Statistics in Medicine

Received 30 April 2010, Accepted 3 May 2010 Published online 4 February 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.3982

Rejoinder: Some methodological issues in biosurveillance‡

Ronald D. Fricker Jr*†

This paper is a rejoinder to the commentaries on Some Methodological Issues in Biosurveillance. Published in 2011 by John Wiley & Sons, Ltd.

Keywords: epidemiologic surveillance; syndromic surveillance; public health surveillance; bioterrorism

1. Rejoinder

I would first like to thank Professor David Buckeridge, Dr Howard Burkom, Dr Krista Hanni, Mr Henry Rolka, and Professors Bill Woodall and Kwok-Leung Tsui for their thoughtful, interesting, and informative commentaries. Given our divergent backgrounds—from local to federal public health, from academia to the CDC, from experience in industrial quality control to applied biosurveillance—I was a bit surprised but very pleased to find that we are all largely in agreement.

In terms of the commentaries, let me begin by agreeing with Professor Buckeridge and Dr Hanni that developing standards and guidelines for biosurveillance studies would be very helpful in advancing the state of the art. As Professor Buckeridge correctly notes, my background and research interests focus on detection algorithms, but he is absolutely correct in suggesting that on-going research and development is necessary across the entire spectrum of biosurveillance functions and activities. For example, in some recent research with a local biosurveillance system based on EARS, we found that changing the text matching algorithms and/or the syndrome logic can have a large and significant impact on the number of individuals coded with a given syndrome.

Specifically, for one year of county-level clinic data, 9093 visits (out of 153 696) were coded for the ILI syndrome using the unmodified EARS program. Modifying the symptom coding logic by allowing expanded text matching, in order to increase sensitivity, resulted in a 53 per cent increase in the number of ILI syndromes coded. In contrast, requiring at least two symptom matches prior to coding the ILI syndrome, in order to increase specificity, decreased the number of ILI syndromes coded to 8 per cent of the original. Making changes to the text matching logic, to improve both sensitivity and specificity, resulted similarly in large swings in the number of coded ILI syndromes.

At issue is that little is known about which of these choices, all of which seem reasonable, would provide the best set of syndromic data upon which to use a detection algorithm. The obvious point is that a detection algorithm is only as good as the data, and there are arguably greater gains to be made with improvements in things such as data collection and management, natural language processing, and syndrome definitions than in trying to develop the next generation of detection algorithms. Directly related to this issue, Dr Hanni further makes the case for a standardized set of *validated* syndrome definitions, as well as standards for assessing the sensitivity and specificity of the definitions. The relevant question, as posed by Mr Rolka, is 'How much reality is reflected in biosurveillance data?' Mr Rolka further notes that the recent government initiatives are likely to lead to an 'influx' of data making the development of methods and procedures for ensuring data quality that much more important. Both Dr Hanni and Mr Rolka suggest that the public health community should take the lead in addressing this issue.

In a similar vein, Professors Woodall and Tsui make the point that biosurveillance applications have considerably more sources of variation compared to the typical statistical process control (SPC) application. However, it is not at all clear, at least to me, that a concerted effort has been made to identify and eliminate as much variation as possible from

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

biosurveillance data and systems. After all, EED is an exercise in identifying a signal amid noise and sensitivity can come from improving signal detection methods, but potentially greater gains may come from reducing the noise (i.e. variation unrelated to the signal) whenever and wherever possible.

Given this, I certainly agree with Professor Buckeridge's call for more studies to understand how surveillance works in real-world public health settings. Just as SPC has its roots firmly planted in the application and performance of methods in practice, so should biosurveillance. As Walter Shewhart, the originator of SPC said, '…the fact that the criterion we happen to use has a fine ancestry of highbrow statistical theorems does not justify its use. Such justification must come from empirical evidence that it works.' [1, p. 18] That said, in my opinion, good simulation studies—informed by how surveillance systems are used in the real-world—are as important or more important because they can provide generalizable results and knowledge. My perspective in this regard is more in line with Professors Woodall and Tsui's discussion of the use of case study data and Mr Rolka's simulation discussion.

