STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:1834–1856
Published online 15 January 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2793

Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: Research needs[‡]

Henry Rolka[1,*,†], Howard Burkom[2], Gregory F. Cooper[3], Martin Kulldorff[4], David Madigan[5] and Weng-Keen Wong[3]

[1] *Centers for Disease Control and Prevention (CDC), Division of Emergency Preparedness and Response, National Center for Public Health Informatics, 1600 Clifton Rd., NE. MS D45, Atlanta, GA 30333, U.S.A.*
[2] *National Security Technology Department, Johns Hopkins University Applied Physics Laboratory, MD, U.S.A.*
[3] *Center of Biomedical Informatics, Division of General Internal Medicine, University of Pittsburgh, PA, U.S.A.*
[4] *Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, MA, U.S.A.*
[5] *Department of Statistics, Rutgers University, NJ, U.S.A.*

SUMMARY

The objective of this report is to provide a basis to inform decisions about priorities for developing statistical research initiatives in the field of public health surveillance for emerging threats. Rapid information system advances have created a vast opportunity of secondary data sources for information to enhance the situational and health status awareness of populations. While the field of medical informatics and initiatives to standardize healthcare-seeking encounter records continue accelerating, it is necessary to adapt analytic and statistical methodologies to mature in sync with sibling information science technologies. One major right-of-passage for statistical inference is to advance the optimal application of analytic methodologies for using multiple data streams in detecting and characterizing public health population events of importance. This report first describes the problem in general and the data context, then delineates more specifically the practical nature of the problem and the related issues. Approaches currently applied to data with time-series, statistical process control and traditional inference concepts are described with examples in the section on Statistics and the Role of the Analytic Surveillance Data Monitor. These are the techniques that are providing substance to surveillance professionals and enabling use of multiple data streams. The next section describes use of a more complex approach that takes temporal as well as spatial dimensions

[*] Correspondence to: Henry Rolka, Centers for Disease Control and Prevention (CDC), Division of Emergency Preparedness and Response, National Center for Public Health Informatics, 1600 Clifton Rd., NE. MS D45, Atlanta, GA 30333, U.S.A.

[†] E-mail: HRolka@cdc.gov, hrr2@cdc.gov

[‡] The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Contract/grant sponsor: National Science Foundation; contract/grant number: IIS-0325581
Contract/grant sponsor: Department of Homeland Security; contract/grant number: F30602-01-2-0550
Contract/grant sponsor: Pennsylvania Department of Health; contract/grant number: ME-01-737

Received 15 October 2006
Accepted 16 November 2006

into consideration for detection and situational awareness regarding event distributions. The space–time statistic has successfully been used to detect and track public health events of interest. Important research questions which are summarized at the end of this report are described in more detail with respect to the methodological application in the respective sections. This was thought to help elucidate the research requirements as summarized later in the report. Following the description of the space–time scan statistical application; this report extends to a less traditional area of promise given what has been observed in recent application of analytic methods. Bayesian networks (BNs) represent a conceptual step with advantages of flexibility for the public health surveillance community. Progression from traditional to the more extending statistical concepts in the context of the dynamic status quo of responsibility and challenge, leads to a conclusion consisting of categorical research needs. The report is structured by design to inform judgment about how to build on practical systems to achieve better analytic outcomes for public health surveillance. There are references to research issues throughout the sections with a summarization at the end, which also includes items previously unmentioned in the report. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: applied statistics; public health bioterrorism surveillance; multiple data streams; biosurveillance

INTRODUCTION

A variety of analytic approaches have arisen and are in use for performing bioterrorism (BT) surveillance using social and other public health indicators from various types of data (e.g. pre-diagnostic/chief complaint ambulatory care encounters, nurse call line data, laboratory test orders, over-the-counter (OTC) sales, absenteeism, Emergency Department (ED) discharge summaries, prescription pharmaceutical sales, 911-emergency calls, etc.). The value of data anomaly investigation and signal detection as technologies in surveillance can be enhanced by more formalized application of data pre-processing methods and applied probabilistic decision science concepts and principles. A full characterization of the usefulness as well as corresponding development of analytic methods for exploiting opportunistic data are rich areas for research, especially in the context of information system integration. The specific focus of this report is on the analytic surveillance component where *multiple sources* of data are utilized to assess the health-related temporal and geographic status of human health risk. The purpose of the report is to provide (1) an overview of the problems, (2) a detailed view of current practices in operation, (3) evolving application areas of promise, and (4) to identify priority topics for a research agenda to address this area of application.[§]

Data issues

Data sources in use, or those which may potentially be used, for public health surveillance range in ability to indicate population events; early or otherwise [1]. There have been attempts to characterize their relative value generally along the lines of timeliness and for their ability to accurately represent true population events and generally provide for actionable situational awareness [2–7]. There are

[§] The concept for this report was promulgated by the *Working Group on Adverse Event and Disease Reporting, Surveillance and Analysis* (http://dimacs.rutgers.edu/Workshops/AdverseEvent/index.html), which was formed as part of a five-year *Special Focus on Computational and Mathematical Epidemiology* (http://dimacs.rutgers.edu/SpecialYears/2002_Epid/) at the Center for Discrete Mathematics and Theoretical Computer Science (http://dimacs.rutgers.edu/). A subgroup of participants from the Working Group constitutes the author list.

