

# Assessing surveillance using sensitivity, specificity and timeliness

**Ken P Kleinman and Allyson M Abrams** Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, Boston, MA, USA

Monitoring ongoing processes of illness to detect sudden changes is an important aspect of practical epidemiology and medicine more generally. Most commonly, the monitoring has been restricted to a unidimensional stream of data over time. In such situations, analytic results from the industrial process monitoring have suggested optimal approaches to monitor the data streams. Data streams including spatial location as well as temporal sequence are becoming available. Monitoring methods that incorporate spatial data may prove superior to those that ignore it. However, analytically, optimal methods for spatial surveillance data may not exist. In the present article, we introduce and discuss evaluation metrics that can be used to compare the performance of statistical methods of surveillance. Our general approach is to generalize receiver operating characteristic (ROC) curves to incorporate the time of detection in addition to the usual test characteristics of sensitivity and specificity. In addition to weighting ordinary ROC curves by two measures of timeliness, we describe three three-dimensional generalizations of ROC curves that result in timeliness-ROC surfaces. Working in the context of surveillance of cases of disease to detect a sudden outbreak, we demonstrate these in an artificial example and in a previously described simulation context and show how the metrics differ. We also discuss the differences and under which circumstances one might prefer a given method.

## 1 Preface

The bioterrorism attacks using anthrax spores in 2001 have generated a greatly increased interest in systems that are intended to detect such attacks as early as possible.<sup>1–13</sup> The rationale for this interest is that with early detection the morbidity and mortality of bioterrorist attacks may be mitigated. Diseases deriving from communicable agents, such as smallpox, can possibly be contained in a small region, whereas diseases caused by non-communicable agents, such as anthrax, may more likely be treated during the limited windows in which treatment is likely to be successful.<sup>14</sup>

One important possibility for early detection is the monitoring of health care encounters such as outpatient visits, emergency department visits or phone calls to physicians' offices. The value of this approach is due to the fact that so many likely bioterrorist agents result in diseases with an indistinct prodrome, causing illness that initially resembles naturally occurring disease. In these important cases, the first sign of disease may

---

Address for correspondence: Ken P Kleinman, Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, 133 Brookline Avenue, Boston, MA 02215, USA. E-mail: ken.kleinman@gmail.com

well be an increased number of health care encounters, without an obvious sign of a terrorist agent.

In this context, improved data management and computerization of health records have allowed the development of systems that record not only the fact that cases with certain indistinct symptoms have been seen but also some locational data about them – at least a postal code but in many cases also a latitude and longitude from a geocoded home address. Using these data may improve the sensitivity of surveillance relative to using the simple count of data summarized across a region. Possibly due to a dearth of such data or to smaller perceived surveillance needs in the past, this area of statistics is underdeveloped. Thus, as systems to gather data have been created, statisticians have often developed *de novo* statistical techniques to detect anomalies and signal changes that might herald attacks.

## 2 Introduction

One important job of public health authorities is to monitor public health data to detect changes in the public's health. Typically, the data so monitored simply show an overall count of the number of people affected, for example, by influenza.<sup>15</sup> However, in an increasing number of applications, the location at which each case was observed is available, in addition to the fact that the case was observed. Surveillance using statistical tools that incorporate spatial data hold the promise of improved sensitivity, timeliness of detection, and possibly specificity.<sup>13,16,17</sup> However, there is a lack of metrics to define and to comprehensively measure the performance of these techniques, despite some attempts.<sup>18</sup>

When spatial data are not available, surveillance often utilizes methods such as Shewhart control charts and related cumulative sum methods that we generically refer to as time-series methods.<sup>6,19,20</sup> Also, in these cases, statistical developments from industrial process control applications have allowed the analytic evaluation of surveillance methods.<sup>21</sup> However, these developments typically rely on assumptions that are too restrictive in the context of public health surveillance. For example, they may require surveillance designed to detect the change from a constant mean to a new constant mean.<sup>22</sup> In contrast, public health surveillance situations are more complex and may violate these assumptions in many ways. For example, the means of public health processes are rarely constant, but depend on features such as season, day of week or even, for example, the pollen count or the pollution level. In these circumstances, the analytical results obtained in simple situations may not apply. Given the complicated environment of public health practice, the simulation of public health events may be the most useful context to evaluate methods.

An alternate to the analytic approach is to use receiver operating characteristic (ROC) curves in specific applications. In assessing the fit of logistic regressions, the ROC curve plots the sensitivity versus 1 less the specificity for a series of decision points based on the predicted probability that results from inverting the predicted logit for each subject. The ROC curve increases monotonically. The closer the curve runs to the point with specificity and sensitivity equal to 1, the better the fit of the model. In addition, the area under the ROC curve (AUROC) is approximately the probability that a true positive will

have a higher probability of being designated a case by the model than a true negative.<sup>23</sup> In the current context, the ‘model’ is a statistical technique for surveillance and the goal is to detect events of interest to the surveillance, as opposed to predicting a case in a logistic regression.

The AUROC is a promising metric for evaluating surveillance, in that it is geared to describe the performance of methods in the context of a specific example data set. However, the ROC curve ignores the crucial dimension of time: the purpose of surveillance is to identify aberrations as soon as possible. The purpose of the present article is to generalize the AUROC to incorporate the time of detection.

In the remainder of the article, in the next section, we present a description of terms and notation, some concerns relevant to simulations and spatial analysis, some proposed metrics, and a description of a simulation context in which the metrics will be applied. In the results section, we show the results of applying the metrics to an artificial data set and to simulated data. In the discussion, we consider the results and criteria by which one might choose the most sensible metric, as well as limitations and directions for future work.

### 3 Methods

We begin with an explicit discussion of terms and notation to be used in the following discussion. A statistical method is considered to be monitoring some stream of data. In each of  $N$  equal-duration time periods, the method generates a  $P$ -value, posterior probability or other numerical assessment of the data with a monotonic relationship to the method-assessed unusualness of the current state of the data stream. In the following, we refer to  $P$ -values, with the smaller  $P$ -values suggesting increasingly unusual observations.

If the  $P$ -value exceeds some threshold, the method has generated a signal. The method includes spatial descriptors in its assessment of the data so that the signal also has a spatial *extent*. In parallel, there may be a temporal duration of the signal; the duration is always assumed to be a contiguous period of time including the time of assessment. The possibility of signal duration is necessary to allow novel signals that include time periods for which no signal was previously generated. The data may contain real or simulated health *events* such as an outbreak of influenza or smallpox; the events may have spatial extent and temporal duration as well.

Throughout the article, we denote thresholds as  $r$ . In some cases, the threshold may be a function of time or other characteristics. If so, we assume that there is some key parameter in that function, which can be changed continuously; this will allow the methods in the following discussion to be applied without loss of generality. Alternatively, if the threshold is discrete, modifications of the following discussion will be needed. If the function depends on multiple covariates, its performance may have to be assessed at a range of fixed values for all but one parameter.

If  $P < r$ , the method has generated a signal at threshold  $r$ . We arbitrarily label the  $M_r$  signals generated at threshold  $r$  as  $s_{m,r}$ ,  $m = 1 \cdots M_r$ . The time at which the  $m$ th signal under threshold  $r$  is generated is referred to as  $T(s_{m,r})$ . The spatial extent of signal  $s_{m,r}$  is denoted  $\text{ex}(s_{m,r})$  and its duration as  $d(s_{m,r})$ . Note that  $\text{ex}(s_{m,r})$  is a generic descriptor

including such specific geometric descriptors as a center and radius, a simple list of included regions or points or other more complicated definitions of a region in space. Additionally, duration is a scalar measuring the number of time periods into the past which are included in the signal.