I also agree with Professor Buckeridge's assertion (echoed by Mr Rolka) that situational awareness is 'intuitively clear but vaguely defined and poorly studied.' Unfortunately, without better definitions it will be hard to make progress in understanding how to best apply biosurveillance systems to enhance situational awareness. And, in spite of my research interest in EED methods, I would suggest that situational awareness is probably the more important function for biosurveillance, since it enhances public health surveillance and management before, during, and after an outbreak or other adverse event, as well as during periods of routine incidence. Dr Burkom's comments seem to support this viewpoint.

In addition, Dr Hanni points out that biosurveillance systems provide benefits other than EED and SA, including establishing relationships between public health departments and healthcare providers (which could be of significant importance during an outbreak), improving public health general preparedness, and providing additional capabilities for providing 'data-driven' analyses. These and other benefits of biosurveillance should not be overlooked as biosurveillance system research, development, and implementation continues.

In terms of Dr Burkom's comments on what he referred to as the 'Figure 2 analysis,' it seems my discussion here was not as clear as it should have been. In it, I am not advocating some sort of competition or either-or choice between biosurveillance and clinicians. Rather, I am suggesting that this type of analysis can help identify those situations in which biosurveillance can 'add value' and those situations in which it is not likely to do so. The intention is not to justify biosurveillance, but rather to better understand in which situations biosurveillance EED is most and least effective. Such knowledge can be useful in guiding practitioners about what they can expect from their biosurveillance systems. It can also be useful for helping researchers and system developers target efforts for improvement.

2. More on metrics

Fraker *et al.* [2] note that 'Substantially more metrics have been proposed in the public health surveillance literature than in the industrial monitoring literature.' In my original paper I noted that a common set of performance metrics is necessary to advance the science and practice of EED, where the plethora of metrics has made it difficult and sometimes impossible to compare results across the literature. I then recommended convening a panel of experts to propose a common set of metrics. This was motivated by the idea that a collaborative effort would have a higher chance of its recommendations being adopted by the wider community. However, Dr Burkom describes the real-world challenges inherent in achieving consensus with such a panel.

In their commentary, Professors Woodall and Tsui provided a more extended discussion about metrics for use in biosurveillance, concluding that more research is required to identify the appropriate metrics, particularly for spatio-temporal methods. This prompted me to give the matter some additional research and thought. Drawing on a 'discussion' that has been ongoing in the literature, most notably Sonesson and Bock [3], Frisén and Sonesson [4], and Fraker *et al.* [2], I would like to expand on Professors Woodall and Tsui's comments here and even go so far as to recommend a set of metrics.

To that end, let us begin by noting that metrics must serve two purposes:

- For researchers and academics, metrics should permit clear and consistent performance comparisons between methods.
- For public health practitioners and biosurveillance system operators, metrics should facilitate effective biosurveillance system implementation and operation.

In terms of the latter purpose to facilitate effective implementation and operation of biosurveillance systems, the metrics should be able to answer the following three questions:

- When there is an outbreak, how fast will the method detect the outbreak?
- When there is no outbreak, what is the time between false positive signals?
- For a particular type of outbreak, what is the chance that the method will detect the outbreak?

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 434–441

As it turns out, metrics that are useful for answering these operational questions are also useful and quite appropriate for the purpose of EED performance comparisons by researchers and academics.

The need for an answer to the first question is obvious and, resource constraints aside, faster is always to be preferred. However, for a given detection method, improving the speed of detection during an outbreak comes at the cost of increasing the number of false positive signals. Thus, when setting a method's signaling threshold it is absolutely critical to address the second question. Specifically, when choosing a threshold for an implementation, it should be set so that the number of false signals is tolerably small, meaning that the biosurveillance system operators have the necessary resources to investigate and adjudicate the signals (meaning determine whether a particular signal it either a true positive signal or a false positive signal).

This approach will trouble some biosurveillance researchers and practitioners, since it makes the speed of detection secondary to daily operational considerations. They appropriately argue that one should set the threshold low enough so that a method signals a true outbreak very quickly. However, since threshold setting drives both true and false positive signaling, if the threshold is set too aggressively it will quickly result in an excessive number of false positive signals that the organization does not have the resources to investigate. And, a critical point to remember is that *an uninvestigated signal is as good as no signal at all*. The only way to control this problem is to first set the threshold to achieve a tolerable false positive rate and then allow that to drive the achievable detection speed for an actual outbreak.

