2011

Some Methodological Issues in Biosurveillance

Fricker, Ronald D. Jr.

Fricker, R.D., Jr. (2011). Some Methodological Issues in Biosurveillance (with commentaries [1] [2] [3] [4] [5] and rejoinder), Statistics in Medicine, 30, 403-441.
https://hdl.handle.net/10945/38755

Research Article

Statistics in Medicine

Received 21 January 2010, Accepted 22 January 2010 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.3880

Some methodological issues in biosurveillance‡

Ronald D. Fricker Jr*†

This paper briefly summarizes a short course I gave at the 12th Biennial Centers for Disease Control and Prevention (CDC) and Agency for Toxic Substances and Disease Registry (ATSDR) Symposium held in Decatur, Georgia on April 6, 2009. The goal of this short course was to discuss various methodological issues of biosurveillance detection algorithms, with a focus on the issues related to developing, evaluating, and implementing such algorithms. The PowerPoint slides from the complete talk can be accessed at `http://faculty.nps.edu/rdfricke/Biosurveillance.htm`**. Published in 2011 by John Wiley & Sons, Ltd.**

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

1. Introduction

Homeland Security Presidential Directive 21 (HSPD-21) [1] defines *biosurveillance* as 'the process of active data-gathering with appropriate analysis and interpretation of biosphere data that might relate to disease activity and threats to human or animal health—whether infectious, toxic, metabolic, or otherwise, and regardless of intentional or natural origin—to achieve early warning of health threats, early detection of health events, and overall situational awareness of disease activity.'

HSPD-21 further defines *epidemiologic surveillance* as 'the process of actively gathering and analyzing data related to human health and disease in a population in order to obtain early warning of human health events, rapid characterization of human disease events, and overall situational awareness of disease activity in the human population.' For the purposes of this paper, I discuss biosurveillance within the context of epidemiologic surveillance (and often in the more specific context of syndromic surveillance), though biosurveillance clearly applies to a broader class of public health surveillance problems.

As HSPD-21 states, biosurveillance systems have two main objectives: early event detection (EED) and situational awareness (SA). I define EED and SA as:

- *EED*: Gathering and analyzing data in advance of diagnostic case confirmation to give early warning of a possible outbreak and, should an outbreak exist, provide early detection.
- *SA*: The real-time analysis and display of health data to monitor the location, magnitude, and spread of an outbreak.

For a more detailed discussion of SA, see `www.satechnologies.com/situation_awareness/`.

Use of biosurveillance, in the form of epidemiological surveillance, is widespread. In 2007–2008, Buehler *et al.* [2] sent surveys to public health officials in 59 state, territorial, and large local jurisdictions. Fifty-two officials responded (an 88 percent response rate) representing jurisdictions containing 94 percent of U.S. population. They found:

- Eighty-three percent reported conducting biosurveillance for a median of three years.
- Emergency room (ER) data are most commonly used for biosurveillance (84 percent), followed by:
  - Outpatient clinic visits (49 percent)
  - Over-the-counter (OTC) medication sales (44 percent)
  - Calls to poison control centers (37 percent)
  - School absenteeism (37 percent)
- Two-thirds said they are 'highly' or 'somewhat' likely to expand the use of biosurveillance in the next two years.

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*

In spite of this widespread and potentially expanding use of biosurveillance, questions remain about its utility and effectiveness. Examples of papers that discuss various issues, challenges, and important research needs associated with effective implementation and operation of electronic biosurveillance systems include [2–11].

In case studies of health departments in eight U.S. states, Uscher-Pines *et al.* [4] found that fewer than half had written response protocols for responding to biosurveillance system alerts and the health departments reported conducting in-depth investigations on fewer than 15 percent of biosurveillance system alerts. Further, Uscher-Pines *et al.* [4] said, 'Although many health departments noted that the original purpose of syndromic surveillance was early warning/detection, no health department reported using systems for this purpose. Examples of typical statements included the following: 'I was a big supporter of syndromic surveillance for early warning early on, but now I am more realistic about the system's limitations.'

In the literature, Reingold [10] suggested that a compelling case for the implementation of biosurveillance systems has yet to be made. Cooper [12] said, 'To date no bio-terrorist attack has been detected in the United Kingdom, or elsewhere in the world using syndromic surveillance systems.' Stoto *et al.* [8] questioned whether biosurveillance systems can achieve an effective early detection capability. And Green [5] said, 'Syndromic surveillance systems, based on statistical algorithms, will be of little value in early detection of bioterrorist outbreaks. Early on in the outbreak, there will be cases serious enough to alert physicians and be given definitive diagnoses.'

The research challenges span many disciplines and problems:

- Legal and regulatory challenges to gain access to data.
- Technological challenges related to designing computer hardware and software for collecting and assembling data.
- Ethical and procedural issues inherent in managing and safeguarding data.
- Analytical challenges of assessing the likelihood of outbreaks and of displaying data to enhance SA.
- Managerial challenges of effectively assembling and operating the entire system.

Much of the continuing controversy surrounding biosurveillance stems from its initial focus on EED, a use that requires a number of still unproven assumptions, including:

- Leading indicators of outbreaks exist in pre-diagnosis health-related data of adequate strength such that they are statistically detectable with satisfactory power.
- The leading indicators occur sufficiently far in advance of clinical diagnoses so that, when found, they provide the public health community with enough advance notice to take action.
- Statistical detection algorithms exist that produce signals reliable enough to warrant continued dedication of public health resources to investigate the signals.

Of course, a myopic focus only on EED in biosurveillance systems misses important benefits that such systems can provide, particularly the potential to significantly advance and modernize the practice of public health surveillance. For example, whether or not biosurveillance systems prove effective at the early detection of bioterrorism, they are likely to have a significant and continuing role in the detection and tracking of seasonal and pandemic flu, as well as other naturally occurring disease outbreaks. This latter function is echoed in an Institute of Medicine report on Microbial Threats to Health by Smolinski, Hamburg, and Lederberg [13]: '[S]yndromic surveillance is likely to be increasingly helpful in the detection and monitoring of epidemics, as well as the evaluation of health care utilization for infectious diseases.' In a similar vein, Uscher-Pines *et al.* [4] quote a public health official: 'Health departments should not be at the mercy of alerts; they need to develop their own uses for syndromic surveillance.'

