificity in the classical hypothesis testing framework. See Fraker *et al.* [33] for additional discussion.

Hence, for each scenario in Table I, we determined the threshold for each method that gave an ATFS of 100 days. The ATFS is a measure of the time between clusters of false signals. It would be equivalent to the average time between signal events (ATBSE) metric of Fraker *et al.* [33] if they allowed the method to be reset to its initial condition after a 'signal event' and the data related to the signal event is removed from the adaptive regression. All other things being equal, larger ATFS values are to be preferred.

The thresholds to achieve a particular ATFS were determined empirically as follows. For a particular scenario and method (without any outbreaks superimposed), we chose an initial h and ran an method r times, recording for each run i the time t_i when the method first signalled. The ATFS was estimated as $\sum_{i=1}^{r} t_i/r$, and we then iteratively adjusted h and re-ran the method to achieve the desired ATFS, eventually setting r large enough to make the standard error of the estimated ATFS acceptably small (less than one day). Once the thresholds were set, the methods were then compared across all the scenarios specified in Table I for all the outbreak types described in Section 3. While we empirically determined the necessary thresholds via simulation, note that approximations have been derived for setting thresholds for the CUSUM under the assumption of *iid* normal observations; see Reynolds [34] and Siegmund [35].

The purpose of setting the thresholds to achieve equal time between false alarms was to ensure a fair comparison between the methods. That is, it is always possible to improve a method's ability to detect an actual outbreak by lowering the threshold, but this comes at the expense of also

decreasing the ATFS. Thus, by first setting the thresholds to achieve equal time between false alarms, we could then make an objective judgement about which method was best at detecting a particular type of outbreak.

Across all the scenarios we evaluated, the CUSUM thresholds ranged from $h=2.9$ to 4.2, including all combinations of c, A, μ, σ and with and without day-of-the-week effects. For the EARS methods, across all of the scenarios we observed for C1: $2.7 \leqslant h \leqslant 8.2$; for C2: $2.6 \leqslant h \leqslant 7.4$; and for C3: $3.0 \leqslant h \leqslant 18.2$.

Having set the thresholds to achieve equal ATFS performance, the ATFOS and fraction missed were calculated as follows. For each iteration i, the methods were run for 100 time periods (using data from $100+n$ time periods so that the adaptive regression could be fit for period 1) without any outbreaks. If a method signalled during this time it was reset and restarted, just as it would be in a real application. This allowed the CUSUM statistics to be in a steady-state condition at the time of the outbreak. Outbreaks began at time 101 and continued for the appropriate duration. If the method signalled at time t_i within the duration of the outbreak, the time to first outbreak signal was recorded as $t_i - 100$ and the ATFS was estimated as $\sum_{i=1}^{s}(t_i - 100)/s$ for the s iterations that signalled within the outbreak duration. The fraction missed was calculated as the number of iterations for which the method failed to signal during the outbreak divided by the total number of iterations run.

4.1. Determining the form of the adaptive regressions

When using regression to predict future observations, the question naturally arises as to how much historical data should be used for the regression's sliding baseline. We evaluated the performance of two alternatives drawn from existing practice and the literature: (1) 7 days of historical data (i.e. $n=7$), which matches what is used in the C1, C2, and C3 EARS methods and (2) 8 weeks of historical data ($n=56$) as recommended by Burkom *et al.* [11].

Of course, all other factors being equal, regressions based on a shorter sliding baseline will less accurately estimate the underlying systematic trends in the data than those based on longer sliding baselines. However, although a longer sliding baseline should allow for a more detailed regression model and presumably a better prediction, often in syndromic surveillance the amount of available data is limited or the older data of questionable relevance due to changing trends or phenomena. Hence, there is a trade-off to be made between the amount of historical data used in a particular model and the predictive accuracy of that model.

As described in Dunfee and Hegler [36], this led us to also determine and evaluate the performance of the 'optimal' sliding baseline (n) for each scenario (c, A, μ, σ combination). For each of the 12 scenarios we studied, we determined the optimal n to later use in the actual regression models for the comparisons. In addition, two separate regression models were evaluated in order to decide on the best form of the model. The first was a linear regression model with an intercept term, a slope term, and, for simulations with day-of-the-week effects, six indicator variables. The second was a regression model that also incorporated a squared term that allowed it to model quadratic trends in the data, where the squared term was simply the time squared.

Figure 2 shows an example of how we assessed the form of the adaptive regression and determined the 'optimal' sliding baseline for scenario 2 ($c=90, A=80, \mu=0, \sigma=10$, with and without day-of-the-week effects). The optimal n was chosen by visual inspection with the criteria that n be not only as small as possible but also as close to achieving the minimum average squared residual as possible. This means that we chose the smallest n that achieved most of the reduction

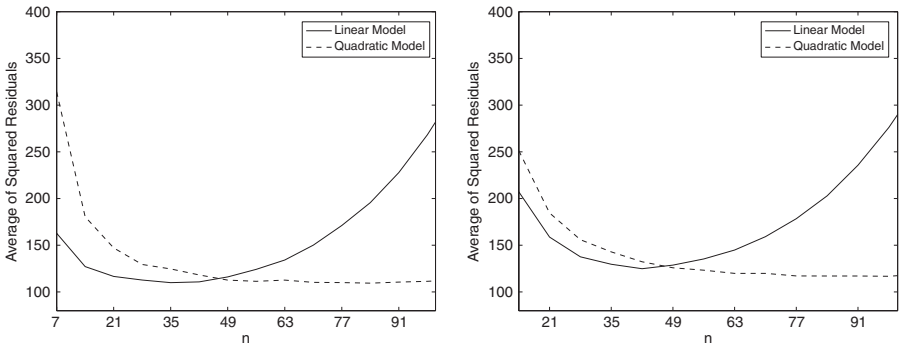

Figure 2. Average squared residuals for linear and quadratic models as a function of the amount of historical data—i.e. the size of the sliding baseline (n)—used to fit the regression models under scenario 2 (see Table I). On the left are the results when there is no day effect in the data and on the right the data have day effects. From this, we determined that the best model was linear with an 'optimal n' of about 30 days for no day effects and about 40 days with day effects.