The  $V$  events are labeled  $e_1, \dots, e_V$ . The time at which event  $e_v$  began is  $T(e_v)$ . The extent and duration of events are  $\text{ex}(e_v)$  and  $d(e_v)$ , respectively;  $d(e_v)$  is measured as the time between the beginning of the event and the time of the last case caused by the event; note that this is time measure forward, as opposed to  $d(s_{m,r})$ . The extent  $\text{ex}(e_v)$  is defined as the region that contains cases caused by the event. The extent of an event may vary with each time period after the event; to ease the development below, we assume instead that the extent is fixed and note that the extension to variable extent is trivial.

In purely temporal data, a signal may be considered a true positive if its duration overlaps with that of an event. The spatial surveillance setting requires consideration of whether the extents of the signal and the event overlap as well. We define a ‘hit’ for a given signal as the condition that the signal overlaps both spatially and temporally with at least one event, even though other definitions are defensible. Symbolically, signal  $s_{m,r}$  is a hit for event  $e_v$  if

$$\begin{aligned} & ([T(s_{m,r}) - d(s_{m,r}), T(s_{m,r})] \cap [T(e_v), T(e_v) + d(e_v)]) \neq \emptyset \\ & \text{and } (\text{ex}(s_{m,r}) \cap \text{ex}(e_v)) \neq \emptyset \end{aligned}$$

using the standard set and logic notation for ‘intersect’ ( $\cap$ ) and the empty set ( $\emptyset$ ) and once again noting that the expression is far from being formally precise, particularly with respect to the description of space. Let  $H(s_{m,r}) = 1$  if the signal is a hit for some event  $e_v$  and 0 if not. We similarly define  $H(e_v, r) = 1$  when there is some signal that is a hit for event  $e_v$  at threshold  $r$  and  $H(e_v, r) = 0$  when no signal is a hit for that event at that threshold. If evaluation of a time-series method within this framework is desired,  $\text{ex}(s_{m,r})$  can be defined as the universe of locations in the surveillance area for all signals.

### 3.1 Simulated events and conditional ROC curves

Owing to the fortunate dearth of bioterrorist attacks, there are few real data sets to assess the performance of this important special case. Data that do exist more commonly contain events of a more mundane and less catastrophic nature.<sup>24,25</sup> Good performance of a method in these data sets may not imply good performance in quickly detecting large bioterrorist-generated events or, indeed, in detecting other types of mundane events. In addition, the heterogeneous nature of the events means that it is not possible to make general statements about what kinds of events a method detects well on the basis of real data. Thus evaluation may be best performed in simulated data sets.

When spatial information is available, the simulation of data without an event is complicated by the complex nature of spatial data. Even in the most accurate simulations, simulated non-event data may bear so little resemblance to real data streams as to render evaluation after simulating an event useless. This observation has led to the middle path of simulating only events and adding the resulting simulated cases to the real data stream.<sup>26,27</sup> This approach is sometimes referred to as ‘injecting’ or ‘spiking’ events into real data streams. False positive signals are those generated in the data before any events are injected, whereas true positives are based on injected events.

However, when assessment via ROC-like methods is contemplated, this shortcut introduces complications. Primarily, as there is only one data set that has no events, the estimation of the false positive rate and hence the specificity can only be done on that original data set. In contrast, the true positive rate and hence the sensitivity can be estimated with arbitrary sample size by increasing the number of simulated events. This means that many results pertaining to ROC curves and the AUROC may not apply. In addition, for all evaluations based on injected events, there is an often-unexpressed assumption that no true (unsimulated) events occurred in the real data stream. This is a quite reasonable assumption if the event in question is a smallpox or anthrax attack and much less so if the event is a naturally occurring outbreak of gastro-intestinal illness. To reflect these special circumstances, we refer to ROC curves based on injected events as ‘conditional ROC’ curves,<sup>26</sup> meaning conditional on there being no relevant real events in the real data and also implying the different sources of information regarding the sensitivity and specificity. The evaluation metrics discussed subsequently can be applied to real data sets, to fully simulated examples or to injected simulated events, but the examples will be for injected events, and thus are based on conditional ROC curves.

### 3.2 Proposed metrics

We define the true positive rate, or sensitivity, as  $TP(r) = \sum_{v=1}^V H(e_{v,r})/V$ , the proportion of events hit by the method at threshold  $r$ . The false positive rate, 1 less the specificity, is  $FP(r) = M_r/N$ , assessed in the  $N$  time periods in the original data, that is, without added simulated cases.

Three features of these definitions attract special attention. First, note that the cases deriving from the simulated events are removed before cases deriving from new simulated events are added. Secondly, due to the geographic requirements of a hit, the true positive rate need not be 1 when the threshold is at its maximum. If injected events are not used, the definition of  $TP(r)$  is unchanged, whereas  $FP(r)$  would be defined as the proportion of time periods with signals that are not hits. Finally, the fundamental unit of the two test characteristics is different: sensitivity is assessed per event, over a series of time periods, whereas specificity is assessed per time period. The last point is another reason that usual results pertaining to ROC curves may not be applicable here.

There are two ways to view timeliness. In the first case, the timeliness is measured only with respect to the event. In this case, the timeliness is defined as the difference between the time of the event and the time of the signal which hits it. Defining the time difference function  $TD(a, b) = T(a) - T(b)$ , this is  $TD(s_{m,r}, e_v)$ , where  $s_{m,r}$  is a signal which is a hit for  $e_v$ . Small values of  $TD(s_{m,r}, e_v)$  are desirable, as they indicate early signals. Note that multiple signals may be hits for event  $e_v$ . If this is so, the timeliness  $TD(s_{m,r}, e_v)$  is the minimum across the signals and the signal generating this minimum timeliness is referred throughout as the signal that hits event  $e_v$ . As  $r$  increases,  $T(s_{m,r}, e_v)$  may decrease but never increase.

In the other case, timeliness is considered with respect to some reference signal for a given event. This could be a constant period of time after a signal or a simulated competing detection method, for example, a gold-standard laboratory confirmation of a bioterrorist agent.<sup>26,27</sup> In this case, the timeliness for a signal that is a hit is defined

as the difference in time between the signal generated by the method and the referent signal. This timeliness is denoted as  $TD(ref_v, s_{m,r}) = T(ref_v) - T(s_{m,r})$ , where  $T(ref_v)$  is the time of the reference signal for event  $e_v$ , and smaller values now indicate reduced advantage over the referent. When  $H(e_{v,r}) = 0$  or  $TD(ref_v, s_{m,r}) < 0$ , we define  $TD(ref_v, s_{m,r}) = 0$ . Whether or not a reference is used, if there are multiple signals that are hits for event  $e_v$ , the one with the minimum  $T(s_{m,r})$  is chosen; this reflects the earliest signal of the event. In most real applications and realistic simulations, some sort of reference signal will exist.