DOI: 10.1002/sim

many other considerations which are not characteristics of the data *per se*, but which are related and must be addressed in order to operationally make practical use of data in real time. Those considerations are important criteria for making policy decisions for system development but are situational so will be included here in general but will not be specifically addressed. These include:

- Cost.
- What data one already has (relative coverage for geography, demographics, etc.).
- Implementation feasibility (given one's information technology infrastructure, staffing capabilities, political factors, administrative and other logistical factors, etc.).
- Privacy/confidentiality factors.

For the sake of the statistical focus of this report, the main criteria for data under consideration will be their ability to (1) detect events of importance early, (2) provide for situational awareness including post-event dynamic characterization of health impact and (3) enable surveillance capabilities for other public health purposes during periods when there is no BT. The interdependency of analytic approaches with data characteristics is fully acknowledged.

Data analytic issues

Particular data factors that are directly associated with the utility of surveillance operations and may be addressed or at least accounted for through creative analytic approaches are 'data lag', 'time alignment' and the 'unlinked data source' problem. The data lag time can be operationally described as (1) the average time between a population event (e.g. patient encounter or some other health-seeking behavioural event) and the event's data representation in an analytic system interface or (2) the proportion of data available at the time a decision is needed (*versus* at some later time). 'Time alignment' refers to the differential health-seeking behaviour times relevant for various data sources that may be available in one analytic system. For example, if one were able to view time series signals in response to a population exposure that caused illness, it may appear earlier in time for OTC sales data than for ED data. The reasoning is that people generally purchase products for self-treatment before symptoms would be severe enough to warrant a trip to the ED. There is little evidence for behavioural response staging patterns to becoming ill.

The unlinked data source problem is an issue for the secondary use of data sources when record linkage is either not possible or is avoided for other reasons. Given that much of the data used in automated surveillance are gathered for some other purpose (e.g. treating patients, billing, market analysis, inventory) and that protecting individual confidentiality is a motive, broad linkage of records is not generally feasible. Therefore, the extent of information overlap is unknown across data streams. For example, if a system uses OTC, ED and laboratory test order data, it is not known to what extent the same people and their reactions to illness are manifest in the different sources.

This is a small subset of the issues that motivate needed research in order to more fully take advantage of secondary data sources for surveillance. These data have been referred to as opportunistic to emphasize that they are used for a purpose other than the original intent of their collection. There is no sampling design that explicitly defines the relationship between the data and the population that they represent. Therefore, the basis for quantifying inferential conclusions using probability concepts is more challenging than, for example, when using sample survey data or data resulting from an experimental design. In contrast, the application of statistical and other analytical approaches is empirical with modern surveillance data. Refinement of methodologies is more dependent on applied experience with data based on creative new applications of statistical theory and concepts. In operational practice, the analytic data monitor in near real time surveillance in an empirical system is employing a multitude of deductive procedures (e.g. to rule out perceived aberrations in expected patterns that may only be indicative of data processing, coding or transmission artefacts) and inductive methodologies (e.g. to enable probabilistic decisions for responding to potential threats to the public). Methodologies, techniques and tools that can incorporate both types of reasoning are needed. In addition, the methodologies would best be adaptable to evolving surveillance requirements for standardization at least in concept. That way, if 'small areas' have differential requirements, standardized concepts could be common ground for combining information across jurisdictions for 'large areas'. Designed flexibility for surveillance system development as a goal is important not only for the front end of data systems, but also for the analytic side.

A characterization of the research needed in order for efficient multiple data stream analytic exploitation for surveillance utility cannot be properly addressed in isolation from the context in which it is translated into practice. To say that we need to get from point A to point B is not complete. We also need to know what size steps to take. We know that we would like to take fuller advantage of available data and prior information. The methods for accomplishing this must be compatible with the public health community's needs and ability to adapt. Communication and interaction between subject matter experts and methodologists is a major success factor for accomplishing this goal.

STATISTICS AND THE ROLE OF THE ANALYTIC SURVEILLANCE DATA MONITOR

The growing availability of data streams for biosurveillance requires corresponding growth in methodologies to analyse them. Analytic data monitors examine these data on a daily basis at all levels of public health. In 2005, the job descriptions and protocols of these monitors are still being established [8, 9]. Investigation of statistical anomalies beyond the database level is labour-intensive and time-consuming [10]. A multiplicity of data sources has appeal because consistent evidence may be employed to suggest inferential accuracy. In practice, however, multiple data sources can be contradictory. Plate 1 shows time series plots of syndromic data taken from a large Maryland county leading into the influenza season of 2004.

The data sources represented were counts of respiratory diagnoses from visits to civilian physician offices ('Office Visits (OV)'), military clinic visits ('MILITARY'), hospital emergency departments ('ED-UI' and 'ED ILI'), and sales of related OTC remedies ('OTC'). Retrospectively, there was a sharp rise in respiratory illness, confirmed by positive laboratory influenza tests, beginning in late November 2003. Public health status in preceding weeks was less clear. The September increase in OTC sales and in civilian OV was not reflected in the other data streams. Sporadic influenza cases were documented in October and early November, but for those weeks, the illustrated increases in the clinical data streams were gradual.