In terms of bioterrorism, Stoto [14] states that electronic biosurveillance systems build links between public health and health care providers—links that could prove to be critical for consequence management should a bioterrorism attack occur. Furthermore, Sosin [11] points out that electronic biosurveillance systems can act as a safety net, should the existing avenues of detection fail to detect an attack, hence countermeasures can be taken swiftly, and then can provide additional lead-time to public health authorities so that they can take more effective public health actions.

As a safety net, a biosurveillance system does not necessarily have to signal earlier than the first clinical diagnosis to be useful. To illustrate, a Dutch biologist conducting automated salmonella surveillance related that the surveillance system has detected outbreaks whose occurrence was somehow missed by sentinel physicians [15]. And unusual indicators in a biosurveillance system (not necessarily a signal from an early event algorithm) may give public health organizations time to begin organizing and marshalling resources in advance of a confirmed case and/or provide critical information about how and where to apply resources.

As someone who conducts research into statistical detection algorithms, I am not convinced by the argument that poor EED performance in existing systems means biosurveillance should not be used for EED. Rather, it suggests that more research is required to identify under what conditions biosurveillance EED can be effectively used, including those types of outbreaks that are detectable using available data and the types of algorithms that are more effective. And that was the purpose of the Centers for Disease Control and Prevention (CDC) and Agency for Toxic Substances and Disease

Published in 2011 by John Wiley & Sons, Ltd. *Statist. Med.* ***2011***, 30 403–415

Registry (ATSDR) Symposium short course: to point out methodological issues in biosurveillance that require additional research. The following section summarizes some of these issues.

2. Some issues in biosurveillance

2.1. Are statistical methods useful for EED?

A critical question that requires further research is under what circumstances statistical detection algorithms can usefully detect an outbreak. Specifically, an outbreak that is sufficiently large, geographically concentrated, or easy to diagnose, will likely be detected by a clinician prior to detection by a statistical algorithm. Similarly, a small or diffuse outbreak is also unlikely to be detected by a statistical algorithm faster than a clinician (assuming the algorithm has sufficient power to detect such an outbreak at all). That means statistical methods are of value only when the outbreak is large or concentrated enough to be detectable but not so large that the outbreak is obvious, combined with the situation where the type of outbreak is sufficiently hard to diagnose that medical professionals are likely to miss it for some time.

Figure 1 from [6] illustrates the point. The relevant research question is: For which diseases and under what medical conditions does the area within the dotted line exist? By 'exist' I mean that an outbreak is more likely to be detected by the statistical algorithm than by a medical or public health practitioner, as in both cases 'detection' is a stochastic event. The implication is that, for those types of outbreaks for which such an area does not exist, EED using biosurveillance systems should not be attempted.

Very little research has been published in this area. In an effort to explore the question, I conducted some simple simulations comparing the probability that a clinician diagnoses a bio-agent prior to a statistical algorithm signaling. The setup is as follows. A random number N of people present at an ER each day with flu-like symptoms, with $\mathbb{E}(N)=100$ and $\sigma_N=20$. In addition, k people present daily at the ER with flu-like symptoms but who have been exposed to a bio-agent, $0\leqslant k\leqslant 50$, where k is fixed for a given simulation run. With probability p those exposed to the bio-agent have extreme symptoms that a clinician can easily diagnose, meaning that if such a person goes to the ER the physician diagnoses the bio-agent with certainty. A biosurveillance system is monitoring the daily counts of people presenting with flu-like symptoms. Using a CUSUM algorithm, the threshold is set so that the average time between false signals is 30 days. The question is, for a particular combination of p and k, what is the probability that the clinician diagnoses a bio-agent case before the biosurveillance system signals?

Figure 2 shows the results of the simulation, where the surface is the probability that the clinician diagnoses a bio-agent case before the biosurveillance system signals for various combinations of p and k. The horizontal axis to the left is p, varying from 0.1 at the far left to 0. The horizontal axis to the right is for k, varying from 0 on the left to 50 at the far right. As expected, Figure 2 shows that as p gets smaller the clinician has less of a chance of detecting first and when p is larger the clinician has a higher probability. Similarly, for $k<8$ the statistical algorithm has a harder

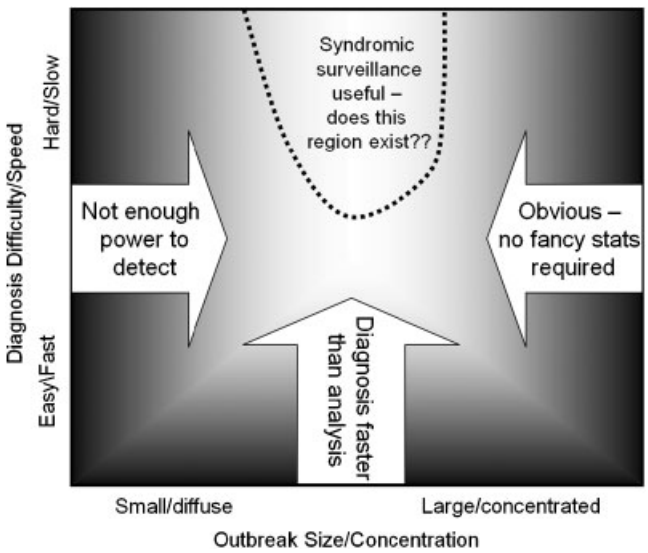

Figure 1. Additional research is required to determine under what conditions statistical detection algorithms can 'add value.' That is, for which types of outbreaks does the region defined by the dotted line exist? Figure reprinted from [6].

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 403–415

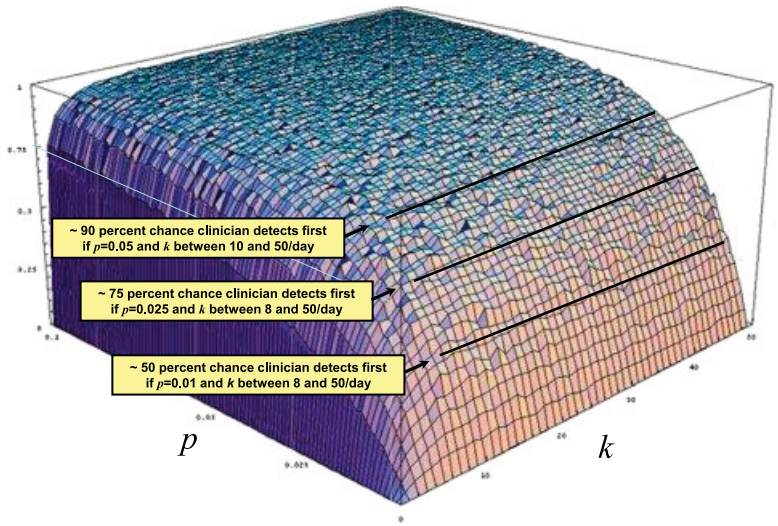

Figure 2. The probability that a clinician diagnoses a bio-agent case prior to a biosurveillance system signaling as a function of the probability of an extreme bio-agent case (p) and the daily number presenting who were exposed to the bio-agent (k).

time signaling before the clinician and, interestingly, as k increases beyond eight the detection algorithm gains little additional power.