### 3.3 ROC curve

The ordinary ROC curve is a curve in the unit square with  $1 - \text{specificity}$  on the horizontal axis and sensitivity on the vertical axis. The points on the curve are determined by finding  $FP(r)$  and  $TP(r)$  for all thresholds  $r$ . This results in a step function, with steps occurring only at each observed  $P$ -value where additional events are hit. Alternatively,  $FP(r)$  and  $TP(r)$  are only calculated at these observed  $P$ -values, and these points are connected with the points at  $FP(r) = 0$  and  $TP(r) = 0$  (at the theoretical minimum  $r$ ) and  $FP(r) = 1$  and  $TP(r) = 1$  (at the theoretical maximum  $r$ ). This version approximates the conceptual smoothness of the curve. The AUROC is calculated as the area in the unit square underneath one or the other construction of the curve. The ROC curve and the AUROC are of limited utility in this context, as discussed earlier, because they do not incorporate timeliness.

### 3.4 Proposed metrics: weighted ROC curves

One approach to incorporating timeliness would be to simply weight each point on the ROC curve by the mean or median timeliness associated with the corresponding threshold. If no referent signal is available, this approach would be misleading, as there would be no penalty for the time elapsed when  $H(e_{v,r}) = 0$ , that is, when the method fails to detect events. For applications with a reference signal, this is not a problem. However, in either case, the weighted values would not be constrained to the (1,1) square and would be difficult to compare across evaluation settings or interpret in an absolute sense.

In order to ensure that all weighted values lie between 0 and 1, a generalizable scale is needed. Assuming a reference signal is available, the proportion time saved is defined as  $TS(e_v, s_{m,r}) = TD(ref_v, s_{m,r}) / TD(ref_v, e_v)$ , the proportion of the time the method generated a signal before the reference signal. This is 0 whenever there is no signal of the event or the signal occurs after the reference signal and is 1 when the signal occurs at the time of the event. If  $T(ref_v) = T(e_v)$ , define  $TS(e_v, s_{m,r}) = 0$ . Then the mean or median time saved across simulated events would be an appropriate weight for the ROC curve, with, for example,  $\overline{TS}(e_v, s_{m,r}) \cdot TP(r)$  plotted against  $FP(r)$ . This weighted ROC, which we refer to a WROC1, has the property in common with the ordinary ROC that a ‘perfect’ method – one which signals all events at the time they occur, for every threshold – will contain an area (AUWROC1) of 1, unless  $T(ref_v) = T(e_v)$  for some event  $v$ .

WROC1, however, uses a very simplistic approach to the timeliness, for which the central tendency may not be an important feature in evaluation. Instead, for a given  $r$ , we can calculate a timeliness weight in a fashion similar to the AUROC. To do this, the unit square with 1 less proportion time saved on the horizontal axis and cumulative

proportion detected on the vertical axis is considered. The points measure, for a given 1 less the proportion time saved, the proportion of events detected within that timeliness, that is,  $y = \sum_{v=1}^V H(e_v) \cdot I(\text{TS}(e_v, s_{m,r}) > x) / N$ , where  $x$  is the abscissa,  $y$  is the ordinate, and  $I(\cdot)$  is 1 when the condition is true and 0 otherwise. As with the AUROC, the points could be computed for every proportion-time-saved value, making a step function, or only calculated at the values when the step rises, meaning when an additional event was detected at less time saved. The area under this curve, denoted  $\text{TW}(r)$ , is the timeliness weight proposed to make WROC2, which plots  $\text{TW}(r) \cdot \text{TP}(r)$  against  $\text{FP}(r)$ . For a perfect method, the area under it, AUWROC2, will be 1.

### 3.5 Proposed metrics: three-dimensional generalizations of the ROC curve

Both WROC1 and WROC2 can easily be converted to three-dimensional versions of the ROC curve by making the mean proportion time saved (WROC1) or the timeliness weight (WROC2) the  $z$ -axis and plotting  $\text{TP}(r)$ ,  $\text{FP}(r)$  and, for example,  $\text{TW}(r)$  for each  $r$ . Then, connecting the points in sequence on increasing  $r$  and drawing a line perpendicular to the  $x, y$ -plane from each plotted point to the  $z = 0$  plane results in a curtain. Finally, connecting each point to the  $(\text{TP} = 0, \text{FP} = 1, \text{TW} = 1)$  corner creates a tent-like surface, which we define as the timeliness-receiver operating characteristic surface (TROS). The surface based on the simple statistic from the timeliness distribution is labeled TROS1. The surface based on the timeliness weight is TROS2. The volumes under each surface are labeled VUTROS1 and VUTROS2, respectively. For a perfect method, the value of both VUTROS1 and VUTROS2 is 1.

A different generalization to three dimensions (TROS3) relies on multiple calculations of a modified ROC curve. For each proportion time saved, we will construct an ROC curve assessing whether the method signaled hits saving at least that proportion of the time. To do this, we begin with the definition of a hit saving at least  $tsp$  of the proportion of time saved as  $H(s_{m,r}, tsp) = H(s_{m,r}) \cdot I(\text{TD}(s_{m,r}, e_v) \leq tsp)$  and the true positive rate as  $\text{TP}(r, tsp) = \sum_{v=1}^V H(e_{v,r}, tsp) / N$  where  $H(e_{v,r}, tsp) = I(\sum_{i=1}^m H(s_{m,r}, tsp) > 0)$ . (A similar construction could be used to measure absolute time, if a fixed referent was used.) With this construction, a new ROC curve,  $\text{ROC}(tsp)$ , is obvious. The  $\text{ROC}(tsp)$  could be calculated only at each of the observed proportion-time-saved values,  $\text{TS}(e_v, s_{m,r})$ , or across values from 0 to 1. As  $tsp$  increases, for any threshold  $r$ , the specificity remains constant and the sensitivity can only increase. To construct the surface, let the  $x$ -axis index 1 less the specificity, the  $y$ -axis the proportion time saved, and the  $z$ -axis the sensitivity. The volume under TROS3 (VUTROS3) will be 1 for a perfect method.

### 3.6 Evaluation methods

One interesting aspect of the metrics described earlier is that there is no obvious way to choose among them. In fact, there is a need for a metric to evaluate evaluation metrics. In the absence of this meta-metric, one way to assess the relative performance of metrics is to show their discriminative power in a practical setting.

We have developed a simulation of how victims of a bioterrorist attack employing anthrax would appear in a surveillance system currently operating in Eastern

Massachusetts.<sup>8,26,28</sup> In brief, the simulation chooses a random spot in either an urban or a suburban area, from which 1 kg of anthrax spores might be dropped from a crop-dusting plane. One of two spore-dispersion functions defines how many spores fall at a given spot on the map.<sup>29,30</sup> One of five probabilities per spore is chosen, from which each individual person in the surveillance is calculated a probability of being a victim; this is simulated for each person.

Victim's symptoms onset time is simulated from a log-normal distribution.<sup>31</sup> Certain persons in the surveillance area are chosen *a priori* to be of the type that will enter the surveillance data set if they become ill. For these victims, a separate simulation of the time from symptom onset until presentation in the system occurs.

The system is based on visits to primary care offices. We use a fixed reference of nine days, meaning that the reference signal to which we will compare the statistical surveillance always takes place on the ninth day after the simulated attack. By this time, it seems likely that some astute physician would make a gold-standard diagnosis of anthrax.