The discussion thus far should make it clear that for EED time is the critical measure of interest, both in terms of the time to detect an outbreak and the time between false positive signals. Those familiar with the SPC literature will note that answering the first two questions leads to metrics that are similar to the *average run length* (ARL) metrics used in industrial process control. However, as pointed out by Professors Woodall and Tsui, unlike in industrial quality control, EED metrics based solely on time are not sufficient because outbreaks are transient.

EED metrics based solely on time can be insufficient because it is possible to create detection methods that have a large time between false signals and a small time to detect outbreaks, but which may also miss too many outbreaks. Of course, one would hope that a well-designed method that has a small time to detect an outbreak would also have a high probability of detecting transient outbreaks, but that may not always be the case. Thus, it is also important to ensure that EED methods also have a high probability of detecting such transient outbreaks. This leads to a third metric that answers the third question.

2.1. Recommended metrics for temporal methods

To be recommended, a set of metrics must effectively address the two purposes and be able to answer the three questions previously posed. In addition, they must also be practical, in the sense that they are easily calculated and understood. The metrics which meet all of these goals are as follows:

- *Average time to first signal* (ATFS) is the mean number of time periods it takes for an EED method to first signal, starting from some initial state, given there are no outbreaks. Thus, the ATFS is the expected time to the first false signal.
- *Conditional expected delay* (CED) is the mean number of time periods it takes for the method to first signal, given that an outbreak is occurring *and* that the method signals during the outbreak. Thus, the CED is the expected number of time periods from the start of an outbreak until the first true signal during that outbreak.
- *Probability of successful detection* (PSD) is the probability that the method signals during an outbreak, where the probability of detection is both a function of the EED method and the type of outbreak.

These metrics answer the three questions since (1) ATFS is a measure of the time between false positive signals, (2) CED is a measure of how fast a method will detect an outbreak, and (3) PSD is a measure of how likely a method will detect a particular type of outbreak.

The metrics are mathematically defined as follows. Let $\mathscr{T}_t$ denote a generic EED method statistic at time t and let h denote the method's threshold, where if $\mathscr{T}_t \geqslant h$ the method signals at time t. Also, let τ_s denote the first day of an outbreak, where the notation $\tau_s = \infty$ means that an outbreak never occurs, and let τ_l denote that last day of an outbreak, where if $\tau_s = \infty$ then by definition $\tau_l = \infty$ as well. Finally, let t^* denote the first time the method signals: $t^* = \min(t : \mathscr{T}_t \geqslant h)$.

Then

$$\text{ATFS} = \mathbb{E}(t^* | \tau_s = \infty),$$

$$\text{CED} = \mathbb{E}(t^* - \tau_s | \tau_s \leqslant t^* \leqslant \tau_l),$$

and

$$\text{PSD} = \mathbb{P}(\tau_s \leqslant t^* \leqslant \tau_l).$$

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. **2011**, 30 434–441

Note that these metrics are neither new nor original. They are motivated by and generally consistent with Sonesson and Bock [3] and Frisén and Sonesson [4], who give a more mathematical treatment of the subject and a more complete review of the associated health-related surveillance literature.

Mathematically, the ATFS metric is defined much the same as the in-control average run length (ARL_0) is in SPC. In fact, if the EED statistic is re-set after each signal, so that we can write

$$\text{ATFS} = \mathbb{E}(t^* | \text{ at time } t+1 \ \mathscr{T}_{t^*} = \mathscr{T}_0, \tau_s = \infty),$$

then the ATFS is the same as the ARL_0 in SPC. Indeed, under these conditions, the acronym 'ATFS' can then be taken to stand for the average time between false signals.

However, if the statistic $\mathscr{T}$ is not re-set then the ATFS is the average time from when the statistic $\mathscr{T}$ returns to its initial state until the first false signal. As such it measures the average time from when a method first starts until it first falsely signals, but that may or may not be representative of the time between false signals and it is not appropriate to interpret it in this fashion.

Under the condition that the statistic is not re-set, Fraker *et al.* [2] have proposed the *average time between signal events* (ATBSE) metric, where a *signal event* is defined as the consecutive time periods during which an EED method signals. Under these conditions, the ATBSE should be used rather than the ATFS metric, although in Section 2.3.2 I argue that EED methods should be re-set after each signal.

The CED is conceptually similar to the out-of-control average run length (ARL_1) in SPC, but in SPC when a process goes out of control it stays out of control. Thus, in SPC $ARL_1 = \mathbb{E}(t^* - \tau_s | t^* \geqslant \tau_s)$. As outbreaks are transient in biosurveillance the definition differs because it must incorporate the idea that the signal occurs sometime during the duration of the outbreak.