What the plot shows is that the clinician has a 50 percent chance of diagnosing first when $p=0.01$ (one person in 100 exposed to the bio-agent develops obvious symptoms) and $8\leqslant k\leqslant 50$. In comparison, the clinician has a 75 percent chance of diagnosing first when $p=0.025$ (one person in 40) and $8\leqslant k\leqslant 50$. And, the clinician has a 90–95 percent chance of diagnosing first when $p=0.05$ (one person in 20) and $10\leqslant k\leqslant 50$. To me, the simulation suggests that there *is* a role for statistical algorithms in biosurveillance when the pathogen is hard to diagnose and/or when small numbers are presenting.

Yet, obviously, this is an overly simplified simulation. A more realistic simulation would allow k to be stochastic, it would allow the p for each individual exposed to the bio-agent to change (presumably increase) over time, etc. Further, as a reviewer pointed out, these simulations may be biased in favor of clinicians as it assumes perfect detection once a patient presents. In the real world the clinician would have to recognize the bioterrorism agent and then be sure enough of his or her finding that he or she would be willing to take action. This would take some time, particularly if it is required with some confirmatory tests, while the simulation assumed immediate diagnosis. And, of course, the clinician might misdiagnose.

Clearly this is an area ripe for further research, since the only other effort of which I am aware is [16] who conducted a very detailed simulation study comparing the performance of a syndromic surveillance system with clinical case findings for a release of inhalational anthrax in the Norfolk, Virginia area. With more research, these types of assessments would begin to definitively answer the question of when and how biosurveillance might be used for effective EED. For example, Buckeridge *et al.* [16] concluded:

> Our results suggest that syndromic surveillance could detect an inhalational anthrax outbreak before clinical case finding. However, we regularly observed a detection benefit only when syndromic surveillance operated at a specificity in the range of 0.9, which corresponds to 1 false alarm every 10 days. When operating at this relatively low specificity with a concomitant high sensitivity, syndromic surveillance detected outbreaks, on average, 1 day before clinical case finding did.

2.2. *Looking for everything means it is harder to find any one thing*

A significant challenge in designing detection algorithms for biosurveillance is that the type of outbreak to be detected is left undefined. Rather, the goal is to find any anomalous deviation from the usual incidence of a disease or syndrome, where 'any anomalous deviation' is not specified.

Consider an analogous situation in which you are given the picture in Figure 3 and asked to identify anything unusual. Clearly, there are a lot of potentially unusual things in the picture so that, without defining what 'unusual' means, one could flag any number of possible items. Contrast that with the more specific goal of identifying whether Osama bin Laden is present in the picture. A more focused goal like this improves detection in two ways: (1) it reduces false positives, as we can now rule out signaling on other unusual events and (2) it allows us to more precisely focus our

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 403–415

Figure 3. Is there anything unusual in this picture? [17].

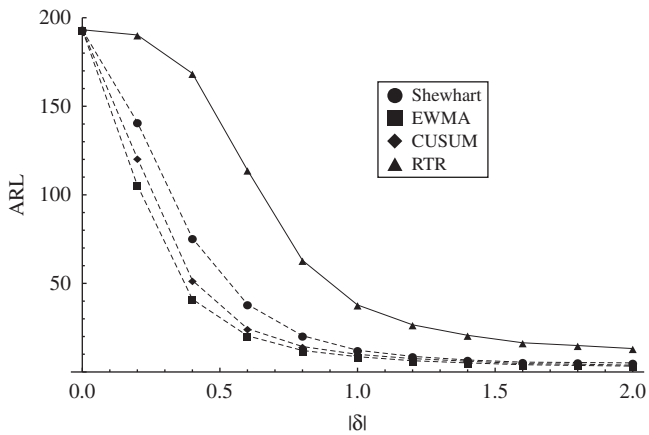

Figure 4. Comparison of expected times to signal for a shift from $f_0 = N(0,1)$ to $f_1 = N(\delta,1)$, $0 \leqslant \delta \leqslant 2$, for four detection algorithms (Shewhart, EWMA, CUSUM, and RTR). See [18] for additional details.

search, potentially making it more efficient and/or effective at detecting bin Laden. (Did you find him? Of course, being more specific in the anomalies to be detected does not imply that the detection will then be easy.)

In hypothesis testing terms, it is as if one were given two sets of data, with the null and alternative hypotheses left undefined, and asked to evaluate whether there is anything unusual in one data set compared with the other. One approach to such a problem is to conduct a nonparametric test of the hypothesis that the two sets of data come from the same distribution, but such a test has less power to detect a particular difference between the two sets of data, such as that the mean of one is greater than the mean of the other (assuming a difference in the means is the desired alternative to be detected). Furthermore, note that even with the nonparametric test, we *are* testing a particular hypothesis. What if it is the wrong one?

Figure 4 gives a more relevant example drawn from [18]. It shows the expected time to signal ('ARL') for four detection schemes based on the Shewhart, EWMA, CUSUM, and RTR algorithms, where the data initially comes from an $N(0,1)$ distribution that subsequently experiences a mean shift to various alternative distributions $N(\delta,1)$, $|\delta| \geqslant 0$. At $\delta = 0$ it shows that the expected time to false signal was set equal for all four schemes. For $\delta > 0$, more effective detection methods have shorter expected times to signal.

As implemented in this simulation, the EWMA was designed to look only for mean shifts. In contrast, the Shewhart and CUSUM algorithms are ensemble schemes designed to look for both mean and variance shifts. And the RTR is a nonparametric methodology designed to look for any distributional change. The result, as the figure shows, is that the EWMA has the shortest expected time to signal for any $\delta > 0$, followed by the CUSUM, then the Shewhart, and then the

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 403–415

RTR. That is, the EWMA, which is only looking for a mean shift, is more effective when such a shift occurs. In contrast, the RTR is looking for any type of distributional change and hence it is less effective than the other three methods.