in the average squared residual and not the n that occurred precisely at the minimum point on the curve. For example, for the curves on the left in Figure 2, which are based on data without a day-of-the-week effect, we determined that the 'optimal n' is 30 days for the linear model and around 35 days for the quadratic model. For the plot on the right, for which the data had day-of-the-week effects in it, we determined that the 'optimal n' is 40 days for the linear model and around 50 days for the quadratic model.

Figure 2 shows that the linear model achieved almost the same minimum average squared residual as the quadratic model but with a smaller n. As described in Dunfee and Hegler [36], this occurred consistently for all of the scenarios leading us to choose a linear adaptive regression model in all of our evaluations. For the linear model, across all the scenarios without day-of-the-week effects, the optimal n values ranged from 15 to 40 days. For the scenarios with day-of-the-week effects, the optimal n values increased, with the largest being around 56 days—the size recommended by Burkom *et al.* [11].

4.2. Determining k for the CUSUM

As discussed in Section 2.2, a common setting for the reference interval is $k=\delta/2$, a setting that results from the derivation of the CUSUM as a sequential likelihood ratio test under such assumptions as stationarity, normality, and independence, as well as the assumption that the mean of the distribution experiences a one-time jump from μ to $\mu+\delta$. Such assumptions are unlikely to apply to raw syndromic surveillance data. In spite of this, we chose to set $k=\delta/2$ for a couple of reasons.

First, we are not applying the CUSUM to the raw data, but rather we are applying it to the residuals from an adaptive regression. While such residuals are not precisely normally distributed, it is reasonable to assume that they are approximately normally distributed, and that to a first order they approximately meet the other assumptions of the CUSUM. Certainly the adaptive regression residuals come much closer to meeting the assumptions than the raw data. To the extent that the residuals meet the assumptions used to derive the CUSUM as a sequential likelihood ratio test, then our choice of k is justified by theory.

Second, note that the choice of k can be based on appealing to the form of the CUSUM,

$$S(t)=\max[0, S(t-1)+Y(t)-\mu-k]$$

and specifically to the expected value of the 'increment,' $Y(t)-\mu-k$. (We use the term increment because at each time t, this is the amount by which the CUSUM statistic is incremented.) In particular, note that the CUSUM will tend to increase when $E[Y(t)]-\mu>k$. Thus, assuming an increase in the mean incidence rate to $\mu+\delta$ is important to detect, then since $E[Y(t)]-\mu=\mu-\delta-\mu=\delta$, setting $k<\delta$ will result in an alarm (eventually) because the CUSUM will tend to accumulate positive increments over time until the signaling threshold is reached.

Thus, using this logic, any choice of $0\leqslant k\leqslant\delta$ is justified. As explained by Page [19] in his seminal work on the CUSUM, *Continuous Inspection Schemes*, 'The system of scoring [i.e. the reference interval value] is chosen so that the mean sample path on the chart when quality is satisfactory is downwards, i.e. of negative gradient, and is upwards when quality is unsatisfactory.' Indeed, Chang and Fricker [37] found $k=\delta$ to work well in a comparison of the CUSUM *versus* a generalized likelihood ratio test for detecting a monotonically increasing mean.

However, there is a trade-off to be made since, for example, setting $k=0$ will result not only in detection of a mean μ' that shifts to $\mu+\delta$ but also in detection to any mean level $\mu'>\mu$. Such a CUSUM will be overly sensitive and thus will require setting a high threshold h in order to achieve the desired ATFS. The result of the high threshold is that CUSUM performance for a shift to $\mu+\delta$ will actually be degraded. Hence, the phrase used in the preceding paragraph is not inconsequential: 'assuming an increase in the mean incidence rate to $\mu+\delta$ is important to detect'. The idea here is that one must make an *a priori* choice of what is important to (quickly) detect and what is not.

In Chang and Fricker [37], the problem was posed such that there was a clear division between the 'good' mean levels and 'bad' mean levels. That is, any mean μ' where $\mu'<\mu+\delta$ was good whereas any $\mu'\geqslant\mu+\delta$ was bad. However, in the syndromic surveillance problem such a clear distinction does not exist: if $\mu+\delta$ is important to detect, then $\mu+\delta-\varepsilon$, for some small ε, is probably just slightly less important to detect. Thus, our choice of $k=\delta/2$ means that we have implicitly set the CUSUM to signal for any $\mu'>\mu+\delta/2$, with an emphasis on quick detection for $\mu'\geqslant\mu+\delta$.

Having chosen to set $k=\delta/2$, in scenarios 1–6 because c was large the quantity $c+s(t)+d(t)+Z_c(t)$ was only very rarely negative so that (1) σ_ε was constant throughout the year and (2) $\sigma_\varepsilon=\sigma$. Thus, for those CUSUMs with larger sliding baselines (e.g. $30\leqslant n\leqslant 60$), because the term under the square root in equation (A1) was nearly one, we set $k=\frac{1}{2}$; for the CUSUMs using a 7-day sliding baseline (designed to match the baseline used in the EARS methods), we set $k=0.65$.