For the purposes of assessing the sensitivity of the proposed metrics, we simulated anthrax attacks in the urban region, with the more concentrated of the two spore distribution functions. Seven methods were used to generate signals; it being beside the point here, we provide no details beyond the names of the methods and relevant references. Three methods were space-time scan statistic approaches with maximum durations of one, three, or seven days, including covariate and spatial non-uniformity adjustment via a mixed effects model.<sup>32-34</sup> Three used a Poisson generalized linear mixed effects model (GLMM)<sup>16</sup> with a fixed window of one, three or seven days.<sup>34</sup> The final method was a time-series approach, not unlike one recently proposed by Reis and Mandl.<sup>19</sup>

## 4 Results

To concretize the abstract discussion, we present a simple manufactured example before proceeding to the results of application in the simulation setting discussed earlier.

### 4.1 Application to artificial data

The setting of the manufactured example is conceived as the surveillance over 20 time periods, in a setting with spatial information containing only region identifiers. Examples of data that might be generated, as well as the accompanying metrics, are presented in Tables 1 and 2. In Table 1, we show the results of the method, meaning the  $P$ -value observed and the descriptors that might result when applying a method. Before an event is injected, there are four  $P$ -values less than 0.99, implying five different false positive rates, depending on the threshold. The remaining 16  $P$ -values are all equal to 0.99. Table 1 also shows what would happen after an event was added, and re-running the method generated a different  $P$ -value during one time period which was a hit when the threshold was large enough. The true positive rate is 0 for thresholds below that  $P$ -value and 1 for larger thresholds. In Table 1, we also show the false positive rate as a function of threshold; this depends only on the  $P$ -values observed before the injection of the event.

**Table 1** Example of an artificial data set with 20 time periods, showing the signals present before any events are injected and an additional result after injecting an event; *P*-values are hypothetical and not based on data analysis

Result of using the method in the original data set (before event was inserted); extent and duration of signals are suppressed for $r > 0.01$							
Time	$p_t^a$	$ex(s_{m,0.01})$	$d(s_{m,0.01})$	$h(s_{m,0.01})$	$h(s_{m,0.2})$	$h(s_{m,0.5})$	$h(s_{m,0.8})$
1	0.2				0	0	0
13	0.0003	Region A	3	0	0	0	0
14	0.5					0	0
20	0.8						0
Description of injected event ( $h(e_{v,r})$ ) based on signal below							
Time	$ex(e_v)$	$d(e_v)$	$h(e_{v,0.01})$	$h(e_{v,r \geq 0.011})$			
9	Region B	2	0	1			
Additional method result after injection of event							
Time	$p_t$	$ex(s_{m,0.011})$	$d(s_{m,0.011})$	$h(s_{m,0.011})$	$h(s_{m,r > 0.011})$		
10	0.011	Region B	2	1	1		
True positive rate as a function of threshold based on single injected event							
$TP(r < 0.011)$	$TP(r \geq 0.011)$						
0	1						
False positive rate for artificial data in Table 1 as a function of threshold							
$FP(r)$							
$r < 0.003$	0						
$0.003 \leq r < 0.011$	0.05						
$0.011 \leq r < 0.2$	0.05						
$0.2 \leq r < 0.5$	0.1						
$0.5 \leq r < 0.8$	0.15						
$0.8 \leq r < 0.99$	0.2						
$r \geq 0.99$	1						

<sup>a</sup>Assume all other time periods have  $P = 0.99$  and  $h(s_{m,0.99}) = 0$ .

In Table 2, we continue the example, showing nine additional simulated events and the minimum threshold at which there is a signal that hits them, ordered by that threshold. We also show the relative times of the signal and reference signal, plus the proportion time saved, mean proportion time saved and timeliness weight. The difference between the mean proportion time saved and timeliness weight is apparently slight in this example. In general, the timeliness weight for a given threshold is somewhat larger than the mean proportion time saved, suggesting a greater weight given to the earlier signals by the timeliness weight.

In Figures 1–6, we display the six described metrics for this example, beginning with the simple ROC curve. The data used to generate these figures is explicitly contained in Tables 1 and 2, except for TROS3, which is implied by Tables 1 and 2. Note that the weighted ROC curves do not incorporate the several timeliness values observed for a given specificity when the sensitivity changes several times across different thresholds that result in the same specificity. For example, the specificity does not change for  $r$  values between 0.003 and 0.2, whereas in that range, the sensitivity increases from 0 to 0.6, taking five intermediate values; only the largest of these values appears in the ROC

**Table 2** Results of applying the hypothetical method in detecting to 10 simulated events added to the data set described in Table 1, ordered by the *P*-value generated by the method, along with the timeliness data and overall true positive rates from this example

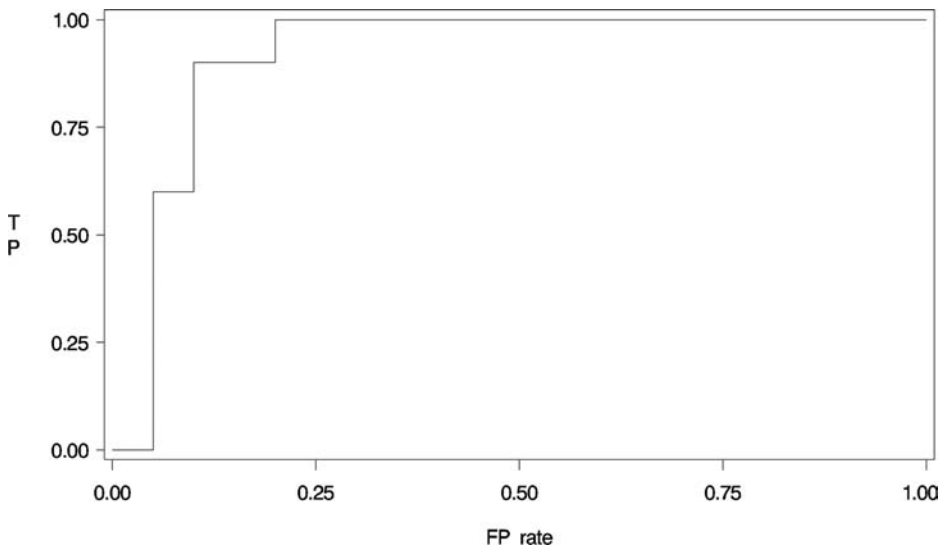
Time	Min <sup>a</sup> $rh(e_{v,r}) = 1$	TP( <i>r</i> )	TD( $s_{m,r}, e_v$ )	TD(ref <sub>v</sub> , $e_v$ )	TD(ref <sub>v</sub> , $s_{m,r}$ )	TS( $e_v, s_{m,r}$ )	$\overline{TS}(e_v, s_{m,r})$	TW( <i>r</i> )
14	0.006	0.1	4	3	0	0	0	0
9	<b>0.011</b>	<b>0.2</b>	<b>1</b>	<b>6</b>	<b>5</b>	<b>0.83</b>	<b>0.08</b>	<b>0.09</b>
12	0.05	0.3	4	6	2	0.33	0.12	0.13
2	0.1	0.4	2	4	2	0.5	0.17	0.20
4	0.12	0.5	7	14	7	0.5	0.22	0.27
16	0.18	0.6	1	1	0	0	0.22	0.27
8	0.23	0.7	4	5	1	0.2	0.24	0.29
6	0.38	0.8	2	14	12	0.86	0.32	0.38
11	0.41	0.9	3	8	5	0.625	0.38	0.42
9	0.998	1.0	12	12	0	0	0.38	0.42

Note: The single event described in Table 1 is given in bold.