For biosurveillance systems it follows that performance can be improved if the events to be detected can be better defined. Dembek *et al.* [19] identified 11 'clues to a deliberate epidemic:'

- A highly unusual event with large numbers of casualties
- Higher morbidity or mortality than expected
- Uncommon disease
- Point-source outbreak
- Multiple epidemics
- Lower attack rates in protected individuals
- Dead animals
- Reverse or unnatural spread
- Unusual disease manifestation
- Downwind plume pattern
- Direct evidence.

This then suggests, for example, that instead of looking for any increase in the rate of illness across many different syndromes, one might look for a significant increase in a particular syndrome *in conjunction with* a significant increase in mortality *or* evidence of an unusual clustering of those with the syndrome. As with the 'Where's Osama?' picture, appropriately defining the alternatives will reduce false positives while making the signals themselves more meaningful.

Of course, as a reviewer pointed out, being overly specific in what a biosurveillance system is looking for is not without its potential problems. In particular, in a bioterrorism setting where there is an intelligent adversary, if the terrorists have knowledge of the precise signal being looked for they will likely use this information to their advantage and mount an attack with a different signal signature. This type of adversarial back and forth is currently most evident in airport passenger screening for detecting bombs and other weapons, but it applies similarly to the application of biosurveillance to detect bioterrorism. This does not negate the previous point that more precisely defining the signals to be detected will increase the sensitivity and specificity of biosurveillance systems. But it does suggest that the signals of interest should not be made public and that biosurveillance systems need to be flexible so that they can adjust to changing terrorist capabilities and tactics.

2.3. Existing methods do not apply directly to biosurveillance

Epidemiology traces its roots back to Dr John Snow who used statistical mapping techniques to help identify the cause of a cholera outbreak in London in the mid-19th century. Since that time, the field has developed many and varied tools for understanding factors affecting the health and illness of populations. Epidemiologists and public health professionals are often called upon to determine the cause or causes of a particular disease outbreak, much as Dr Snow first did almost two centuries ago. A defining feature of such an investigation is that the outbreak has already been identified. The investigation is thus a *retrospective* exercise in trying to determine the cause of a particular outbreak.

In contrast, biosurveillance is a *prospective* exercise in monitoring populations for possible disease outbreaks. The goal is to routinely evaluate data for evidence of an outbreak prior to the existence of a confirmed case (in fact, even prior to any suspicion of an outbreak) and, as a result, the epidemiological tools and techniques developed over the past two centuries generally do not apply to the biosurveillance problem. Public health practitioners have thus turned to the field of industrial quality control, sometimes applying and adapting those tools directly to biosurveillance.

Industrial quality control traces its roots to 1931 when Walter A. Shewhart wrote *Economic Control of Quality of Manufactured Product*, where he developed the concept of the control chart, a graphical statistical tool most commonly used to control manufacturing processes. The success of Shewhart's method in the industrial world lies in its simplicity. Essentially one establishes control limits and as long as a statistic derived from the data, sequentially observed over time, falls within the control limits the manufacturing process is assumed to be 'in control.' If one or more fall outside the control limits then the process is examined to determine whether it is 'out of control' and requires adjustment. Shewhart's work gave rise to the field of *statistical process control* (SPC) and a large and still growing literature of research into myriad statistical methods for controlling processes.

In industrial quality control, and thus SPC, it is often reasonable to assume that:

- as one controls the manufacturing process, the in-control distribution is (or can reasonably be assumed to be) stationary;
- observations can be drawn from the process so they are independent (or nearly so);
- monitoring the process mean and standard deviation is usually sufficient;

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. **2011**, 30 403–415

- the asymptotic distributions of the statistics being monitored are known and thus can be used to design appropriate control charts;
- shifts, when they occur, remain until they are detected and corrective action taken; and
- temporal (as opposed to spatial) detection is the critical problem.

However, the general biosurveillance problem violates many, if not all, of these assumptions. For example:

- there is little to no control over disease incidence and thus the distribution of disease incidence is usually non-stationary;
- observations (often daily counts) are autocorrelated, and the need for quick detection works against the idea of taking measurements far apart to achieve (near) independence;
- in biosurveillance there is little information about what types of statistics are useful for monitoring—one is often looking for anything that seems unusual;
- because individual observations are being monitored, the idea of asymptotic sampling distributions does not apply, and the data often contain significant systematic effects that must be accounted for;
- outbreaks are transient, with disease incidence returning to its original state once an outbreak has run its course; and
- identifying both spatial and temporal deviations are often critical.

This gap between existing SPC methods and the biosurveillance problem provides an area ripe for new research and, indeed, quite a bit has been going on for the past decade or so. However, frequently epidemiologists and biostatisticians have attempted to re-invent existing SPC methodology or have failed to incorporate decades of lessons learned from the SPC literature. Perhaps this is because the research cultures and norms of medical and SPC communities have hindered research collaboration. For example, the medical journals and literature emphasize evaluating algorithm performance on the actual data whereas the SPC community emphasizes extensive simulation. Similarly, epidemiologists and biostatisticians are steeped in the medical hypothesis testing language of sensitivity and specificity whereas SPC researchers generally use measures of expected time to signal for assessing algorithm performance. The result is that biosurveillance research done by medical researchers is largely unpublishable in the SPC journals and vice versa.

Yet, biosurveillance research would greatly benefit from collaboration between the two communities. And SPC research would certainly benefit from advances in biosurveillance monitoring as there are manufacturing processes that have characteristics similar to the biosurveillance problem (e.g. continuous monitoring of chemical processes). The challenge to the research community is to find ways to break down the barriers to collaboration. This has been happening more recently, with a number of SPC researchers in attendance at the 12th Biennial CDC & ATSDR Symposium, but finding ways to encourage greater collaboration will facilitate advances in biosurveillance.