However, in scenarios 7–12 the quantity $c+s(t)+d(t)+Z_c(t)$ can be frequently negative (for some σ and A combinations) during those portions of the year when the seasonal component $s(t)$ is at its lowest point. Per equation (9), this means that there are periods of the year when the simulated observations are often zero and other times of the year corresponding to the peaks in the seasonal component the simulated observations are always non-zero with sometimes fairly large counts. This type of behavior results in a non-constant σ_ε. Under these conditions, one can either define an adaptive CUSUM that estimates σ_ε for various times of the year or fix a value. We did the latter, fixing $\sigma_\varepsilon=\sigma_Z$; thus, we used σ_Z to standardize the residuals for the CUSUM in equation (7) and in the calculations for the standard deviation of the prediction error in equation (8). Thus, having fixed $\sigma_\varepsilon=\sigma_Z$, as in scenarios 1–6, with larger sliding baselines (e.g. $30\leqslant n\leqslant 60$) we set $k=\frac{1}{2}$ and for the CUSUMs using a 7-day sliding baseline we set $k=0.65$.

Fixing $\sigma_\varepsilon = \sigma_Z$ means that the performance of the CUSUM now varies by time of year. In particular, although on an annual basis the threshold was set to achieve ATFS = 100 days, during those periods where the seasonal component is large (in the 'winter') the ATFS is lower than the annual average (approximately 66 days) and during those periods of the year when the seasonal component is small (the 'summer') the ATFS is larger (approximately 181 days). Furthermore, this means that outbreaks that occur in the winter are relatively easier to detect but those that occur in the summer are relatively harder to detect. As we can see in Section 5, this differential did not seem to have a major impact on overall performance.

5. RESULTS

Figure 3 in many ways shows the results of all the evaluations we conducted. In it, the plots on the left side show the ATFOS *versus* various outbreak durations (D) for scenario 2 starting with a small outbreak at the top ($M = 9$), a medium outbreak in the middle ($M = 22.5$), and a large outbreak at the bottom ($M = 45$). The plots on the right side show the fraction of times a method missed detecting an outbreak *versus* outbreak duration. Each plot gives the results for six methods, the C1, C2, and C3, as well as three CUSUMs using various sliding baseline lengths (n values): 7, 15 (the 'optimal' for scenario 2), and 56 days.

Figure 3 shows that the C1, C2, and C3 methods do not perform as well as the CUSUM methods with the larger sliding baselines. Focusing for a moment just on the C1, C2, and C3 methods, we see that the C1 and C2 methods perform somewhat similarly, with the C1 generally having a slightly lower ATFOS compared with the C2 but at the expense of having slightly higher fraction missed than the C2. However, when comparing the C1 and C2 with the CUSUMs, we see that they all have similar ATFOS performance, but the CUSUMs with longer sliding baselines miss significantly fewer outbreaks. This difference in performance is evident for all the outbreak magnitudes but is most striking with the larger magnitude outbreaks. For example, in the middle row of plots, the C1 and C2 ATFOS can be up to a day or two shorter than the longer sliding baseline CUSUMs, but they only catch between about 25 and 35 per cent of the outbreaks, whereas the 56-day sliding baseline CUSUM catches nearly 80 per cent of the outbreaks of the longest duration. For this scenario, it is clear that the CUSUM with a 56-day sliding baseline is the preferred method.

A note about the ATFOS plots is in order for those used to looking at graphs of average run lengths in the SPC literature. Such readers may be surprised that the ATFOS curves increase as outbreak duration increases. Remember in this problem that the time to first outbreak signal is constrained to the interval $[1, D]$. That is, the earliest a 'true signal' can occur is on the first day of the outbreak and the latest is on the last day of the outbreak (D). Thus, for $D = 3$, ATFOS is constrained to be between 1 and 3 and, as we see in the plot, is about 2 for all the methods. On the other hand, for $D = 15$ the ATFOS can be much larger and, in fact, falls anywhere from about 4 days to about 7 days for the various methods.

Also in Figure 3, we see that the C1 and the CUSUM with a 7-day sliding baseline suffer from being contaminated by the outbreak data in the largest magnitude outbreak scenarios. That is, in the lower right plot we see that the fraction missed by these two methods actually *increases* for longer duration outbreaks (as eventually does the C2 and C3, as well as the CUSUM with a 15-day sliding baseline ever so slightly). If these methods fail to detect the outbreak early on, they begin to incorporate the outbreak data into their calculations (either the moving average for the C1