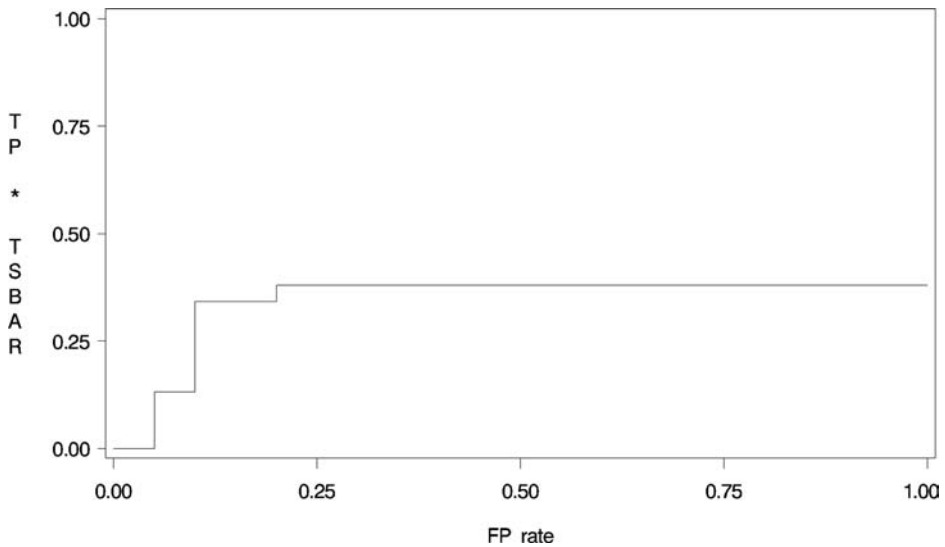
<sup>a</sup>Assume no earlier detection results from larger thresholds.

curves. In contrast, the three-dimensional TROS1 and TROS2 surfaces separate these values. These can be seen in Figures 4 and 5.

Table 3 shows the resulting measures of each metric. Clearly, the ROC misrepresents the value of the method by not using the information on timeliness. The AUROC has a large value, suggesting that the detection method performs rather well at signaling the events.<sup>36</sup> The newly proposed metrics, in contrast, provide a much less rosy view of the method, suggesting in the best case only 59% of a perfect score. This is likely due to the three events that were detected no sooner than the reference signal.



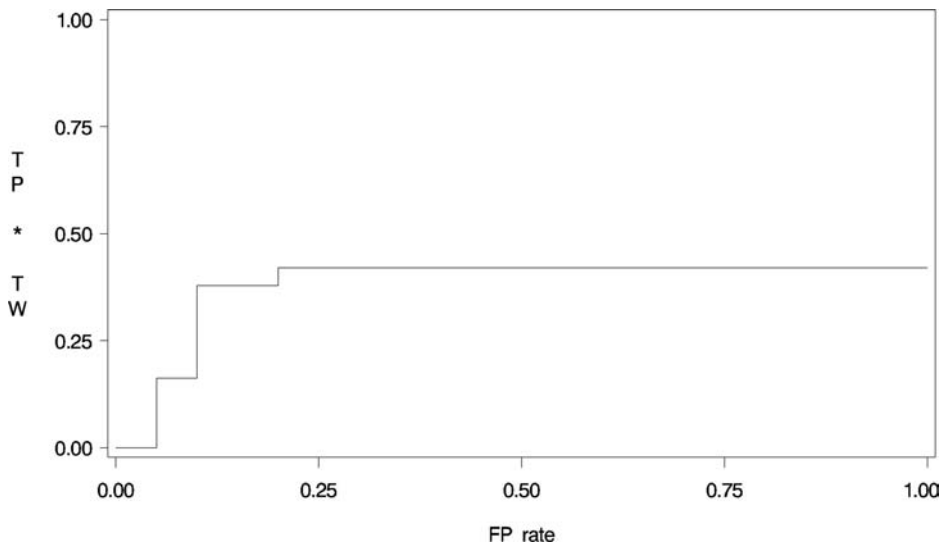
**Figure 1** Conditional ROC curve corresponding to the data in Tables 1 and 2.



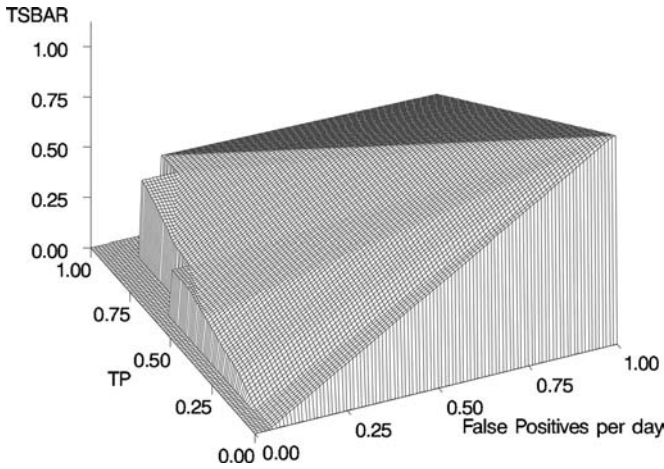
**Figure 2** Conditional weighted ROC curve corresponding to the data in Tables 1 and 2, weighted by the mean proportion time saved.

### 4.2 Application to the simulation

In this example, there is a fixed reference of nine days; even for the ROC curve, hits after this point are not included. Thus, the AUROC values presented subsequently are not, in fact, values computed from values  $TP(r)$  but only from true positives signaling



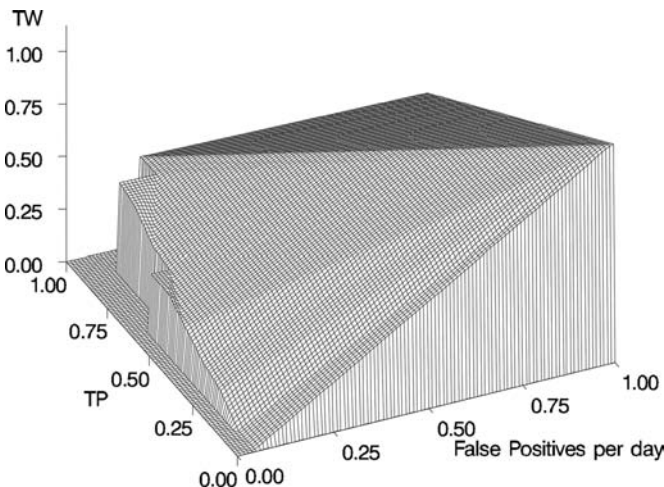
**Figure 3** Conditional weighted ROC curve corresponding to the data in Tables 1 and 2, weighted by the timelines weight.