2.4. Algorithm performance evaluations need to be standardized and expanded

This leads to the next issue, which is that the lack of commonly accepted standards for evaluating detection algorithms creates major impediments to research progress, including: (1) biosurveillance algorithm performance evaluations published in the literature cannot be replicated, (2) the results often cannot be generalized to other settings, and (3) the various results cannot easily be compared with one another. Published results often cannot be replicated or generalized because many of the evaluations are based on demonstrating performance on a single data set that is usually not publicly available. Performance comparisons are hindered by the same factors and compounded by a lack of consensus on the appropriate performance metrics. Improvement is necessary in (at least) three areas: (1) the use of simulation, (2) standardization of evaluation metrics, and (3) encouraging those proposing new methods to make performance comparisons with existing competing methods.

2.4.1. Using simulation to compare performance. As the performance of most proposed methods is demonstrated on data that are not publicly available, it is very difficult and often impossible to compare the performance of the various detection methods across the biosurveillance literature. This is often driven by the public health community's desire to see methods demonstrated on real data. Yet, precisely because the data are real, there is a lack of general availability of such data to the research community due to confidentiality and privacy concerns.

One solution is to make real data more widely available. Shmueli and Burkom [3] say, 'Currently syndromic data are only available to researchers affiliated with a particular biosurveillance system or research group, for reasons of data confidentiality and non-disclosure agreements. This a major obstacle in the way of scientific progress in both temporal and spatio-temporal biosurveillance, and hopefully some data will be made available to academic researchers.'

However, seemingly unrecognized in this discussion is the fact that any real data set is simply one realization of a stochastic process. Focusing only on a particular stream of data one fails to recognize and account for the full randomness

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 403–415

of the underlying phenomenon. As per Fraker *et al.* [20]:

> Evaluations and comparisons of statistical performance in public health surveillance often involve the use of real surveillance over a past time period of interest. The outbreak locations in time are either assumed to be known or outbreaks are artificially superimposed on the data. As pointed out by Woodall [21], this is rarely, if ever, the case in the industrial literature where case study-type data are used only to illustrate the application of methods, not to evaluate statistical performance.

In addition, even if some real data are made available, it will provide little to no information about what outbreaks look like, particularly those associated with bioterrorism related events. The challenge, as Rolka *et al.* [22] have said, '… is to develop improved methods for evaluating detection algorithms in light of the fact that we have little data about outbreaks of many potential diseases that are of concern.'

Simulation is an alternative. Of course, it is very difficult to (stochastically) characterize, and thus simulate, all the detailed features characteristic of the normal or baseline state of disease incidence, as well as the various outbreak conditions. However, one could also make similar statements about industrial quality control problems. Yet that field, over time, has come to use various data abstraction conventions that facilitate simulation and, as a result, allow comparisons between methods and across the literature. As Rolka *et al.* [23] said, 'Reliance on the use of Monte Carlo simulation in the field of Statistics is well known. It has been this author's experience that the technique is undervalued in the field of Public Health because it has previously not been required.' Monte Carlo simulation can:

- facilitate evaluating algorithms across many scenarios;
- eliminate unneeded/distracting real world complexities;
- allow clean and clear comparisons of algorithms; and
- make it easier to get at generalizable conclusions/results.

Developing appropriate simulation conventions will not be easy. The biosurveillance problem is complicated and not well defined. Breaking out of the 'my data are unique' and 'only real data are valid' paradigms also will not be easy. Furthermore, it is important to *abstract* the most important features of the problem to facilitate comparisons and, in particular, being able to distinguish when and why various methods perform differently. Agreeing on what the 'important features' of biosurveillance data will be challenging. However, it can be done. For examples of efforts in this direction, see [24, 25].

2.4.2. Common performance metrics. The biosurveillance research community must settle on a commonly accepted set of metrics for evaluating detection algorithm performance. Only then will it be possible to synthesize across the literature to better understand which algorithms work better or worse under particular conditions. At issue, as Fraker *et al.* [20] say, is that '[s]ubstantially more metrics have been proposed in the public health surveillance literature than in the industrial monitoring literature' and no consensus has yet arisen about which metrics are to be preferred. Proposed metrics include:

- Sensitivity, specificity, and timeliness
- Sensitivity and predictive value positive
- Recurrence interval
- Expected delay and conditional expected delay (CED)
- Probability of successful detection (PSD)
- Area under the response operating characteristic (ROC) curve, activity monitoring operating characteristic (AMOC) curve, and free response operating characteristic (FROC) curve
- Average run length (ARL), average overlapping run length (AORL), average time to signal given an outbreak, average time between signal events (ATBSE), and average signal event length (ASEL)

In comparison, the SPC community almost exclusively uses the expected time to signal metrics to evaluate and compare the algorithmic performance. I do not suggest that the ARL is appropriate for biosurveillance, but rather use it to illustrate that another community has settled on a common measure for the benefit of that research. Unfortunately, there is little similar consensus in the biosurveillance. See [3] for additional discussion.

2.4.3. More evaluation comparisons needed. As Woodall *et al.* [26] said, 'The body of literature on health-related surveillance is smaller than that on industrial surveillance, and is somewhat less mathematical in nature.' Perhaps more importantly, the trend in the biosurveillance detection algorithm literature is to propose a new method and illustrate its performance on a set of data that is not publicly available *without also comparing its performance with other commonly accepted methods*. We have already discussed how using a private data set hinders the community from being able to generalize results across the literature. Not requiring individual authors to compare and contrast a new method's performance with existing methods exacerbates the problem.

Published in 2011 by John Wiley & Sons, Ltd. *Statist. Med.* ***2011***, 30 403–415

In a search of the biosurveillance literature, very few papers were found that conducted such comparisons. While I surely missed some, the only papers I found were [23–25, 27–33]. That is only 10 papers. Even if I missed an equivalent amount in my (somewhat cursory) search, that is substantially fewer than the various papers proposing new methods, which means there has been relatively little comparison between methods, at least as published in the literature.

To again compare with the SPC community, one essentially cannot publish a paper on a new methodology in that literature without an extensive set of (simulation) comparisons with the standard methods. Indeed, the literature encourages such comparisons and one can even publish papers that simply compare existing methods under new conditions. Our understanding of biosurveillance algorithm performance would be immeasurably improved by encouraging a whole raft of such publications, even if they were done on proprietary data sets that could not be publicly released.

2.5. Monitoring for bioterrorism versus natural diseases: it is not necessarily the same

One benefit of biosurveillance systems designed for bioterrorism detection is that they can also be used to detect and monitor natural diseases. However, it does not follow that a system operated for detecting natural diseases will be most effective at detecting bioterrorism.