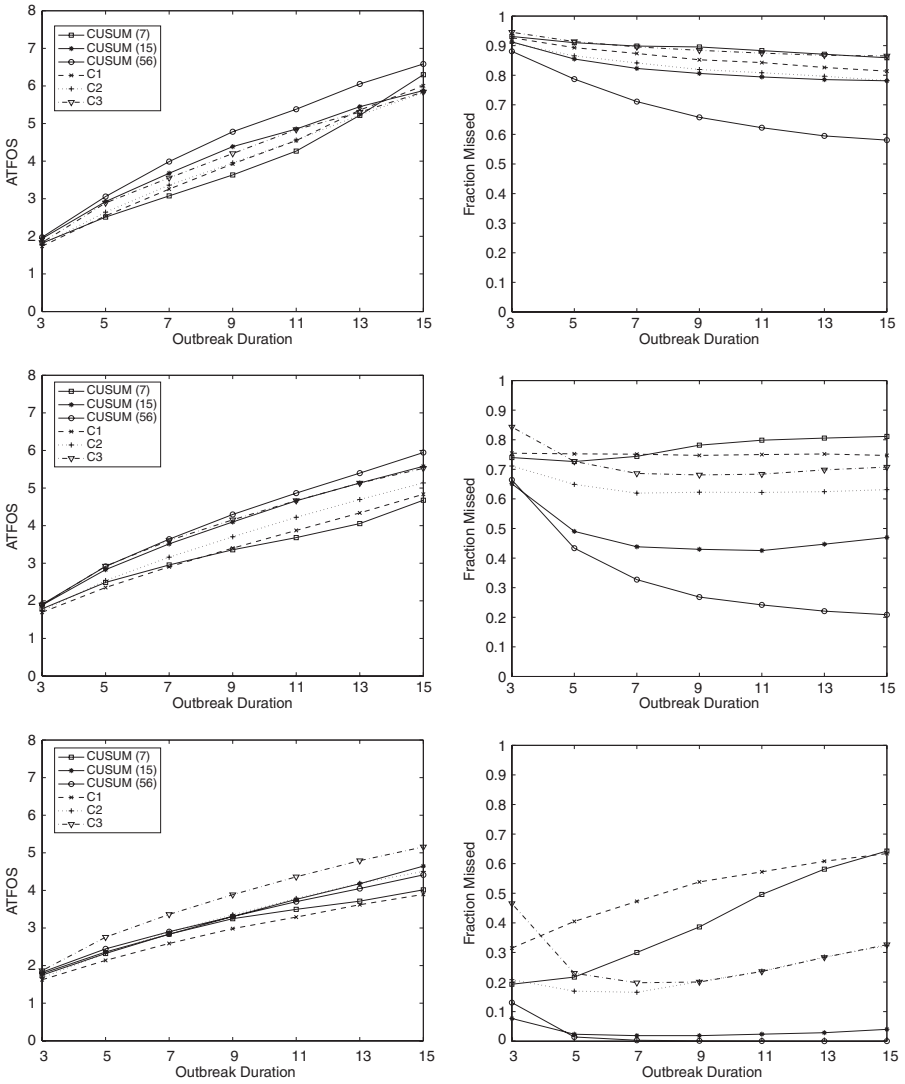

Figure 3. Performance of the methods for scenario 2 for three magnitudes of outbreaks—$M = 9$, 22.5, and 45, shown from top to bottom—*versus* various outbreak durations.

or the adaptive regression predictions for the CUSUM), making it increasingly more difficult to distinguish the outbreak from the normal background disease incidence. In comparison, the 2-day lag in the C2 method seems to be sufficient to mitigate much of this problem for that method (and the C3 that is a function of the C2 statistics).

Figure 4 shows the results for scenario 7. What is immediately striking between Figures 3 and 4 is the overall similarity of the CUSUM performance results. The CUSUMs with the longer sliding baselines are clearly the best performing methods (where note that the 'optimal' sliding baseline in

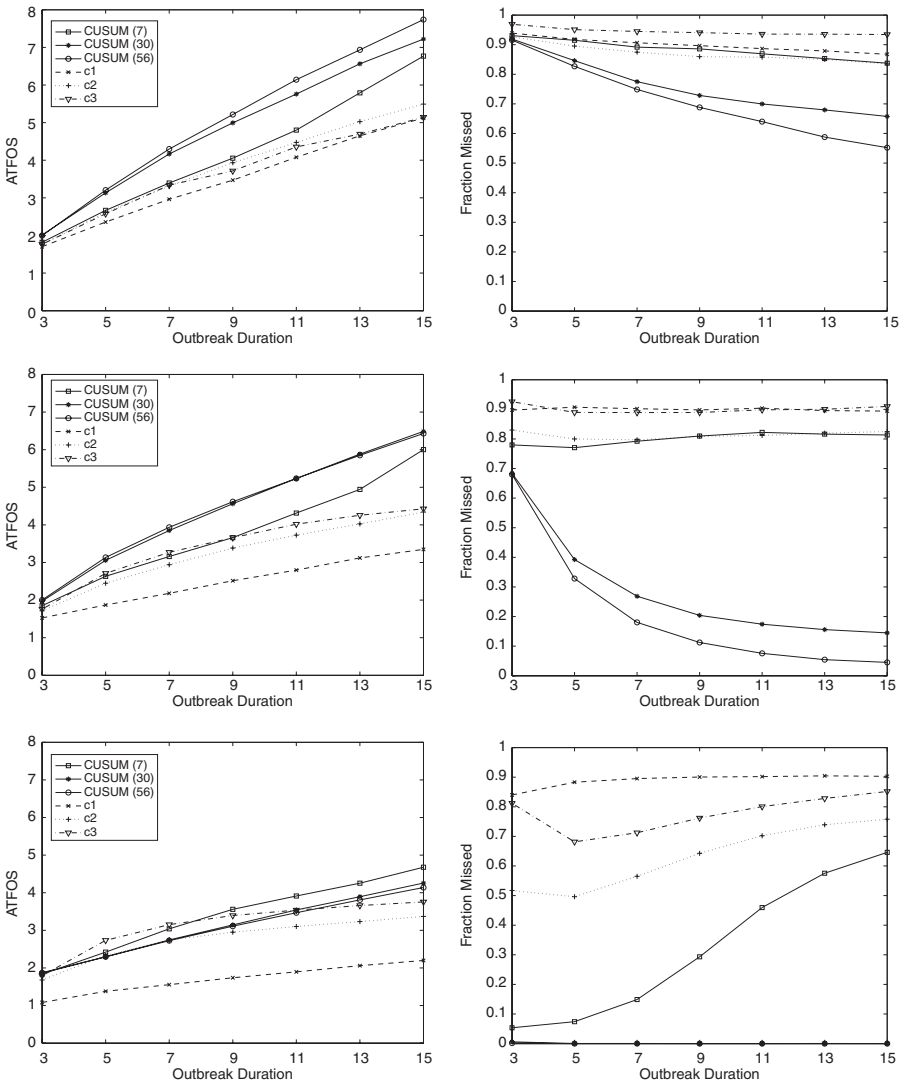

Figure 4. Performance of scenario 7 for three magnitudes of outbreaks—$M = 4$, 8, and 16, shown from top to bottom—*versus* various outbreak durations.