Note that it will sometimes be useful to set $\tau_s = 1$ when estimating the CED. When doing so, the resulting estimate is referred to as the *zero state* or *initial state* CED because the EED method's statistic $\mathscr{T}_0$ is in its initial state (which is zero for many EED methods) just prior to the outbreak occurring. Alternatively, if the estimates are based on a large τ_s, where the method has been allowed to run for some time, so that $\mathscr{T}$ is not in its initial state, then it is referred to as the *steady state* CED.

PSD does not have an analogue in the SPC literature. As defined above, it is the probability of detecting an outbreak at any time during the outbreak. For longer outbreaks this definition may be too loose, meaning that detection later in the outbreak may not be medically useful. If that is the case, the definition by Sonesson and Bock [3] may be more operationally relevant: $\text{PSD} = \mathbb{P}(t^* - \tau_s \leqslant d | \tau_s \leqslant t^* \leqslant \tau_l)$, where d is the maximum delay required for a successful detection, and where 'successful detection' means early enough in the outbreak cycle that an intervention is medically effective.

2.2. *Recommended metrics for spatio-temporal methods*

As Professors Woodall and Tsui point out, the performance metrics for spatio-temporal methods are more challenging. The issue is that in the temporal problem a successful detection simply requires a method signal during an outbreak (or perhaps within d days of the start of the outbreak). However, in the spatio-temporal problem this is insufficient since a successful detection requires a method both signal during an outbreak *and* accurately identify the location of the outbreak.

However, note that whether the method is purely temporal or spatio-temporal, the two goals and the three questions originally posed remain unchanged. What changes is how the metrics must be defined in order to answer the second and third questions. That is, they must be modified to account for the fact that any definition of a successful detection must have both a temporal and spatial component.

To begin with, let $\mathscr{O}_t$ denote the outbreak region at time t and $\mathscr{S}_t$ denote the signal region at time t. That is, $\mathscr{O}_t$ is the location or locations in (two-dimensional) space of the outbreak at time t, where the area of $\mathscr{O}_t$ is 0 when there is no outbreak. Similarly, $\mathscr{S}_t$ is the location (or locations) where a spatio-temporal method indicates that an outbreak is occurring at time t, where the area of $\mathscr{S}_t$ is 0 when there is no signal. Finally, denote the area of $\mathscr{O}_t$ as o_t, the area of $\mathscr{S}_t$ as s_t, the area of $\mathscr{O}_t \cap \mathscr{S}_t$ as os_t and the area of $\overline{\mathscr{O}}_t \cap \mathscr{S}_t$ as $\overline{o}s_t$.

Sensitivity- and specificity-like measures can then be used to quantify the performance of the method for time period t^* when the spatio-temporal method signals. That is, given a signal, the fraction of the outbreak region contained in the signal region is os_{t^*}/o_{t^*}. Similarly, the ratio of the signal region outside the outbreak region to the non-outbreak region is $\overline{o}s_{t^*}/\overline{o}_{t^*}$. For an ideal signal from a spatio-temporal method, $os_{t^*}/o_{t^*} = 1$, although $0 \leqslant os_{t^*}/o_{t^*} \leqslant 1$, and $\overline{o}s_{t^*}/\overline{o}_{t^*} = 0$, although $0 \leqslant \overline{o}s_{t^*}/\overline{o}_{t^*} \leqslant 1$.

Given this, we can now define the spatio-temporal metrics as:

$$\text{ATFS} = \mathbb{E}(t^* | \tau_s = \infty),$$

$$\text{CED} = \mathbb{E}(t^* - \tau_s \,|\, \tau_s \leqslant t^* \leqslant \tau_l, os_{t^*}/o_{t^*} \geqslant a_1, \text{ and } 1 - \overline{o}s_{t^*}/\overline{o}_{t^*} \geqslant a_2),$$

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. **2011**, 30 434–441

and

$$\text{PSD} = \mathbb{P}(\tau_s \leqslant t^* \leqslant \tau_l, os_{t^*}/o_{t^*} \geqslant a_1, \text{ and } 1-\overline{o}s_{t^*}/\overline{o}_{t^*} \geqslant a_2).$$

What remains is to determine a_1 and a_2, which are parameters that specify how well a spatio-temporal method must identify an outbreak region to be useful and, like d in the PSD metric, they must be specified by the public health community based on medical and operational considerations. Furthermore, the CED and PSD performance of a spatio-temporal EED method can and should be assessed over a range of a_1 and a_2 values, where it is likely that trade-offs will occur between the speed of a true signal and the spatial accuracy of the estimated outbreak location.