Perhaps Shmueli and Burkom [3] said it best:

> Determining whether an abnormality is present in the data requires defining normal behavior. One complication arises from the intended dual-use of biosurveillance systems for detecting natural and bioterror-related or pandemic disease outbreaks, because the data footprint of a seasonal influenza epidemic is a target signal in the former context but part of the background clutter in the latter. In the bioterrorism monitoring context, all usual seasonal influences should be removed for sensitivity even at the peak and aftermath of a usual influenza outbreak.

At issue is that if a system that is signaling during a natural disease outbreak is not 're-set,' then it cannot detect a bioterrorism attack during that time. An analogy: the smoke alarm that goes off every time the oven is used is of little use detecting a stove top fire when you are baking a cake.

This suggests that additional emphasis needs to be placed on operational paradigms and algorithm design. In particular, during natural disease outbreaks systems must be designed to adjust for revised background incidence rates so that system can look for further anomalies. In addition, the current practice of not resetting algorithms after a signal and, in fact, using multiple sequential signals as evidence of an actual outbreak, should be recognized as practices detrimental to detecting a bioterrorism attack.

Of course, this begs the question of whether it is even possible to adjust for a transient disease outbreak in any reliable fashion that would then allow for further anomaly detection. If the main purpose of biosurveillance systems is bioterrorism protection, then this is an important area of future research. After all, when would a smart bioterrorist attack? During the flu season, of course.

2.6. Research into biosurveillance system design needed

Biosurveillance systems need to be designed as just that, *systems*. For example, a typical characteristic of biosurveillance systems is that the data collection locations (generally hospitals and clinics) are in fixed locations that may or may not correspond to a particular threat of either natural disease or bioterrorism. In order to provide comprehensive population coverage, biosurveillance system operators are inclined to enlist as many hospitals and clinics as possible. However, as the sources and types of data being monitored proliferate in a biosurveillance system, then so do the false positive signals from the systems. Indeed, false positives have become an epidemic problem for some systems. As one researcher [34] said, '…most health monitors…learned to ignore alarms triggered by their system. This is due to the excessive false alarm rate that is typical of most systems—there is nearly an alarm every day!'

Furthermore, if the main purpose of biosurveillance is to guard against bioterrorism, then one should design algorithms that can incorporate information about the adversary, particularly information about likely modes and locations of attack. For example, Fricker and Banschbach [35] have developed a methodology for optimizing a (very simple) threshold detection-based biosurveillance system that uses information about the probability of attack by location. In so doing, they maximize the system-wide probability of detecting an 'event of interest' subject to a constraint on the expected number of false signals by differentially setting detection thresholds at each location in accordance with the probability that each location is attacked.

Using this or some similar methodology would allow public health officials to 'tune' their biosurveillance systems to optimally detect various threats while explicitly accounting for organizational resource constraints available for investigating and adjudicating signals. This would then allow practitioners to focus their public health surveillance activities on locations or diseases that pose the greatest threat at a particular point in time. And, as the threat changes, using the same hospitals and clinics, the system could then subsequently be tuned to optimally detect other threats. With this approach large biosurveillance systems are an asset (as opposed to a burden because of excessive alarms).

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 403–415

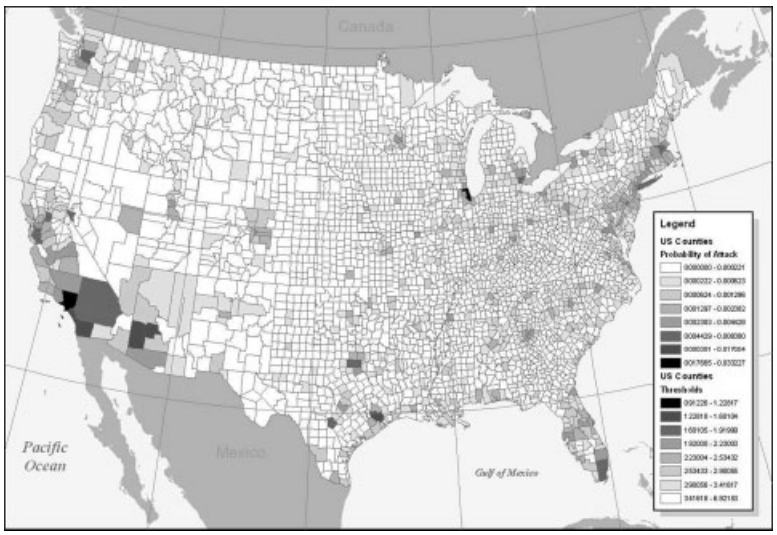

Figure 5. Map of the counties in the contiguous continental United States with their associated optimal thresholds and associated probabilities of attack for the hypothetical example.

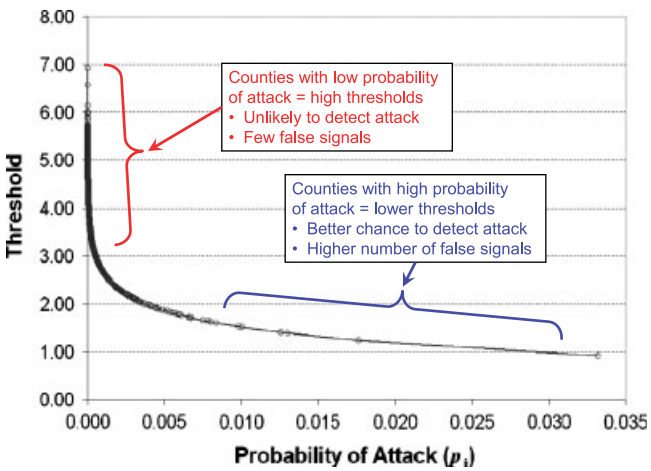

Figure 6. Plot of the optimal thresholds against probability of attack for the hypothetical example.

In an example, Fricker and Banschbach [35] show how one might optimally set thresholds on a hypothetical biosurveillance system designed to simultaneously monitor all 3141 counties in the United States. For the purposes of illustration, they used the proportion of the total population in a county as a surrogate for the probability that the county is attacked. Assuming a maximum expected number of false signals of four per day (since this national system was assumed to have resources sufficient to investigate four false positives per day), the system has an expected time to detect an attack of three days. This is achieved with thresholds ranging from 0.91 for Los Angeles county, the county with the highest probability of attack, to 6.92 for Loving county, the county with the lowest probability of attack. Figure 5 is a map showing the probability of attack and thresholds by county.