Decision requirements for the prospective analytic data monitor involve when and how deeply to investigate a data anomaly as well as when to escalate the information (as an alert) for action. Unambiguous, corroborated data spikes are the exception rather than the rule. For single data streams, univariate algorithms employ data modelling and hypothesis tests to provide systematic signal escalation protocols. In the multivariate data environment, the statistical decision requirements of the analytic data monitor also include: (1) which combinations of data sources to test, (2) which algorithms to use with respect to characteristics of the data background, (3) how to achieve sensitivity over many locations within manageable false alert rate frequency, and (4) how much corroboration among data streams is required to achieve a threshold for escalating the information. The following paragraphs describe an approach for adapting multivariate testing methodologies from other disciplines to meet these requirements.

The parallel and consensus analytic monitoring problems

In a classical hypothesis test, values of an observed quantity are treated as realizations of a random variable, and the null hypothesis is that this variable satisfies membership in an assumed distribution. A test statistic is computed from the observed values. The mean or some other property of the assumed distribution is used to calculate the probability, or p-value, of randomly occurring values at least as unlikely as those observed. The null hypothesis is rejected if this p-value falls below a predetermined threshold α. In the current applied biosurveillance context, the null hypothesis is generally assumed to be the absence of a disease outbreak. An outbreak is suspected if the null hypothesis is rejected. An outbreak, however is not a necessary condition but only one possible cause for this type of observed signal in data. Other possible causes include changes or errors in diagnostic coding, increases in participating data providers, and database problems. However, if a single data stream adequately represents the care-seeking behaviour of the monitored population for a given syndrome group, and the care-seeking behaviour of a population reflects its disease status, an outbreak may be a sufficient condition for a signal in data. Thus, alerts based on such signals in data may be used to focus the attention of health officials to potential outbreaks if there is a reasonable rate of false positives. The question here is how to extend hypothesis testing to the multi-source, distributed surveillance context.

We consider two prototype monitoring problems for the multivariate context [11]. The *parallel monitoring problem* pertains to time series representing distributed locations, such as counties or treatment facilities, possibly stratified by other covariates such as syndrome type or age group. The statistical challenge is to maintain sensitivity while limiting the number of false signals arising from testing the resulting time series. The second problem, the *consensus monitoring problem*, is the testing of a single hypothesis using multiple sources of evidence. For example, the combination of syndromic counts of ED visits, outpatient clinic office appointments, and sales of OTC remedies may be used to test the hypothesis that there is no current outbreak of gastrointestinal disease in the monitored population.

Parallel monitoring methods

Signal and subsequent false alert rates can grow to a nuisance level as the number of monitored data streams increases [12]. We may preserve a nominal overall background alert rate α by replacing the individual test threshold $\underset{\cdot\cdot}{\alpha}$ with the Bonferroni bound $\underset{\cdot\cdot}{\alpha}/N$, where N is the number of monitored data streams [13]. The resulting criterion may result in a severe loss of sensitivity, especially if the data streams are correlated. Several published methods [14–17] relax the Bonferroni criterion to maintain the overall error rate with less stringent rejection criteria. Let $P_{(1)}, \ldots, P_{(N)}$ be the p-values sorted in ascending order. For correlated streams, Hommel rejected the combined null hypothesis [14] if for any j, $j = 1, \ldots, N$