2.3. *Alternative metrics*

While the foregoing section recommends a suite of EED method metrics, that naturally begs the question of why the other metrics commonly used in the biosurveillance literature were not recommended. This section briefly addresses that question. This discussion draws from and compliments a similar discussion in Fraker *et al.* [2].

2.3.1. Sensitivity, specificity, and timeliness metrics. Sensitivity, specificity, and timeliness are the most commonly used biosurveillance metrics, where their popularity likely stems from the community's familiarity with and the widespread use of binary classification and classical hypothesis testing. While these metrics are clearly defined and appropriate in these settings, they are less so in the biosurveillance setting for a number of reasons.

First, in binary classification tests one is generally classifying individual patients (sick or not) whereas in biosurveillance one is classifying periods of time (an outbreak exists or not). One critical difference is that in the former the test uses some type of 'gold standard' to definitively classify patients and against which to compare the results of the classification test. In biosurveillance no such gold standard exists, at least in the online operation of the system. If it did, there would be little point in using the other data. In addition, in binary classification tests, with appropriate experimental design, one can reasonably assume that the patients are independent, but independence is generally violated in biosurveillance. Specifically, biosurveillance data is usually autocorrelated and, even if such autocorrelation can be removed via modeling, the signaling statistics for EED methods that use historical data are still autocorrelated.

Second, because the EED data and statistics are rarely independent, and because EED methods never accept the null hypothesis, metrics such as sensitivity and specificity are not well defined. As Fraker *et al.* [2] say, 'Specificity is not useful for the ongoing monitoring scenario because almost all surveillance schemes combine information over time and thus have autocorrelated surveillance statistics. In this situation it is difficult to interpret a false alarm probability because the decisions at each time period to signal or not are dependent and the 'false alarm probability' is not well-defined or directly related to the more meaningful time-to-signal performance.'

To overcome this, researchers have struggled to redefine them in the biosurveillance context. Some definitions that have appeared in the literature include:

- 'Sensitivity is defined as the number of days with true alarms divided by the number of days with outbreaks.' [5]
- 'Sensitivity can be assessed by estimating the proportion of cases of a disease or health condition detected by the surveillance system. Sensitivity can also be considered as the ability of the system to detect unusual events.' [6, p. 14]
- 'Sensitivity is the probability that a public health event of interest will be detected in the data given the event really occurred.' [7, p. 45]
- 'Sensitivity is the probability of an alarm given an outbreak.' [7, p. 413]

However, none of these definitions overcome the fundamental problem of the lack of independence. In addition, as Frisén and Sonesson [4] say, 'Evaluation [of EED methods] by the significance level, power, specificity, and sensitivity which is useful for a fixed sample is not appropriate in a surveillance situation without modification since they have no unique value unless the time period is fixed.'

Third, 'timeliness' is a concept, not a metric. It very appropriately suggests that EED methods should signal quickly, but there is no standard metric that precisely defines what it means to be timely. As previously discussed, the lack of a standard definition exacerbates the problem of synthesizing results across the biosurveillance literature.

These issues extend to other binary classification and classical hypothesis testing metrics, such as PVP and PVN, as well. In fact, suites of metrics that do not include any time-based metrics can give misleading results. Fraker *et al.* [2] give an example in which two EED methods have exactly the same sensitivity, PVP, and PVN, yet one method has significantly better time to detection performance and is thus clearly to be preferred, a fact that is not evident without a time-based metric.

Now, this is not to say that the conceptual approach to measuring EED performance in these three dimensions is undesirable, just that these terms are too imprecise to serve as *metrics* for biosurveillance. This lack of definitional

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. **2011**, 30 434–441

precision has allowed each researcher to define his or her own set of metrics, some of which have been ill-suited to the nature of EED as a sequential decision problem. However, the recommended metrics are both appropriate for the sequential nature of EED and *they map directly onto the sensitivity, specificity, and timeliness performance dimensions*:

- CED is a timeliness metric. It measures the average time it takes an EED method to signal after the start of an outbreak, conditional on the method signalling sometime during the outbreak.
- PSD is a sensitivity-like metric. It measures the probability that an EED method signals during an outbreak.
- ATFS is a specificity-like metric. It measures the frequency that an EED method does not signal when there is no outbreak.