this scenario was 30 days, compared with 15 days for scenario 2). Within the EARS methods, the C1 method has the lowest ATFOS but misses more outbreaks than the C2, and neither performs as well as the CUSUMs with longer sliding baseline. In particular, while the C1 does have the shortest ATFOS of all the methods, it misses from 85 to 90 per cent of all the outbreaks. In comparison, in the bottom plots with the larger magnitude outbreaks, for example, the CUSUMs using either a 30- or 56-day sliding baseline catch virtually all of the outbreaks with a 2-day ATFOS for a 3-day outbreak duration up to a 4-day ATFOS for a 15-day outbreak duration.

6. DISCUSSION

On the basis of our comparisons of the EARS methods *versus* CUSUM methods applied to the residuals of adaptive regressions, we conclude that the CUSUM methods perform better. We reach this conclusion based on the fact that the EARS methods frequently failed to catch a majority of the outbreaks across a wide variety of background disease incident patterns (large and small daily counts; large, medium, small, and no seasonal cycles; large and small random daily fluctuations; with and without day-of-the-week effects) and a wide variety of outbreak magnitudes and durations. In fact, the EARS methods generally caught less than 30 per cent of the outbreaks except in the largest outbreak cases. In contrast, the CUSUM methods, particularly with the 8-week sliding baseline, performed much better.

The similarity in the performance of the long sliding baseline CUSUMs across all the various types of background disease incidence (both in the results shown here as well as the results for the other scenarios—see Dunfee and Hegler [36]) indicates that the adaptive regression methodology is effective at removing the systematic effects from the background disease incidence. (See also Fricker *et al.* [29] and Hu and Knitt [30] for other examples in which adaptive regression applied in a multivariate syndromic surveillance setting was also effective at removing systematic effects.) The longer sliding baseline, along with the linear form of the model, is also effective at ensuring that the adaptive regression does not get contaminated in longer duration outbreaks if it uses some of that data in the regression model.

These conclusions are based on extensive comparisons of the methods using simulated syndromic surveillance data that were designed to mimic the major features of a wide cross section of syndromic surveillance data. However, as we noted earlier, the simulations were purposely idealized depictions that assumed a sinusoidal shape for the annual background variation, linearly increasing and decreasing outbreaks, and particular error term distributions. In addition, our analysis used a fixed ATFS of 100 days, which we considered a reasonable false alarm rate for a syndromic surveillance system, and a particular choice for the CUSUM reference interval parameter k. We find these assumptions reasonable and fully expect our conclusions to hold for other forms of data and choices of ATFS, but confirmation of this is left to future research.

Based on this work, it is not clear whether the specific form we used for the adaptive regression will necessarily be the best choice for all types of syndromic surveillance data. In particular, we found that a linear regression model (without a quadratic term) worked best on our simulated data. While we have used this methodology on one set of actual syndromic surveillance data (see Fricker *et al.* [29]), the means of this data and the simulated data changed quite smoothly over time. For other types of data, it may be that polynomial regression, exponential smoothing, or some other model form is more effective than linear regression.

Given the performance of the CUSUM methods, particularly those using longer sliding baselines, one might be tempted to simply attribute the success to the additional information being used in the adaptive regressions. This is certainly part of the reason, but some additional preliminary simulations seem to indicate that is not the complete answer. Specifically, it seems that the Shewhart and Shewhart-based methods may be less well suited for the syndromic surveillance problem in which outbreaks do not occur instantaneously and are transient.

Figure 5 compares the performance of a CUSUM method and a Shewhart method, both applied to the residuals from an adaptive regression with a 56-day sliding baseline. Each day, the Shewhart method compares the standardized residual from the adaptive regression for that day to a threshold (chosen so that the ATFS $=100$ days; the same as with the CUSUM). In Figure 5, the top plots