Now, consider the system performance if one were to have used a common threshold for all the counties of 3.018, which achieves the same expected number of false signals (four per day), the expected time to detection would more than double to $6\frac{1}{2}$ days. This decrease in sensitivity occurs because the system is less able to detect an attack in those locations most likely to be attacked. Conversely, setting a common threshold of 2.433 to achieve an expected time to detect of three days results in an almost *sixfold increase* in the expected number of false signals to 23.5 per day.

Figure 6 shows a plot of the optimal thresholds versus the probability of attack for all of the counties. What the plot shows is that the counties with a high probability of attack have low thresholds. These low thresholds make false positives in those locations more likely, but they also make it much more likely that an actual attack will be readily detected. Conversely, the counties with very low probabilities of attack have high thresholds making it very difficult to detect an attack but, on the other hand, these counties are being monitored at a level consistent with their risk of attack. That is,

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 403–415

the optimization has made the necessary trade-off of probability of detection against the likelihood of false signals to maximize the probability of detecting an attack somewhere in the country within a manageable false signal rate.

Much more work needs to be done in this area.

3. Discussion and recommendations

In the short course at the 12th Biennial CDC & ATSDR Symposium and in this paper I have attempted to lay out various biosurveillance methodological challenges. The goal was to discuss some current issues in biosurveillance, particularly with respect to detection algorithms. My viewpoint and suggestions are based on an industrial SPC background and, in my opinion, biosurveillance research has yet to fully tap the SPC literature and expertise. In addition, other disciplines have much to offer to the biosurveillance problem as well:

- *Operations research*: Optimizing biosurveillance system performance is a non-trivial problem.
- *Systems engineering*: Biosurveillance systems are complex systems that require careful design.
- *Game theory*: In a bioterrorism context, there is an autonomous, willful adversary who makes specific choices about when and where to attack.

This paper is not meant to be critical of the current state of the art, which has come a long way in the past decade, but rather to provide something of an organized set of issues as a potential research 'road map' for the community. In so doing, I have posed more problems than solutions, highlighting some of the open issues, including the:

- benefits of better specifying the events to be detected;
- lack of standard evaluation methods and metrics in the literature;
- utility of using more Monte Carlo methods for algorithm evaluation; and
- need to take a systems-design approach to improve EED performance.

In this paper, I focused quite a bit on the need for more and better comparisons between detection methods, and I have suggested that simulation has a role to play in such comparisons. At issue is that the sheer plethora of methods that have been proposed in the literature raise questions that are, thus far, largely unanswered:

- Under what conditions do the various detection algorithms work best?
- Which methods are most sensitive at detecting a particular type of outbreak?
- Are some methods less effective with certain types of background data?
- Etc.

To answer these and other questions, I recommend encouraging on-going research that conducts comparisons between methods under various conditions, both on real and simulated data. In a related vein, the community should promote research into stochastically characterizing data (both normal background and outbreak) so that simulated data can be made as realistic as possible. As a referee pointed out, an ancillary benefit is that 'if the stochastic structure of the data is understood (e.g. Autoregressive Integrated Moving Average (ARIMA) model) then the data can often be transformed to yield independent data.' This alone would go a long way in bridging the gap between traditional SPC methods and their application to the biosurveillance problem.

In addition, in my opinion, competitions (e.g. Defense Advanced Research Projects Agency (DARPA)-sponsored Bio-Advanced Leading Indicator Recognition Technology (ALIRT) competition, 2001–2004) are of limited utility. The biosurveillance problem does not lend itself to a single 'solution' arising from a competition using one set of real data since the best performer on one particular data set does not mean that it will be the best under other conditions or that the results are generalizable.

Further, to be able to effectively synthesize results from these comparisons, the community needs to begin using a common set of metrics. To that end, I suggest convening a panel of experts akin to (or perhaps exactly) a National Academies panel to conduct the requisite research and produce a report with authoritative recommendations on the metrics that the community should use.

Returning to the original question of whether statistical methods are useful for EED, I suggest that we really do not know yet. That is, while some current systems have given EED a 'bad rap' because of an excessive number of alarms, whether the systems and their associated detection algorithms can be modified so that they appropriately minimize false positive signals while maintaining sufficient sensitivity to actual outbreaks/attacks is still an open question. Certainly the simple simulation presented in Section 2.1 and the work by Buckeridge *et al.* [16] suggest that biosurveillance for EED has a role in some situations:

- As a primary detection tool for rare, hard to diagnose diseases/agents.
- As a back-up to clinicians for moderately sized outbreaks that are moderately hard to diagnose.

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 403–415

Further rigorous, scientific studies are still required to clearly define and refine that role, as well as the limitations of biosurveillance.

Acknowledgements

I would like to thank an anonymous reviewer for a number of excellent points that improved the paper. I would also like to thank the CDC, particularly Henry Rolka, Myron Katzoff, Wendy Wattigney and the 2009 Symposium Committee, for the opportunity to present the short course at 12th Biennial CDC and ATSDR Symposium. Thanks also to Kathy O'Connor for editing this special issue and for all her help in coordinating the discussion papers. Finally, I would be remiss if I did not acknowledge my debt to the broader community of biosurveillance researchers and practitioners from whom I have benefitted over the years. Any errors or omissions and all opinions are purely my own.