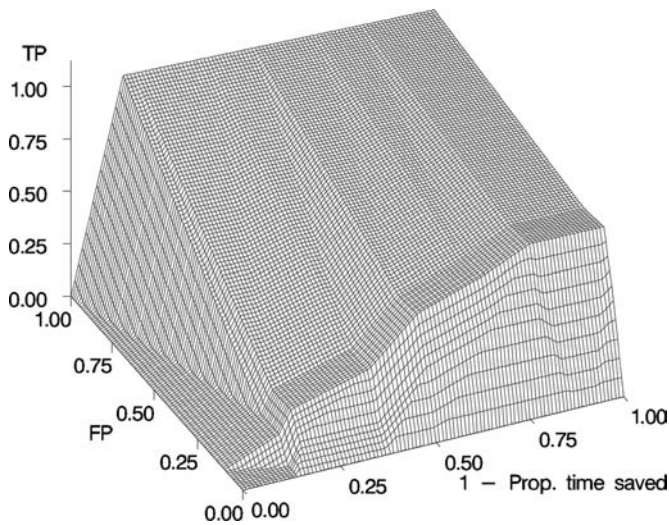
**Figure 4** Timeliness-ROC surface corresponding to the data in Tables 1 and 2; vertical dimension is the mean proportion time saved.

at threshold  $r$  before the ninth day. If they had been based on the former, they would be larger than in the following tables.

In Tables 4–7, we show the results of applying the metrics in the simulation study. Table 4 shows the result of each metric for a range of probability of illness per spore when using a scan statistic with a three-day maximum duration. As in the artificial data example, the largest values are always produced by the time-insensitive AUROC. However, the weighted ROC curves and the volumes based on the weights suggest almost no value for the method for the smaller probabilities of illness per spore, whereas the volume



**Figure 5** Timeliness-ROC surface corresponding to the data in Tables 1 and 2; vertical dimension is the timeliness weight.



**Figure 6** Timeliness-ROC surface corresponding to the data in Tables 1 and 2; with usual ROCs repeatedly recalculated using hits only saving as much as a given proportion time saved.

based on repeated modified ROC curves suggests some usefulness. Note the counter-intuitive discovery that the VUTROS1 and VUTROS2 metrics achieved a high value with the next-to-highest probability of illness. If a single statistical detection method was evaluated in an absolute sense to assess the worth of a system, a table such as Table 4 (or a subset of its columns), might be used to show how well a system detected a given set of event scenarios.

**Table 3** The values of six metrics in the artificial data example described in Tables 1 and 2, with a brief description of each metric

Metric	Description	Value
AUROC	Area under the ROC curve, ignoring timeliness	0.920
AUWROC1	Area under the weighted ROC curve, weight is average timeliness	0.345
AUWROC2	Area under the weighted ROC curve, weight is timeliness weight	0.395
VUTROS1	Volume under the surface with average timeliness for the third axis	0.415
VUTROS2	Volume under the surface with timeliness weight for the third axis	0.489
VUTROS3	Volume under the surface with hits defined by occurring by a time	0.590

**Table 4** Six metrics applied to simulation results with five probabilities of illness per spore in simulated releases in the urban region around Boston, Massachusetts, with a class A spore dispersal pattern, with statistical signals generated via a scan statistic with three-day maximum duration signal

Pr(ill)	AUROC	AUWROC1	AUWROC2	VUTROS1	VUTROS2	VUTROS3
$10^{-10}$	0.067	0.020	0.004	0.014	0.010	0.025
$5 \times 10^{-10}$	0.235	0.087	0.087	0.064	0.045	0.103
$10^{-9}$	0.378	0.150	0.105	0.110	0.086	0.175
$5 \times 10^{-9}$	0.865	0.445	0.408	0.212	0.219	0.492
$10^{-8}$	0.935	0.552	0.412	0.181	0.169	0.594

**Table 5** Six metrics applied to seven signaling methods assessed on to simulation results with probability of illness per spore equal to  $10^{-8}$  in simulated releases in the urban region around Boston, Massachusetts with a class A spore dispersal pattern

Method	AUROC	AUWROC1	AUWROC2	VUTROS1	VUTROS2	VUTROS3
Scan1	0.921	0.563	0.381	0.209	0.198	0.603
Scan3	0.935	0.552	0.412	0.181	0.169	0.594
Scan7	0.932	0.521	0.437	0.207	0.120	0.567
GLMM1	0.891	0.499	0.399	0.270	0.260	0.543
GLMM3	0.912	0.499	0.429	0.283	0.275	0.545
GLMM7	0.860	0.416	0.458	0.246	0.263	0.465
TS	0.798	0.563	0.222	0.215	0.164	0.590

*Note:* The seven methods are scan statistics with one, three, and seven-day maximum duration, generalized linear mixed models with fixed one, three and seven-day durations and a time-series method with one-day duration only.

**Table 6** Six metrics applied to seven signaling methods assessed on to simulation results with probability of illness per spore equal to  $10^{-9}$  in simulated releases in the urban region around Boston, Massachusetts with a class A spore dispersal pattern

Method	AUROC	AUWROC1	AUWROC2	VUTROS1	VUTROS2	VUTROS3
Scan1	0.410	0.171	0.130	0.081	0.059	0.198
Scan3	0.378	0.150	0.105	0.110	0.086	0.175
Scan7	0.349	0.130	0.091	0.120	0.098	0.154
GLMM1	0.235	0.081	0.048	0.082	0.064	0.098
GLMM3	0.276	0.094	0.067	0.115	0.093	0.114
GLMM7	0.271	0.087	0.068	0.104	0.091	0.108
TS	0.650	0.410	0.200	0.130	0.090	0.437

*Note:* The seven methods are scan statistics with one, three, and seven-day maximum duration, generalized linear mixed models with fixed one, three and seven-day durations and a time-series method with one-day duration only.

**Table 7** Six metrics applied to seven signaling methods assessed on to simulation results with probability of illness per spore equal to  $10^{-10}$  in simulated releases in the urban region around Boston, Massachusetts with a class A spore dispersal pattern

Method	AUROC	AUWROC1	AUWROC2	VUTROS1	VUTROS2	VUTROS3
Scan1	0.083	0.023	0.008	0.008	0.007	0.033
Scan3	0.067	0.020	0.004	0.014	0.010	0.025
Scan7	0.068	0.023	0.004	0.019	0.013	0.028
GLMM1	0.016	0.005	0.000	0.006	0.004	0.006
GLMM3	0.028	0.008	0.001	0.012	0.008	0.010
GLMM7	0.046	0.017	0.008	0.019	0.016	0.020
TS	0.392	0.160	0.152	0.039	0.026	0.194

*Note:* The seven methods are scan statistics with one, three, and seven-day maximum duration, generalized linear mixed models with fixed one, three and seven-day durations and a time-series method with one-day duration only.

Tables 5–7 show the results of applying competing methods to a single set of simulation parameters. Among the spatial methods, the scan methods generally outperform the GLMM methods for the metrics in the case of the two smaller probabilities of illness (Tables 6 and 7), whereas the GLMM methods are superior for the simulations with

the highest probability of illness (Table 5) for some metrics. In contrast, the time-series method is poorer than some scan methods when the probability of illness per spore is largest, but performs better, sometimes remarkably better, when the probability of illness is smaller.

## 5 Discussion

In the simulated data example and in general, all of these evaluation metrics really measure the performance of a surveillance system – the means by which data are collected and arranged for analysis – not just a statistical method. For example, in the presented simulation application, individuals visit their physician on the day their symptoms appear, but the data are not abstracted and analyzed until the next day. Thus, even a ‘perfect’ statistical method cannot generate signals until the day after additional cases enter the system – a heavy timeliness penalty for the system against a hypothetical perfect *system* that detects and analyses incident cases instantly. In addition, not all affected individuals are included in the surveillance data stream, and the probability of inclusion varies over the surveillance region. Therefore, there will be some simulated anthrax releases that never cause illness in study subjects, implying that those events also cannot be detected by even a perfect method.