$$P_{(j)} < j \cdot \alpha / C \cdot N \quad \text{where } C = \sum 1/j$$

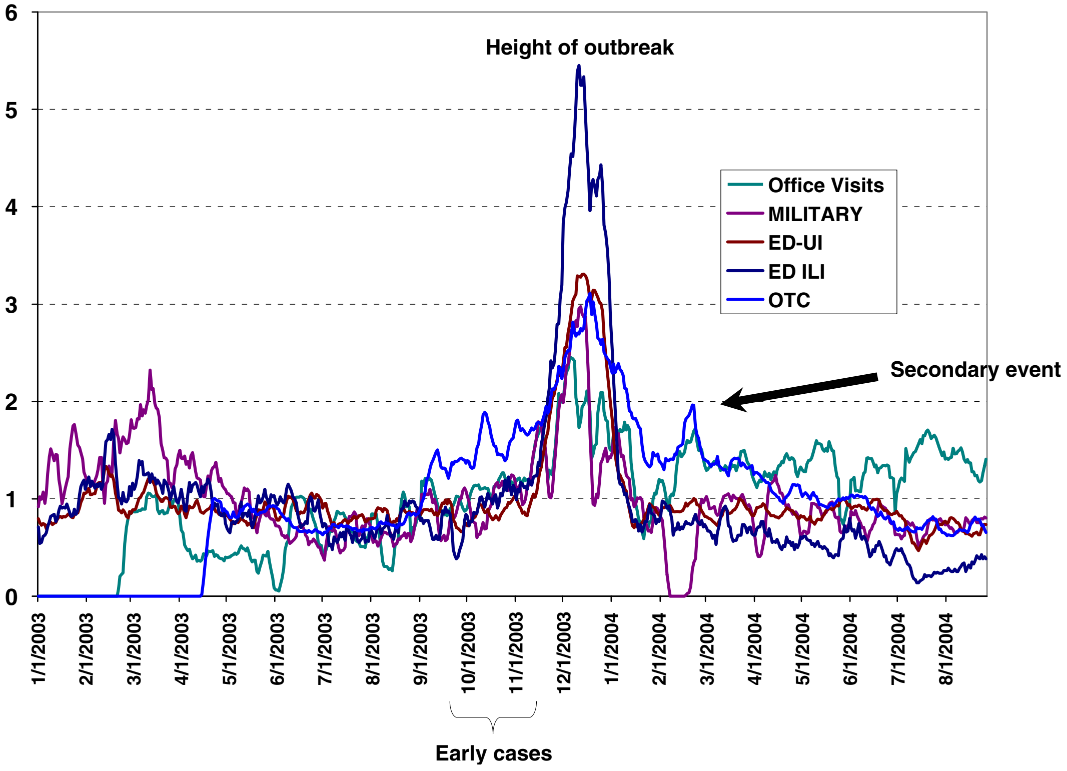

Plate 1. Recent respiratory syndrome data.

DOI: 10.1002/sim

For tests that are independent, Simes [15] and others showed that this criterion gives an overall error rate of α for $C=1$ if the tests are independent, and this relaxed criterion has been shown to maintain this error rate for many common multivariate data sets with positive correlation [17]. These improvements were widely applied after it was shown [18] that they control the false discovery rate (FDR), or expected ratio of false alerts to the total alert count. For example, National Health Service of the United Kingdom has used FDR methods to monitor results of CUSUM charts applied to hospital data streams from numerous districts [19]. The benefit of FDR methods increases as data streams are added, as their correlation increases, and as the alerting threshold α is raised. These factors should be considered in the choice of a parallel monitoring method intended to control alert rates.

For practical considerations, note that the Simes criterion may be formulated to alert if

$$\mathrm{Min}\{N \cdot P_{(j)}/j\} < \alpha, \quad j = 1, \ldots, N$$

Note that only p-values below the nominal threshold α can contribute to an alert. There is no consensus effect as in the next section—the method applied to 10 p-values of 0.06 returns 0.06. It has been noted [14–16] that the Simes criterion does not specify which of the data streams should be investigated; a popular procedure [12] is to reject the null hypothesis for all streams with p-values below the largest one that satisfies the inequality. More conservative closed-form criteria [16] have been developed that indicate which component hypotheses to reject, and the designers of large, complex systems with hundreds of simultaneous data streams should consider these criteria.

Consensus monitoring methods

The consensus problem is the combination of various sources of clinical and non-clinical evidence to gain sensitivity in disease monitoring. A critical issue in the choice of methods is whether the combination of time series adds more to the background noise—leading to excessive alerting—or to potential signals of interest. In the former situation, multiple univariate strategies are preferable, while in the latter, fully multivariate algorithms may add sensitivity at practical alert rates. We consider both strategies in the following paragraphs.

The multiple univariate methods are similar to parallel monitoring ones except that the p-values are combined to produce a single p-value $p^* = f(p_1, \ldots, p_n)$, where f has a consensus property that several near-critical values can produce a critical one. Many functions f have such a property; this section considers two methods used in independent, sequential clinical trials. The first is Fisher's rule [20], a function of the product of the p-values. The statistic is

$$F = 2\sum_{j} \ln(p_j)$$

If the separate hypothesis test results are independent, this quantity is distributed χ^2 with $2n$ degrees of freedom. As a multiplicative method, it is more sensitive to a few small p-values than to a broader number of moderate values. It is recommended if the objective is to extract a single decision on whether to avoid the overall null hypothesis and avoid considering the individual p_j.

DOI: 10.1002/sim

The second statistic is Edgington's method [21], an additive method that calculates the resultant p-value as

$$\mathbf{p_E} = \frac{\mathbf{S^n}}{\mathbf{n!}} - \binom{\mathbf{n}}{\mathbf{1}} \frac{\mathbf{(S-1)^n}}{\mathbf{n!}} + \binom{\mathbf{n}}{\mathbf{2}} \frac{\mathbf{(S-2)^n}}{\mathbf{n!}} - \binom{\mathbf{n}}{\mathbf{3}} \frac{\mathbf{(S-3)^n}}{\mathbf{n!}} + \cdots$$

where **S** is the sum of the **n** **p**-values. The summation continues until $(\mathbf{S}-\mathbf{j})$ is no longer positive. This additive method is more sensitive to multiple, near-critical values. For more than a few dozen data streams, this formula cannot be computed accurately. In such cases, the expression

$$(\text{mean}(p) - 0.5)/(0.2887/\sqrt{n})$$

gives a z-score whose Gaussian probability is a close approximation to this formula [22].