References

1. U.S. Government, Homeland Security Presidential Directive 21: Public Health and Medical Preparedness, 2007. Accessed on-line at: http://www.dhs.gov/xabout/laws/ gc_1219263961449.shtm on July 15, 2009.
2. Buehler JW, Sonricker A, Paladini M, Soper P, Mostashari F. Syndromic surveillance practice in the United States: findings from a survey of state, territorial, and selected local health departments. *Advances in Disease Surveillance* 2008; **6**:(3):1–20.
3. Shmueli G, Burkom HS. Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *Technometrics* (Special Issue on Anomaly Detection). Accessed on-line at: www.rhsmith.umd.edu/faculty/gshmueli/web/images/statchallengesbiosurveillancerevised-iii.pdf on 27 September 2009.
4. Uscher-Pines L, Farrell CL, Babin SM, Cattani J, Gaydos CA, Hsieh Y, Moskal MD, Rothman RE. Framework for the development of response protocols for public health syndromic surveillance systems: case studies of 8 US states. *Disaster Medicine and Public Health Preparedness* 2009; **3**:S29–S36.
5. Green M. Syndromic surveillance for detecting bioterrorist events—the right answer to the wrong question? Presentation given at the Naval Postgraduate School, Monterey, CA, 9 June 2008.
6. Fricker Jr RD, Rolka HR. Protecting against biological terrorism: statistical issues in electronic biosurveillance. *Chance* 2006; **19**:4–13.
7. Rolka HA. Data analysis research issues and emerging public health biosurveillance directions. In *Statistical Methods in Counterterrorism*: *Game Theory*, *Modeling*, *Syndromic Surveillance*, Wilson A, Wilson G, Olwell DH (eds). Springer: Berlin, 2006; 101–107.
8. Stoto MA, Schonlau M, Mariano LT. Syndromic surveillance: is it worth the effort? *Chance* 2004; **17**:19–24.
9. Bravata DM, McDonald KM, Smithe WM, Rydzak C, Szeto H, Buckeridge DL, Haberland C, Owens DK. Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Annals of Internal Medicine* 2004; **140**(11):910–922.
10. Reingold A. If syndromic surveillance is the answer, what is the question? *Biosecurity and Bioterrorism*: *Biodefense Strategy*, *Practice*, *and Science* 2003; **1**:1–5.
11. Sosin DM. Syndromic surveillance: the case for skillful investment view. *Biosecurity and Bioterrorism*: *Biodefense Strategy*, *Practice*, *and Science* 2003; **1**:247–253.
12. Cooper DL. Can syndromic surveillance data detect local outbreaks of communicable disease? A model using a historical Cryptosporidiosis outbreak. *Epidemiology and Infection* 2006; **134**:13–20.
13. Smolinski MS, Hamburg MA, Lederberg J. *Microbial Threats to Health*: *Emergence*, *Detection*, *and Response*. National Academic Press: Washington, DC, 2003.
14. Stoto MA. Syndromic surveillance in public health practice. Presentation to Institute of Medicine Forum on Microbial Threats, Washington, DC, December 12, 2006.
15. Burkom H. Personal Communication. December 22, 2006.
16. Buckeridge DL, Owens DK, Switzer P, Frank J, Musen MA. Evaluating detection of an inhalational anthrax outbreak. *Emerging Infectious Diseases* 2006; **12**(12):1942–1949.
17. wht3.com. Where's Osama? Accessed on-line at: http://wht3.com/Osama1.html. on September 28, 2009.
18. Fricker Jr RD, Chang JT. The repeated two-sample rank (RTR) procedure: a nonparametric multivariate individuals control chart, in draft. Available at: http://faculty.nps.edu/rdfricke/frickerpa.htm [10 July 2009].
19. Dembek ZF, Kortepeter MG, Pavlin JA. Discernment between deliberate and natural infectious disease outbreaks. *Epidemiology and Infection* 2007; **135**:353–371.
20. Fraker SE, Woodall WH, Mousavi S. Performance metrics for surveillance schemes. *Quality Engineering* 2008; **20**:451–464.
21. Woodall WH. Use of control charts in health-care and public-health surveillance (with Discussion). *Journal of Quality Technology* 2006; **38**:89–104.
22. Rolka H, Burkom H, Cooper GF, Kulldorff M, Madigan D, Wong W. Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs. *Statistics in Medicine* 2007; **26**:1834–1856.
23. Rolka H, Bracy D, Russell C, Fram D, Ball R. Using simulation to assess the sensitivity and specificity of a signal detection tool for multidimensional public health surveillance data. *Statistics in Medicine* 2005; **24**:551–562.
24. Fricker Jr RD, Hegler BL, Dunfee DA. Comparing biosurveillance detection methods: EARS' versus a CUSUM-based methodology. *Statistics in Medicine* 2008; **27**:3407–3429.
25. Fricker Jr RD, Knitt MC, Hu CX. Directionally sensitive MCUSUM and MEWMA procedures with application to biosurveillance. *Quality Engineering* 2008; **20**:478–494.
26. Woodall WH, Grigg OA, Burkom HS. Research issues and ideas on health-related monitoring. In *Frontiers in Statistical Quality Control 9*, Lenz JJ, Wilrich P-Th (eds). Springer: Berlin, 2006; 101–107.
27. Fricker Jr RD. Directionally sensitive multivariate statistical process control methods with application to syndromic surveillance. *Advances in Disease Surveillance* 2007; **3**. Available on-line at: www.isdsjournal.org.

28. Groenewold MR. Comparison of two signal detection methods in a coroner-based system for near real-time mortality surveillance. *Public Health Reports* 2008; **122**:521–530.
29. Stoto MA, Fricker Jr RD, Jain A, Diamond A, Davies-Cole JO, Glymph C, Kidane G, Lum G, Jones L, Dehan K, Yuan C. Evaluating statistical methods for syndromic surveillance. In *Statistical Methods in Counterterrorism*: *Game Theory*, *Modeling*, *Syndromic Surveillance*, *and Biometric Authentication*, Wilson A, Wilson G, Olwell DH (eds). Springer: Berlin, 2006; 141–172.
30. Hutwagner LC, Thompson WW, Seeman GM, Treadwell T. A simulation model for assessing aberration detection methods used in public health surveillance systems with limited baselines. *Statistics in Medicine* 2005; **24**:543–550.
31. Hutwagner LC, Browne T, Seeman GM, Fleischauer AT. Comparing aberration detection methods with simulated data. *Emerging Infectious Diseases* 2005; **11**:314–316.
32. Rogerson PA, Yamada I. Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine* 2004; **23**:2195–2214.
33. Siegrist D, Pavlin J. Bio-ALIRT biosurveillance detection algorithm evaluation. *Morbidity and Mortality Weekly Report* 2004; **53**(supplement): 152–158.
34. Shmueli G. Accessed on-line at: https://wiki.cirg.washington.edu/pub/bin/view/Isds/ SurveillanceSystemsInPractice. on October 8, 2008.
35. Fricker Jr RD, Banschbach D. Optimizing biosurveillance systems that use threshold based event detection methods. *Information Fusion* 2010; in press. Available online at: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W76-4Y35TKC-1&_user=3326500&_coverDate=01%2F04%2F2010&_alid=1215071323&_rdoc=1&_fmt=high&_orig=search&_cdi=6618&_sort=r&_docanchor=&view=c&_ct=1&_acct=C000060280&_version=1&_urlVersion=0&_userid=3326500&md5=b92f4583a90bd68c95b10f9230f37353.