The differences between the weighted AUROC metrics and the volume-based three-dimensional metrics based on the weights are not remarkable. Any may be used, and we suggest they may be of most use when the goal is an absolute assessment of the value of a system-method, as in Table 4. These methods may more accurately reflect the small actual value of these systems relative to the reference signals. The anomalous result for the VUTROS1 and VUTROS2 volume metrics suggests avoiding these metrics, however. Of the two weighted ROC areas, the simpler is the one based on a simple statistic of the distribution of the time saved proportions; as there seems to be little difference between the two, we suggest this metric is preferred.

In contrast, the VUTROS3 volume metric is of most use in comparing methods’ performance in a given system under a given simulation method, as in Tables 5–7. In this case, relative values are at least as useful as absolute values, and the volume metrics can be more sensitive to differences in methods. In application to relative comparisons, it may also be worthwhile to rescale the values so that the maximum attainable value in a particular system is 1. This could be done by omitting simulated events that do not cause cases to enter the surveillance data stream and by changing the time of the event discussed earlier to refer to the time the first simulated case enters the data stream.

The motivation, development and examples are framed in the setting of spatial data. However, the use of the proposed metrics in settings without spatial data or ignoring what spatial information exists may also be considered. It would be easy to do, as noted earlier, because the extent of such signals may be treated as all of the available area. A prior question, however, is whether there is a need for such evaluation, in light of the known properties and analytic optimality developments regarding time-series methods. There are two reasons why the evaluation of time-series methods through these metrics is still worthwhile. First, it is quite likely that the assumptions required for the analytic

results will be violated by the complex nature of public health processes. Secondly, side-by-side comparison of time-series and spatial methods in common data sets will help to describe the relative strengths and weaknesses of the methods, as in the example.

The proposed definitions of the test characteristics are somewhat unusual; in screening applications they refer to persons, and the sensitivity is the probability that, for example, a tumor will be detected. In the present application, the time periods are treated as the individuals. It may seem odd to declare a true positive result when a signal's spatial extent covers 40 000 non-cases and 1 positive case or to consider as equivalent two false positive signals with spatial extent including 500 and 50 000 persons. For people who institute and perform surveillance, the main question which they must consider in practical application is whether the signals they receive are likely to be false positives and whether the method will likely generate a signal if an event occurs. This is because much work is incurred by any signal and the incremental cost of an additional case is smaller in comparison with the costs of beginning an investigation. In addition, they usually feel that they will detect an event even if there are few cases contributing to a signal. Thus, the proposed definitions of the test characteristics are natural for those who perform public health surveillance.

Other unsuspected differences from the usual ROC setting also occur. One simple result for the AUROC is that no method can have a value below 0.5. Such a test can be interpreted as a sign of 'no event', meaning that for example,  $P$ -values greater than 0.05 suggest events; the AUROC for this test is 1 less the value of the original test and therefore greater than 0.5. However, for the definitions of the sensitivity and specificity discussed earlier this is not the case, because the area requirement of a hit makes reversing the definition of the test undefined – essentially, the reversed test would have to declare an event in all regions not included in the original signal. A corollary of this observation is that the weighted ROC curves need not include the point where the sensitivity is 1 and the specificity is 0. Meanwhile, reversing the test would not have the salutary effect on the specificity observed in the usual setting.

The discussion is framed only for public health surveillance to detect bioterrorism. In fact, there is no need to see the proposed methods in such a limited light. Returning to the canonical example of screening for a tumor, a new screening method might improve over an older method by detecting tumors at an earlier stage or simply earlier than a current gold-standard screen. For particularly aggressive conditions, a smaller area under the curve might be acceptable if timeliness gains could be achieved. The proposed metrics would allow assessing this trade-off.

## **5.1 Limitations and directions for future work**

One common criticism of the ROC curve and the AUROC is that they treat all specificities equally. In contrast, frequently only part of the ROC curve is of interest, namely the part with specificities in an 'acceptable' range for the application to hand. For example, in the current applications, some proportion of time periods with false positives might be untenable, regardless of the sensitivity achieved at those thresholds. This has led some to suggest that 'partial' ROC curves, and the area underneath them, should be used instead. This is a reasonable complaint and approach, and the methods proposed might be adapted to the examination of a particular range of specificities.

One limitation of the proposed metrics as a means of comparing methods is that they are susceptible to ‘gaming’, meaning altering the statistical method intentionally so as to perform well on the chosen metric. For example, one might add several miles to the radius of a signal so as to include more area. Unintentionally generating larger extent signals would have the same effect. For example, the scan methods in the example, which generate signals with an extent up to 50% of the surveillance area, tend to outperform the GLMM methods, which are limited to a single postal zip code, out of about 200 in the surveillance area. On the other hand, the time-series method, with the largest possible extent signals, actually performed worst on most metrics for the most virulent simulated anthrax release. For the present, we leave the effect of the relative extent of signals from different methods as a question for secondary data analysis.

In each of the metrics defined earlier, it is implied that the value of each incremental time saved has the same value. That is, it is implied that saving the first 10% of the time, assuming a reference signal exists, has the same value as saving the last 10%. Alternatively, without a reference signal, it is implied that detecting a signal in the first versus the second time period has the same additional value as detecting in the tenth rather than the eleventh. In reality, this is not the case. Instead, there is often a premium on earlier detection, so that detecting an event on the first day versus the second may be worth 10 times as much as detecting it on the day before the reference versus the reference. It is worthwhile therefore to outline methods by which the various metrics may weight the time periods.

Assume weights between 0 and 1. The WROC1 and TROS1 can be weighted by dividing the time of signal by a function of proportion time saved. If earlier detection has increased the incremental value relative to the late detection, the function should be increasing in proportion time saved. WROC2 and TROS2 can be weighted by rescaling the horizontal axis of the timeliness weight curve according to the same function before calculating the timeliness weight. TROS3 can similarly be weighted by transforming the timeliness axis before the VUTROS3 is calculated.

Another implication of the earlier discussion is that the number of people affected by the event, with versus without a signal, is ignored. In fact, this could and probably should be a factor in evaluating a method. If a method performs perfectly in detecting events with total affected persons <0.01% of the population, but is not sensitive to events affecting 10% of the population, the evaluation method should probably assign less value to that method than the affected-population-independent metrics discussed here.

A more basic question is whether the reliance on the proportion of time saved is a reasonable strategy. The result of this approach is that detecting a signal in halfway between the event time and the reference signal has a fixed impact, regardless of whether the reference signal is generated two, ten, or twenty days after the event. In fact, whether this is plausible or in what way it is implausible is a value judgment. The alternative is to value each day saved equally, regardless of whether the signal occurs the nineteenth of twenty days after the event or the first of two days. This seems less likely to match the needs of the public health practitioners who should use metrics like these to choose statistical methods. In the event of a fixed reference signal, there is happily no conflict, as the proportion time saved is equivalent to the number of days saved. However, a fixed referent seems unlikely to be a realistic scenario for event detection.