If the data streams are independent, Edgington's method gives fewer alerts than Fisher's method at nominal thresholds but is more sensitive to data correlation. Edgington's method is recommended if the number of data streams is modest—say less than a dozen—and the user wishes a sensitive consensus indicator in addition to the individual test results. This need has been expressed by epidemiologist users who require some summarization but are skeptical of bottom-line results that hide the contributions of individual evidence sources. Note that very small single p-values do not necessarily cause alerts in Edgington's method because it is an additive method. If the system is not also monitoring single streams, either the use of Fisher's method or both methods is recommended.

An example of the potential benefit of combining univariate alerting results is shown in Plate 2.

Input data were the time series shown in Plate 1 for emergency department visits (ED), physician OV, and OTC sales. The ordinate for these plots is the daily computed algorithm p-value, with markers below the red line indicating alerts at the $\alpha = 0.01$ level. Univariate EWMA algorithms were applied to each individual time series, and the derived p-values are shown for the office visit data. Also plotted are composite daily p-values computed by combining univariate values for the three time series with the Edgington method. The composite algorithm showed added sensitivity for the influenza event. There were sporadic data signals during the early outbreak interval with stronger and more consistent signals during the height of the outbreak.

Multivariate methods

Alerting algorithms that combine values from separate time series in a single computation have the potential to detect evidence of faint outbreaks. While strong correlation among data sources tends to dilute the benefit of FDR-like methods, prospective multivariate algorithms can exploit consistent correlation. Published work on multivariate methods [23, 24] based on weekly data from large regions has focused on multivariate statistical process control (MSPC) [25]. Little published research deals with more complex multivariate hypothesis tests based on wavelets, Bayesian statistics, etc. Most MSPC methods are based on Hotelling's T^2 as applied in monitoring efforts in related fields [26]. The T^2 statistic may be written as

$$(X-\mu)S^{-1}(X-\mu)$$

where X, multivariate data from the test interval; μ, vector mean estimated from the baseline interval; and S, estimate of covariance matrix calculated from the baseline interval.

While Hotelling's T^2 may be viewed as a multidimensional z-score, this method has been generalized to obtain other multivariate control charts. A multivariate EWMA chart (MEWMA)

DOI: 10.1002/sim

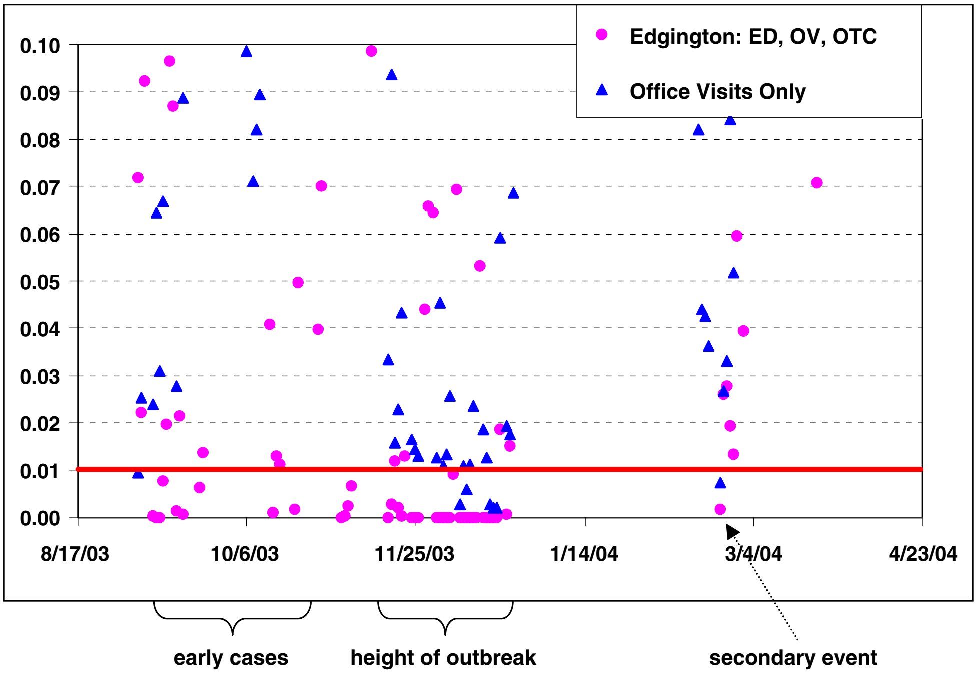

Plate 2. Effect of combining evidence with multiple univariate methods.