Thus, one can think about ATFS, CED, PSD as well-defined metrics that are appropriately measuring the 'sensitivity,' 'specificity,' and 'timeliness' performance of EED methods.

2.3.2. Run length-based metrics. Run length-based metrics date back to Wald [8, p. 34] who said, 'Restricting ourselves to sequential tests of a given strength (α, β), a test may be regarded as more desirable the smaller the expected number of observations required by the test.' In this vein, the ARL is the standard metric for evaluating control charts. ARL_0 is referred to as the *in-control ARL* and ARL_1 is the *out-of-control ARL*. They are essentially equivalent to the ATFS and CED metrics for SPC data. A closely-related metric is the *average time to signal* (ATS). Where the ARL is based on the average number of observations until a signal, the ATS is the average number of time periods until a signal.

While ARL and ATS have been used to evaluate the performance of EED methods, these metrics fail both to account for the transient nature of outbreaks and that the EED methods' statistics are often not re-set after they signal. In comparison, as pointed out by Professors Woodall and Tsui in their commentary, when an SPC process goes out-of-control it stays in that condition until the control chart signals and the cause is identified and corrected.

To overcome the issues associated with applying the control chart ARL metrics to biosurveillance, various modifications have been proposed, including the ATBSE by Fraker *et al.* [2]. They also define the *average signal event length* (ASEL) as how long, on average, an EED method signals over consecutive time periods.

The ATBSE and ASEL metrics were designed for how biosurveillance systems are often currently operated, where the EED statistics are not re-set after they signal. In this situation, biosurveillance system operators allow the EED methods to continue to run after they signal and interpret the resulting sequence of signals (or lack thereof) as additional information about a potential outbreak. Under these conditions, the ATBSE should be used rather than the ATFS metric.

While these metrics are just as applicable to the EED problem, I prefer ATFS *under the assumption that EED methods should be re-set after each signal*. The reasons for my preference are twofold:

- If the system is being used for bioterrorism, a bioterrorist attack cannot be detected during a natural disease outbreak with a system that is allowed to continuously signal.
- Even when used for natural disease surveillance, sequences of EED signals are crude indicators of the progression of an outbreak. In my opinion, given a true EED signal, other data about the outbreak are likely to be more informative.

But, perhaps most fundamentally, I am of the opinion that biosurveillance systems should be structured so that every EED signal is taken as an alarm worthy of (and requiring) investigation. It should be operated like any other alarm system. I surely would not want a fire alarm system structured so that the firemen waited for the second or third alarm before deciding the signal was real enough to respond. The same with a burglar alarm, or an aircraft collision early warning system, or a tsunami warning, or any other type of alarm system.

2.3.3. Recurrence interval metrics. The recurrence interval (*RI*) is another biosurveillance metric that has been multiply defined. Intuitively, it is often thought of as the average number of time periods (or observations) between signals or signal events. However, this is the definition of the ATS, not the RI. Rather, the RI is 'the fixed number of time periods for which the expected number of false alarms is one' [2]. See [2] and [9] for additional discussion on this point.

In contrast, Kleinman [10] defined the RI as 'the expected number of days of surveillance required so that exactly one p-value as small as the one observed would be expected.' Mathematically,

$$\mathrm{RI} = \frac{1}{p\text{-value} \times S},$$

where S is the number of regions being simultaneously monitored and the p-value is the probability of observing a regional count as extreme or more extreme than the one observed in a particular time period.

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 434–441

When $S=1$, and in the case where the data and monitoring statistic are independent over time and the EED method re-sets after each signal, then ATFS $=1/p$, where p is the probability of a signal under a specific set of circumstances (such as that no outbreak is occurring). Under these conditions, the ATFS is the average time between false signals for a specific pre-set p, where p is the probability of a false signal, and where a signal is generated when p-value $\leqslant p$. However, this relationship between the RI and ATFS (or ARL) does not hold in general, either because of spatial correlation (when $S>1$) or because the monitoring statistics are autocorrelated. In these situations there is no relationship between the ATFS and the RI. As Fraker *et al.* [2] say, EED methods 'with the same RI value can have widely different in-control [i.e. non-outbreak] time-to-signal properties.' Further, Fraker *et al.* [2] say, 'If the marginal probabilities of signaling are not equal, then the number of time periods for which the expected number of signals is one is not constant and the RI is not well-defined.' Nonetheless, the RI is used in a number of biosurveillance EED applications and methods, including the small area regression and testing (SMART) method of Kleinman [10] and the SatScan method of Kulldorff [11].