As with the observation that real data examples provide imperfect assessments of performance for the purposes of detecting bioterrorism, we find here that a better performance in detecting a simulated anthrax attack need not imply a better performance in detecting naturally caused illness of small or large proportion or indeed a bioterrorism attack using a different agent. In fact, we suggest that simulation of the illness to be detected is the appropriate way to choose the statistical method best suited to detect that sort of attack.

Among the myriad details not addressed earlier, we note that many methods may generate more than one  $P$ -value and thus more than one signal per day. In this case, the above calculations would include a summation over the signals per day as well as the days, complicating the notation discussed. However, the heuristics of the proposed methods will be easily extended to accommodate this situation.

## 6 Conclusion

We discussed the need for a metric to evaluate surveillance methods in public health and more generally. For the case of ‘injected’ signals in real public health surveillance data streams, we discussed how ROC-like curves can be calculated and attendant caveats. We then developed six metrics that extend the ROC curve either via weighting for timeliness, resulting in weighted ROC curves, or via adding timeliness as a third dimension, resulting in timeliness-ROC surfaces. In discussion, we argued that the weighted measures give a better sense of absolute performance, whereas the volume metrics provide a better insight into the relative performance of statistical methods.

By incorporating sensitivity, specificity and timeliness into single metrics, we hope to simplify the comparison of methods and help the developers and users of monitoring systems make rational choices about which methods to implement.

## Acknowledgements

This work was supported, in part, by CDC contract UR8/CCU115079 and by NIH grant NL0008707.

## References

- 1 Bravata DM, McDonald KM, Smith WM, Ryzak C, Szeto H, Buckeridge DL *et al.* Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Annals of Internal Medicine* 2004; 140(11): 910–22.
- 2 CDC. Framework for evaluating public health surveillance systems for early detection of outbreaks. *Morbidity and Mortality Weekly Report* 2004; 53(RR-5): 1–16.
- 3 ESSENCE. ESSENCE: Electronic Surveillance System for the Early Notification of Community-based Epidemics.
- 4 Greenko J, Mostashari F, Fine A, Layton M. Clinical evaluation of the emergency medical services (EMS) ambulance dispatch-based syndromic surveillance system, New York City. *Journal of Urban Health* 2003; 80(2 Suppl. 1): i50–6.

- 5 Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. *Emerging Infectious Disease* 2004; 10(5): 858–64.
- 6 Hutwagner L, Thompson W, Seaman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *Journal of Urban Health* 2003; 80(2 Suppl. 1): i89–96.
- 7 Irvin CB, Nouhan PP, Rice K. Syndromic analysis of computerized emergency department patients' chief complaints: an opportunity for bioterrorism and influenza surveillance. *Annals of Emergency Medicine* 2003; 41(4): 447–52.
- 8 Lazarus R, Kleinman K, Dashevsky I, Adams C, Kludt P, DeMaria A Jr *et al.* Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. *Emerging Infectious Diseases* 2002; 8(8): 753–60.
- 9 Lober WB, Trigg LJ, Karras BT, Bliss D, Ciliberti J, Stewart L *et al.* Syndromic surveillance using automated collection of computerized discharge diagnoses. *Journal of Urban Health* 2003; 80(2 Suppl. 1): i97–106.
- 10 Lombardo J, Burkom H, Elbert E, Magruder S, Lewis SH, Loschen W *et al.* A systems overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II). *Journal of Urban Health* 2003; 80(2 Suppl. 1): i32–42.
- 11 Mandl KD. Infrastructure and methods to support real time biosurveillance. Proceedings of the National Science Foundation Next Generation Data Mining Workshop 2002.
- 12 Platt R, Bocchino C, Caldwell B, Harmon R, Kleinman K, Lazarus R *et al.* Syndromic surveillance using minimum transfer of identifiable data: the example of the National Bioterrorism Syndromic Surveillance Demonstration Program. *Journal of Urban Health* 2003; 80(2 Suppl. 1): i25–31.
- 13 Wong W-K, Moore A, Cooper G, Wagner M. WSARE: What's Strange About Recent Events? *Journal of Urban Health* 2003; 80(2 Suppl. 1): i66–75.
- 14 CDC. Public health emergency preparedness and response; Biological diseases/agents. Update: influenza activity – United States and Worldwide, 2004–05 season. *Morbidity and Mortality Weekly Report* 2005; 54: 631–34.
- 15 Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *American Journal of Epidemiology* 2004; 159(3): 217–24.
- 16 Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine* 2005; 2(3): e59.
- 17 Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences of the United States of America* 2003; 100(4): 1961–5.
- 18 Reis BY, Mandl KD. Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making* 2003; 3(1): 2.
- 19 Chatfield C. *Time-series forecasting*. Chapman and Hall/CRC, 2000.
- 20 Frisen M, Demare J. Optimal surveillance. *Biometrika* 1991; 78: 271–90.
- 21 Frisen M, Sonesson C. Optimal surveillance. In Lawson AB, Kleinman K eds. *Spatial and syndromic surveillance for public health*. Wiley, 2005: 51.
- 22 Hosmer DW, Lemeshow S. *Applied logistic regression*, second edition. John Wiley and Sons, 2000: 162.
- 23 Yih W, Abrams A, Danila R, Green K, Kleinman K, Kulldorff M *et al.* Ambulatory care diagnoses as potential indicators of outbreaks of gastrointestinal illness in Minnesota. *Morbidity and Mortality Weekly Report* 2005, to appear.
- 24 Mostashari F, Kulldorff M, Hartman JJ, Miller JR, Kulasekera V. Dead bird clusters as an early warning system for West Nile virus activity. *Emerging Infectious Diseases* 2003; 9(6): 641–6.
- 25 Kleinman K, Abrams A, Mandl KRP. Simulation for assessing statistical methods of bioterrorism surveillance. *Morbidity and Mortality Weekly Report* 2005, to appear.
- 26 Buckeridge D, Burkom H, Moore A, Pavlin J, Cutchis P, Hogan W. Evaluation of syndromic surveillance systems: development of an epidemic simulation model. *Morbidity and Mortality Weekly Report* 2004; 53(Suppl.): 137–43.
- 27 Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification

- of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC Public Health* 2001; **1**(9).
- 29 Spijkerboer HP, Beniers JE, Jaspers D, Schouten HJ, Goudriaan J, Rabbinge R *et al.* Ability of the Gaussian plume model to predict and describe spore dispersal over a potato crop. *Ecological Modeling* 2002; **155**: 1–18.
- 30 Meselson M, Guillemin J, Hugh-Jones M, Langmuir A, Popova I, Shelokov A *et al.* The Sverdlovsk anthrax outbreak of 1979. *Science* 1994; **266**(5188): 1202–8.
- 31 Brookmeyer R, Blades N. Prevention of inhalational anthrax in the US outbreak. *Science* 2002; **295**(5561): 1861.
- 32 Kulldorff M. Prospective time periodic geographic disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A* 2001; **164**: 61–72.
- 33 Kleinman K, Abrams A, Kulldorff M, Platt R. A model-adjusted space–time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection* 2005; **133**: 409–19.
- 34 Kleinman K. Generalized linear models and generalized linear mixed models for small-area surveillance. In Lawson AB, Kleinman K eds. *Spatial and syndromic surveillance for public health*. Wiley, 2005: 77–94.
- 35 Hosmer DW, Lemeshow S. *Applied logistic regression*, second edition. John Wiley and Sons, 2000.

Copyright of *Statistical Methods in Medical Research* is the property of Sage Publications, Ltd. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.