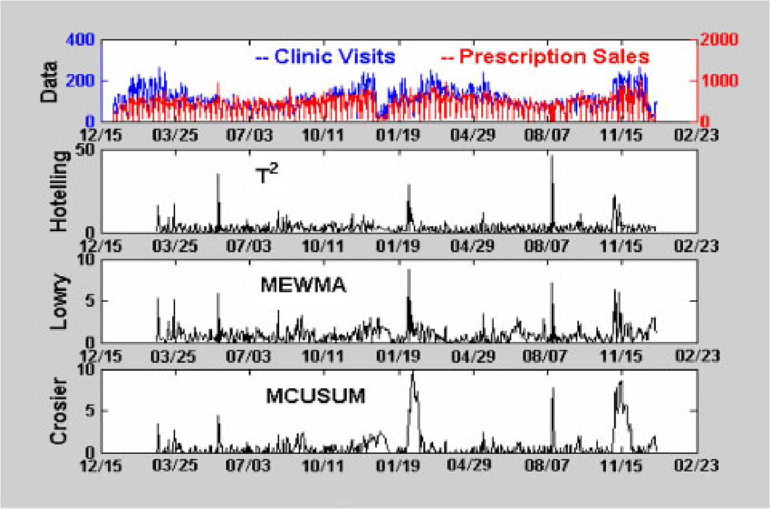

Plate 3. Application of MSPC methods to two correlated data streams.

DOI: 10.1002/sim

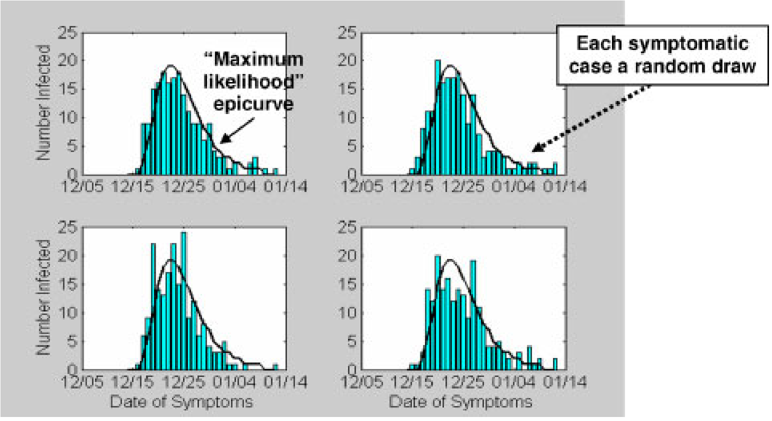

Plate 4. Modelling the data epicurve: 4 stochastic realizations.

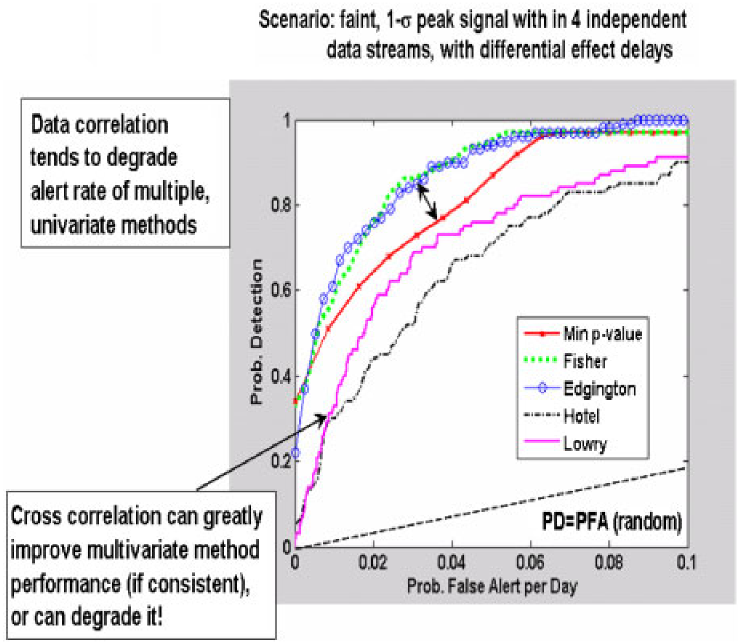

Plate 5. ROC curve comparisons for multivariate methods applied to simulated data.

has shown improved run length characteristics [27] and has yielded promising results with health surveillance data [28]. In Lowry's MEWMA, the data vector is replaced by the exponentially weighted moving average

$$Z_j = RX + (1 - R)Z_{j-1}$$

where R is a diagonal matrix of smoothing coefficients, and the covariance matrix is a scalar multiple of the data covariance matrix S in the usual application where equal smoothing coefficients are used. Analogous multivariate CUSUMs have also been applied to surveillance data, with Pignatiello's MCUSUM [29] applied to yearly, spatially distributed counts of breast cancer incidence [24]. Hawkins [30] describes Hotelling's T^2 as 'particularly bad at distinguishing location shifts from scale shifts.' Rogerson and Yamada [24] notes that combined univariate methods are 'directional' in that they may be quick to detect shifts in just a few data sources but less sensitive to shifts in more general directions. These methods are omni-directional—a property that can be useful in detecting an earlier signal, but which can also cause false alerts if there is a change in the covariance matrix that is irrelevant to any outbreak signal of interest. Plate 3 illustrates this problem with applications of three MSPC methods, Hotelling T^2 [26], Lowry's MEWMA [26], and Crosier's MCUSUM [31], to two syndromic time series that are highly correlated.

The spikes in the algorithm outputs were in general agreement and plausible signals except for those seen for August 7. These spikes are purely the result of a change in day-of-week behaviour for one of the two data streams. For practical analytic data monitoring to take advantage of the added sensitivity of these MSPC techniques, procedures and protocols must be developed to avoid irrelevant alert escalation.