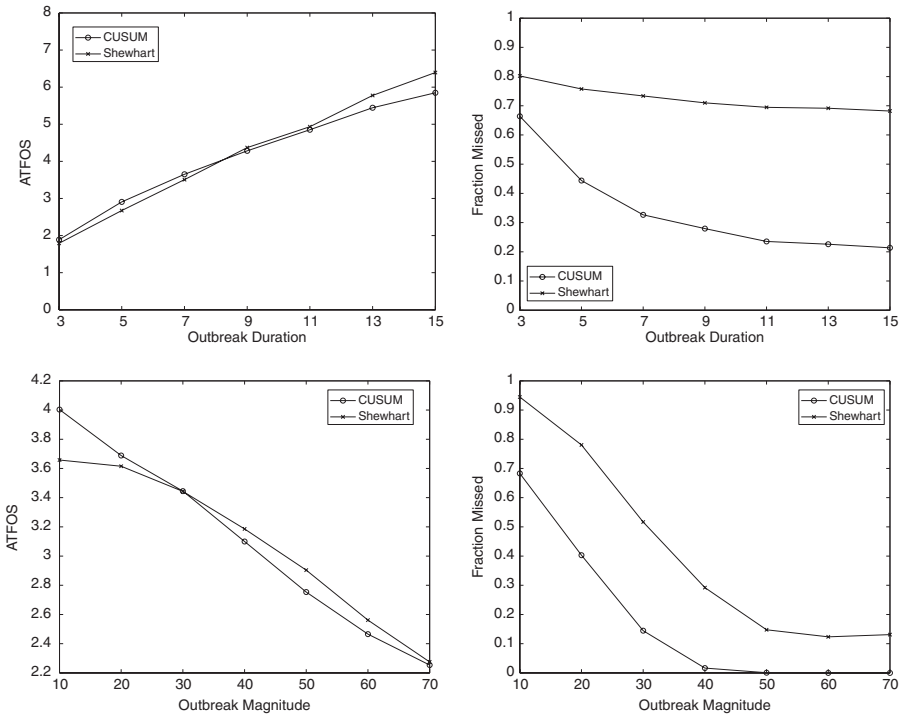

Figure 5. Performance comparison of the CUSUM *versus* the Shewhart methods applied to the residuals of an adaptive regression with an 8-week sliding baseline. The top plots compare the methods for a fixed outbreak magnitude ($M = 22.5$) for varying outbreak durations. The bottom plots compare the methods for a fixed outbreak duration ($D = 7$ days) for various outbreak magnitudes. The background disease incidence was generated via scenario 2.

compare the methods for a fixed outbreak magnitude ($M = 22.5$) for varying outbreak durations. The bottom plots compare the methods for a fixed outbreak duration ($D = 7$ days) for various outbreak magnitudes. The background disease incidence was generated via scenario 2.

In the top plots we see that the ATFOS is roughly equal between the two methods, with the Shewhart seeming to have a slight advantage for outbreaks of short duration and the CUSUM for outbreaks with long durations. This is consistent with the literature on the performance of these two methods in industrial SPC applications. However, in the upper right plot, we see that the Shewhart is much poorer at actually catching outbreaks than the CUSUM.

In the bottom two plots, rather than varying outbreak duration we vary the magnitude. That is, we fixed the duration at $D = 7$ days and compared the performance of the two methods as M varied from 10 to 70. In terms of ATFOS, we see that the Shewhart does slightly better than the CUSUM for smaller outbreaks and slightly worse for larger outbreaks. However, once again, it does significantly poorer in terms of the fraction of missed outbreaks. From this, we surmise that the poorer performance of the EARS methods is due both to the additional data used in the CUSUMs with the longer sliding baselines *and* to the Shewhart-like design of the C1 and C2 methods.

In addition, we note that in practice the thresholds for the C1, C2, and C3 are fixed quantities, which means that the average time between false alarms varies in each implementation depending on the patterns of the background disease incidence data. For this work we set the thresholds individually to achieve an ATFS of 100 days which, across all the scenarios, resulted in thresholds for the C1, C2, and C3 methods ranging from 2.7 to 18.2. This should help explain why in certain applications the methods experience so many false alarms.

In summary, the CUSUM applied to residuals from an appropriately employed adaptive regression model with an 8-week sliding baseline outperformed the EARS methods in all the scenarios we evaluated. These scenarios were chosen to mimic the major features of syndromic surveillance data over a wide variety of conditions. For standard syndromic surveillance systems using the EARS methods, this work suggests such systems would benefit from replacing the EARS methods (and the W2 in BioSense) with the CUSUM (applied to residuals from an adaptive regression) and from setting the CUSUM thresholds appropriately to minimize the false alarm burden as much as is appropriate.

Of course, the EARS methods were originally designed for a drop-in surveillance system with little or no baseline data available. In these situations, the use of an 8-week sliding baseline may be impossible, at least upon initiation of the drop-in system. However, our simulations showed that a CUSUM with a 7-day sliding baseline performed about the same as the EARS methods, and as the length of the sliding baseline increased the performance of the CUSUM quickly improved. This suggests a strategy for drop-in surveillance systems of starting with a CUSUM with a 7-day sliding baseline and, as time progresses and more data accumulates, allowing the baseline to increase until such time as enough data are accumulated so that baseline can be allowed to slide.

APPENDIX A: STANDARD DEVIATION OF THE PREDICTION ERROR

The standard deviation of the predication error for a new observation y_0 in a simple linear regression (cf. Montgomery and Peck [38]) is

$$\sigma_{\text{p.e.}} = \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{A1}$$

where

- n is the number of observations in the regression;
- x_0 is the observed value of the independent variable for the new y_0;
- S_{xx} is the sum of the squared differences of the x values in the regression from their mean; and,
- σ_ε is the standard deviation of the error term in the regression model.

Equation (A1) can be simplified since x values are sequential integers representing time relative to the current day. That is, n is the number of days we are regressing on, with yesterday being 'day n' and going back in relative time to 'day 1.' Hence, $x_0 = n+1$ and $\bar{x} = (n+1)/2$;

thus,

$$(x_0-\bar{x})^2=(n+1)^2/4 \tag{A2}$$

Similarly, we can express S_{xx} as follows:

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2=2\sum_{i=1}^{(n-1)/2}i^2$$

$$=\frac{1}{3}\left(2\left[\frac{n-1}{2}\right]^3+3\left[\frac{n-1}{2}\right]^2+\left[\frac{n-1}{2}\right]\right)$$

where the third step follows assuming n is odd and the last step follows since $\sum_{i=1}^{k}i^2=(2k^3+3k^2+k)/6$. Simplifying gives

$$S_{xx}=\frac{n(n^2-1)}{12} \tag{A3}$$

Substituting the results from (A2) and (A3) into equation (A1) and further simplifying give

$$\sigma_{\text{p.e.}}=\sigma_\varepsilon\sqrt{\frac{(n+2)(n+1)}{n(n-1)}} \tag{A4}$$