2.3.4. Other metrics. A number of metrics focused on generalizing the ROC curve have been proposed. Generalizations are necessary in order to incorporate some measure of timeliness. For example, Kleinman and Abrams [12] proposed weighted ROC curves, where each point on the curve is weighted by a measure of timeliness. They also proposed a three-dimensional timeliness-receiver operating characteristic surface or TROS. Buckeridge *et al.* [13] used an activity monitoring operating characteristic (AMOC) curve which is a plot of a timeliness metric versus a false signal rate metric. For spatio-temporal methods, the free response operating characteristic (FROC) curve has been proposed, which is a plot of the fraction of outbreak locations detected versus the false positive detection rate [13].

These measures tend to have the same definitional issues as the 'sensitivity,' 'specificity,' and 'timeliness' metrics. Furthermore, note that ROC curves are plotted over the entire possible range of thresholds ($-\infty\leqslant h\leqslant\infty$) when usually only a subset of the thresholds are of practical use. Thus, when judging the performance it is usually only necessary to compare curves over a smaller range of thresholds. Sometimes these ROC-based metrics are motivated by a desire to distill the performance assessment of two EED methods down to a comparison between two univariate metrics. However, not only is such a comparison subject to the thresholds issue just described, but they also only marginally reduce the complexity since multiple comparisons are still required for various types of outbreaks, background disease incidence patterns, etc.

3. Conclusion

Once again, I thank David Buckeridge, Howard Burkom, Krista Hanni, Henry Rolka, and Bill Woodall and Kwok-Leung Tsui for their commentaries. I hope that the similarity of our opinions is an indication of some general agreement in the biosurveillance research community about important open research questions.

In my opinion, for EED, metrics remains the most important area in which consensus is required. Without a standard set of metrics, advances in biosurveillance EED will (continue to) be hindered. Hopefully this is the first step in a conversation that will lead to an agreed-upon set of rigorous, well-defined metrics.

As part of this conversation, it will be important to remember that metrics must serve both the research community and the public health practitioner and biosurveillance system operator communities. In particular, the latter communities should not be forgotten because the EED method's performance is only as good as biosurveillance system operation, and good metrics will facilitate good system operation.

References

1. Shewhart W. *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company: Princeton, NJ, 1931.
2. Fraker SE, Woodall WH, Mousavi S. Performance metrics for surveillance schemes. *Quality Engineering* 2008; **20**:451–646.
3. Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society, Series A* 2003; **166**:5–21.
4. Frisén M, Sonesson C. Optimal surveillance. *Spatial & Syndromic Surveillance for Public Health*, Lawson AB, Kleinman K (eds). Wiley: New York, 2005; 31–52.
5. Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences* 2003; **100**:1961–1965.
6. Lawson AB, Kleinman K (eds). *Spatial & Syndromic Surveillance for Public Health*. Wiley: New York, 2005.
7. Lombardo JS, Buckeridge DL (eds). *Disease Surveillance: A Public Health Informatics Approach*. Wiley: New York, 2007.
8. Wald A. *Sequential Analysis*. Wiley: New York, 1947.
9. Woodall WH, Marshall B, Joner MD, Fraker SE, Abdel-Salam AG. On the use and evaluation of scan methods for health-related surveillance. *Journal of the Royal Statistical Society, Series A* (*Statistics in Society*) 2008; **171**:223–237.

Published in 2011 by John Wiley & Sons, Ltd.

10. Kleinman K. Generalized linear models and generalized linear mixed models for small area surveillance. In *Spatial & Syndromic Surveillance*, Lawson AB, Kleinman K (eds). Wiley: New York, 2005; 77–94.
11. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A* (*Statistics in Society*) 2001; **164**:61–72.
12. Kleinman K, Abrams A. Assessing surveillance using sensitivity, specificity, and timeliness. *Statistical Methods in Medical Research* 2006; **15**:445–464.
13. Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics* 2005; **38**:99–113.