Evaluation methodology

This section presents a simulation approach for evaluating the detection performance of the monitoring algorithms discussed above. To challenge the algorithms, data epicurves of injected cases attributable to a presumed outbreak are added to background data. These epicurves are stochastically drawn from an ideal incubation period distribution [32]. This procedure is adopted to produce plausible outbreak-like signals, in contrast to the standard method of adding a fixed quantity to the process mean to find the average run length of a control chart [13].

We first describe the procedure for testing univariate alerting algorithms. The lognormal incubation period distribution of Sartwell [33] is used to estimate the idealized curve for the expected number of new symptomatic cases on each outbreak day. Stochastic epicurves are drawn as follows. In order to construct a difficult signal to detect, the number of cases on the peak outbreak day is set at a fixed multiple k ($1 \leqslant k \leqslant 3$) of the standard deviation of the background data. The total outbreak size N is this peak value divided by the maximum of the lognormal probability density function and rounded up. Individual incubation periods are then chosen with a set of N random lognormal draws, one draw for each simulated case. The number of cases to add for each day after onset is then the number of draws rounded to that day. Four typical data epicurves drawn with this procedure are shown in Plate 4.

To test a univariate algorithm, repeatedly draw one of these stochastic signals and then add it to the background time series at a randomly chosen start day (beyond a warm-up period that is kept constant for all tested methods). Algorithms to be evaluated are then applied to each resultant time series, and for each threshold level of interest, sensitivity or probability of detection (PD) is estimated as the fraction of all trials for which the algorithm output exceeds that level

DOI: 10.1002/sim

during the interval of injected cases. For the same threshold, the daily false alarm probability (PFA) is the fraction of non-outbreak days for which the level is exceeded. (The signals may be injected into simulated or authentic background data. While an authentic background may be preferable, it makes the identification of non-outbreak days more subjective.) A receiver operating characteristic (ROC) curve may be formed by plotting PD *versus* PFA as the algorithm threshold varies. Algorithm performance can be precisely measured in this process because the start and duration of each simulated outbreak are known.

For testing multivariate algorithms, four additional simulation decisions reflecting the complexity of multivariate detection affect algorithm performance: (1) the number of data streams analysed, (2) the number of data streams an outbreak affects, (3) the relative magnitudes of the outbreak effect among the impacted streams, and (4) the time lag between the separate data signals. The first three have been shown to affect the decision of whether a multiple univariate or a multivariate algorithm is preferable [24].

We present ROC curve results of a single experiment to compare the straightforward method of choosing the minimum univariate p-value with multiple univariate methods (Fisher and Edgington) and with multivariate methods (Hotelling's T^2 and Lowry's MEWMA). To represent background data streams, four independent time series simulating 700 days of syndromic data counts were formed by random draws from a Poisson distribution with a mean of 100. The lognormal parameters of the epicurve distribution were chosen to give a median incubation period of 3.5 days, consistent with associated symptoms of known weaponized diseases [34], and a temporal case dispersion consistent with past observed outbreaks [33]. The expected number of attributable cases on the peak day of the outbreak was set at one standard deviation of the background, or 10 cases, for each data stream. The stochastic signals were computed separately and injected into all four data streams with effect delays of 0, 2, 3 and 4 days. Plate 5 shows the resulting ROC curve for 100 trials.

Daily false alert rates, or background recurrence rates, may be read from the x-axis. An expected alert every 2, 6 and 10 weeks corresponds to PFA values of 0.071, 0.024 and 0.014, respectively. Corresponding sensitivity estimates may be read from the plot. The independence of the background streams gave a clear advantage to the multiple univariate methods and degraded the MSPC performance. As noted above, the Edgington method is more consensus-oriented, and it suffers from the relative signal delays. It outperformed the Fisher method when effects were simultaneous, but its background alert rate was far more degraded by cross-correlation. Additional experiments with correlated data streams have pushed the MSPC curves above the others, but these advantages are highly scenario-dependent and require consistent correlation. For some scenarios, the ROC curve for the minimum p-value gives the best performance, indicating an FDR-like combination rule. The overall message is that the optimal algorithm choice requires an understanding of the multivariate data environment.

The effectiveness of monitoring with multiple data sources depends on user acceptance and visualization tools as well as on algorithm development and data analysis. The utility of multivariate methods in biosurveillance systems [35] involves epidemiologist, analytic data monitors, database/website designers and statisticians. This cross-disciplinary collaboration requires a technical implementation approach in which familiar data plots, univariate charts, and regression predictions are employed while newer fusion methods are gradually introduced along with combination views. This process involves evolution of technologies as well as the knowledge base of the analytic data user community at all levels of public health [8, 36]. The data analytic methods described in the following sections are successively less implemented but hold promise for effectiveness in the modern public health surveillance context as more insight is gained about the nature of multiple relevant data sources and how best to incorporate more refined approaches.