For the multiple regression with day effects, such as that shown in equation (6), solving for an expression equivalent to equation (A4) is a bit more complicated. The general expression for the standard deviation of the prediction error is

$$\sigma_{\text{p.e.}}=\sigma_\varepsilon\sqrt{1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \tag{A5}$$

where $\mathbf{x}_0$ is the next observation vector, which without loss of generality we have parameterized as $\mathbf{x}_0=\{1,n+1,0,0,0,0,0,0\}$, and $\mathbf{X}$ is the $n\times 8$ matrix of historical data:

$$\mathbf{X}=\begin{pmatrix}1 & n & 1 & 0 & 0 & 0 & 0 & 0\\ 1 & n-1 & 0 & 1 & 0 & 0 & 0 & 0\\ 1 & n-2 & 0 & 0 & 1 & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & 3 & 0 & 0 & 0 & 0 & 1 & 0\\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 1\\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}$$

Note that, to simplify the calculations of $\mathbf{X}'\mathbf{X}$, we fixed n as a multiple of 7 in the $\mathbf{X}$ matrix above. Given this, and after a bit of algebraic manipulation, we have

$$\mathbf{X}'\mathbf{X}=\begin{pmatrix} n & n(n+1)/2 & n/7 & n/7 & n/7 & n/7 & n/7 & n/7 \\ n(n+1)/2 & \frac{n(n+1)(2n+1)}{6} & \frac{n(n+7)}{14} & \frac{n(n+5)}{14} & \frac{n(n+3)}{14} & \frac{n(n+1)}{14} & \frac{n(n-1)}{14} & \frac{n(n-3)}{14} \\ n/7 & \frac{n(n+7)}{14} & n/7 & 0 & 0 & 0 & 0 & 0 \\ n/7 & \frac{n(n+5)}{14} & 0 & n/7 & 0 & 0 & 0 & 0 \\ n/7 & \frac{n(n+3)}{14} & 0 & 0 & n/7 & 0 & 0 & 0 \\ n/7 & \frac{n(n+1)}{14} & 0 & 0 & 0 & n/7 & 0 & 0 \\ n/7 & \frac{n(n-1)}{14} & 0 & 0 & 0 & 0 & n/7 & 0 \\ n/7 & \frac{n(n-3)}{14} & 0 & 0 & 0 & 0 & 0 & n/7 \end{pmatrix}$$

After inverting this matrix, substituting the results into equation (A5), and simplifying we ultimately obtain

$$\sigma_{\text{p.e.}}=\sigma_{\varepsilon}\sqrt{\frac{n^2+3n-28}{n(n-7)}} \tag{A6}$$

Figure A1 compares the results of equations (A4) and (A6) for various n, where the 'multiplicative factor' on the vertical axis refers to the square root terms in the equations. One half of this factor would be the precise value to set k in the CUSUM assuming it is important to quickly detect an increase in the mean of one standard deviation of the prediction error. For example, in Figure A1 if we were to use a sliding baseline of $n=14$, then with a simple linear regression we would set $k=0.58$ and for a multiple regression incorporating day effects we would set $k=0.72$. In comparison, with a sliding baseline of $n=56$, we would set $k=0.52$ for a simple linear regression and $k=0.55$ for a multiple regression incorporating day effects. Of course, the choice of designing

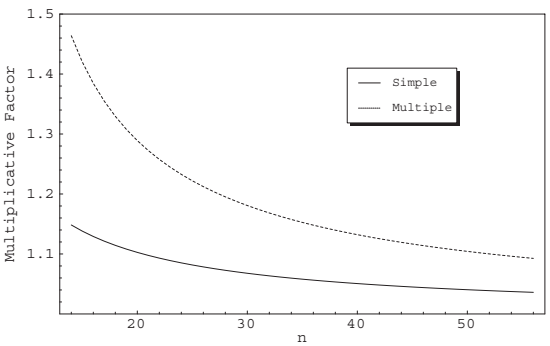

Figure A1. A comparison of the square root terms in equations (A4) and (A6) for various n, $14\leqslant n\leqslant 56$. The solid line is the result for a simple linear regression ('simple' in the key) and the dotted line for a multiple regression with six indicator variables to model day-of-the-week effects ('multiple' in the key).

the CUSUM to detect an increase of one standard deviation of the prediction error is subjective, as was our choice to set $k=\delta/2$.

For the sake of simplicity in the simulations, we fixed $k=0.65$ for $n=7$ and $k=0.5$ for $n>7$. The net effect of this was to make the CUSUM based on adaptive regressions with either shorter sliding baselines or that modeled day effects slightly more sensitive to smaller shifts in the mean. However, given that the better performing CUSUMs had $n=56$, the net effect of these simplifications on the conclusions is negligible. For an actual application of the methodology, on the other hand, a practitioner would wish to give some careful thought to the desired size of an outbreak to detect (in terms of the prediction error standard deviations) in order to optimize the performance of the CUSUM in their application.

ACKNOWLEDGEMENTS

We would like to thank Bill Woodall, Jerry Tokars, Lori Hutwagner, two anonymous reviewers, and an associate editor for their comments on earlier drafts. Their suggestions resulted in substantial improvements to this paper. R. Fricker's work on this effort was partially supported by Office of Naval Research grant N0001407WR20172.

REFERENCES

1. Centers for Disease Control and Surveillance. *Syndromic Surveillance*: *Reports from a National Conference*, *2003*. *Morbidity and Morality Weekly Report*, vol. 53 (Supplement). Centers for Disease Control and Surveillance, 2004.
2. International Foundation For Functional Gastrointestinal Disorders. www.iffgd.org/GIDisorders/glossary.html. (accessed on 21 November 2006.)
3. Centers for Disease Control and Prevention. *Syndrome Definitions for Diseases Associated with Critical Bioterrorism-associated Agents*. Centers for Disease Control and Prevention, 23 October 2003. (Available from: http://www.bt.cdc.gov/surveillance/syndromedef/, accessed on 21 November 2006.)
4. Henning KJ. Overview of syndromic surveillance: what is syndromic surveillance? *Morbidity and Mortality Weekly Report*, vol. 53 (Supplement). Centers for Disease Control and Prevention, 2004; 5–11. (Available from: www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm.)
5. Fricker Jr RD. Syndromic surveillance. *Encyclopedia of Quantitative Risk Assessment* 2007; in press.
6. Fricker Jr RD. Directionally sensitive multivariate statistical process control methods with application to syndromic surveillance. *Advances in Disease Surveillance* 2007; **3**:1. (Available from: www.isdsjournal.org.)
7. Fricker Jr RD, Rolka H. Protecting against biological terrorism: statistical issues in electronic biosurveillance. *Chance* 2006; **91**:4–13.
8. Stoto MA, Fricker Jr RD, Jain A, Diamond A, Davies-Cole JO, Glymph C, Kidane G, Lum G, Jones L, Dehan K, Yuan C. Evaluating statistical methods for syndromic surveillance. In *Statistical Methods in Counterterrorism*: *Game Theory*, *Modeling*, *Syndromic Surveillance*, *and Biometric Authentication*, Wilson A, Wilson G, Olwell D (eds). Springer: New York, 2006.
9. Centers for Disease Control and Surveillance. *Early Aberration Reporting System*. Centers for Disease Control and Surveillance, 2007. (Available from: www.bt.cdc.gov/surveillance/ears, accessed on 30 April 2007)
10. Centers for Disease Control and Surveillance. *BioSense*. Centers for Disease Control and Surveillance, 2007. (Available from: www.cdc.gov/biosense, accessed on 30 April 2007.).
11. Burkom HS, Murphy SP, Shmueli G. Automated time series forecasting for biosurveillance. *Statistics in Medicine* 2007; **26**(22):4202–4218.
12. Shmueli G. Statistical challenges in modern biosurveillance. *Technometrics* 2006; submitted.
13. Woodall WH. The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology* 2006; **38**:1–16.
14. Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health*: *Bulletin of the New York Academy of Medicine* 2003; **80**:89i–96i.

15. Zhu Y, Wang W, Atrubin D, Wu Y. Initial evaluation of the early aberration reporting system—Florida. *Morbidity and Mortality Weekly Report*, vol. 54 (Supplement). Centers for Disease Control and Prevention, 2005; 123–130.
16. Shewhart WA. *Economic Control of Quality of Manufactured Product*. van Nostrand: Princeton, NJ, 1931.
17. Montgomery DC. *Introduction to Statistical Quality Control* (4th edn). Wiley: New York, 2001.
18. Centers for Disease Control and Surveillance. *BioSense Bulletin*. Centers for Disease Control and Surveillance, September 2006.
19. Page ES. Continuous inspection schemes. *Biometrika* 1954; **41**:100–115.
20. Lorden G. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics* 1971; **42**: 1897–1908.
21. Hawkins DM, Olwell DH. *Cumulative Sum Charts and Charting for Quality Improvement*. Springer: Berlin, 1998.
22. Moustakides GV. Optimal stopping times for detecting a change in distribution. *Annals of Statistics* 1986; **14**:1379–1388.
23. Ritov Y. Decision theoretic optimality of the CUSUM procedure. *Annals of Statistics* 1990; **18**:1464–1469.
24. Brillman JC, Burr T, Forslund D, Joyce E, Picard R, Umland E. Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. *BMC Medical Informatics and Decision Making* 2005; **5**.
25. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society*, *Series A* (*Statistics in Society*) 1996; **159**:547–563.
26. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. *BMC Medical Informatics for Decision Making* 2003; **3**.
27. Lotze T, Murphy SP, Shmueli G. Preparing biosurveillance data for classic monitoring. *Advances in Disease Surveillance* 2006; submitted.
28. Hutwagner L. *Personal communication*. 12 December 2006.
29. Fricker Jr RD, Knitt MC, Hu CX. Comparing directionally sensitive MCUSUM and MEWMA procedures with application to biosurveillance. *Quality Engineering* 2007; submitted.
30. Hu CX, Knitt MC. A comparative analysis of multivariate statistical detection methods applied to syndromic surveillance. *Master's Thesis*, Naval Postgraduate School, Monterey, CA, 2007.
31. Kleinman KP, Abrams A, Mandl KD, Platt R. Simulation for assessing statistical methods of biologic terrorism surveillance. *Morbidity and Mortality Weekly Report*, vol. 54 (Supplement). Centers for Disease Control and Prevention, 2005. (Accessed from: http://dacppages.pbwiki.com/f/mmwr2005.pdf, accessed on 12 May 2007.)
32. Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences USA* 2003; **100**:1961–1965.
33. Fraker SE, Woodall WH, Mousavi S. Recurrence interval and time-to-signal properties of surveillance schemes. *Quality Engineering* 2007.
34. Reynolds Jr MR. Approximations to the average run length in cumulative sum control charts. *Technometrics* 1975; **17**:65–71.
35. Siegmund D. *Sequential Analysis Tests and Confidence Intervals*. Springer: New York, 1985.
36. Dunfee DA, Hegler BL. Biological terrorism preparedness: evaluating the performance of the early aberration reporting system (EARS) syndromic surveillance algorithms. *Master's Thesis*, Naval Postgraduate School, Monterey, CA, 2007.
37. Chang JT, Fricker Jr RD. Detecting when a monotonically increasing mean has crossed a threshold. *Journal of Quality Technology* 1999; **31**:217–233.
38. Montgomery DC, Peck EA. *Introduction to Linear Regression Analysis* (2nd edn). Wiley: New York, 1992.