SPACE–TIME SCAN STATISTICS WITH MULTIPLE DATA STREAMS

A short review

Scan statistics [37–39] have been used to detect and evaluate temporal disease clusters since the early 1980s [40, 41]. Spatial scan statistics [42] have also become popular for the evaluation of geographical disease clusters in a wide range of application areas including cancer [43, 44], infectious diseases [45, 46] and pediatrics [47, 48]. More recently, space–time scan statistics have been used in a prospective setting for the early detection of disease outbreaks [49, 50], with uses ranging from West Nile virus [51] to syndromic surveillance [52]. The basic idea in the prospective setting is to use a variable size cylinder where the circular base represents space and the height represents time. This concept is then used algorithmically, to scan across the geographical and temporal study region, comparing counts of the observed and expected number of cases within cylinders. Only those cylinders where the temporal height reaches the current time are used. Based on a likelihood criterion, the most unusual cylinder is noted, and its statistical significance is evaluated using Monte Carlo hypothesis testing [53], adjusting for the multiple testing inherent in the many cylinder sizes and locations evaluated.

In order to simultaneously search for and evaluate clusters in more than one data set, a multivariate scan statistic that incorporates multiple data sets from the same geographical area into one single analysis can be used [54, 55]. For example, one may be interested in spatial clusters with excess incidence of leukaemia only, of lymphoma only or of both simultaneously. As another example, one may be interested in detecting a gastrointestinal disease outbreak that affects children only, adults only or both simultaneously. The different data set could also be from different sources, such as (1) ED visits from hospitals, (2) ambulatory care visits from a health insurance plan and (3) over the counter drug sales from a pharmacy chain. As discussed in the previous section of this report, each of these data sources may contain evidence for a specific disease outbreak as different people seek different types of health care.

The rationale for the multivariate scan statistic is that together there may be sufficient joint evidence to signal an outbreak even when neither constituent generates a signal on its own. This idea may be interpreted (although we may not have a valid independence assumption between data sources) as increasing the detection 'power' by increasing the 'sample size'. If a space–time scan statistic is used to analyse one single combined data set, one may miss a cluster that is only present in one of the subgroups. On the other hand, if two analyses are performed, one for each data set, there is a loss of power if the true cluster is about equally strong in both data sets.

The multivariate scan statistic with multiple data sets works as follows when searching for clusters with high rates:

1. For each window location and size, the log likelihood ratio is calculated for each data set.
2. The log likelihood ratios for the data sets with more than expected counts are summed up. This sum is the likelihood for that particular window.
3. The maximum of all the summed log likelihood ratios, taken over all the window locations and sizes, constitutes the most likely cluster. This is evaluated in the same way as for a single data set.

When searching for clusters with low rates, the same procedure is performed, except that we instead sum up the log likelihood ratios of the data sets with fewer than expected counts within the window in question. When searching for both high and low clusters, both sums are calculated, and the maximum of the two is used to represent the log likelihood ratio for that window. A detailed description of the multivariate scan statistic has been presented by Kulldorff *et al.* [55], who also applied it to three different types of physician encounters: (1) telephone calls, (2) regular ambulatory care visits and (3) urgent care visits in the Boston metropolitan area. Analyses using the multivariate scan statistic can be done using the freely available SaTScan software (www.satscan.org).

Open research questions

When working with multiple data types, the ideal is to have independent data sets, where no person is counted more than once. If not, temporal correlation in the data may trigger unnecessary false alarms. (This issue is referred to summarily in the conclusions section of this report as 'record linkage'.) When incorporating multiple data types within the same analysis, and the same person has more than one encounter of either the same or a different type, the duplicate encounter records should ideally be removed from the count. For example a patient first visits her regular physician and then goes to the ED the next day. When all information comes from the same data provider, such as a health insurance company, it is sometimes possible to remove such duplicates [56], but it is a much more challenging problem when there are multiple data providers. Modifications to approaches using multiple data sources are needed to take the duplicate encounters by the same person into account. If not, a single person could theoretically generate a signal if he has many health encounters in a very short period of time.

When we *are* able to remove duplicates, it is not clear that it is always the latter encounter that should be removed. For example, if a person visits her doctor on Tuesday due to fever and a subsequent laboratory test is found to be positive on Thursday, the latter encounter may contain more 'information'. In such a situation, one could include the fever encounter in the analyses performed on the Tuesday and Wednesday data but eliminate it from subsequent analyses when the laboratory test result is available. The best approach will obviously vary depending on the data used, and to some extent one can use intuition, but formal investigations in different settings would be very valuable. For the multivariate space–time scan statistic one option is to use the same cylinder for each data set when adding up the log likelihoods. That may not be the best approach though. When there is asynchronous reporting, with, for example, laboratory tests of blood cultures that take longer than gram stains or urine cultures, the various data sets should be appropriately synchronized. Depending on the data types and the infectious disease outcome under study, one could incorporate different temporal lag times, but this has not yet been tried. Experimentation to scan with different spatial windows for the different data sets may also prove informative.

There is nothing in the application of the multivariate space–time scan statistic that requires uniform geographical coverage for each data set. At the same time, little is known about how such differential coverage will affect the analysis and the ability to pick up a disease outbreak in different regions. The use of irregularly shaped scanning windows [57–60] rather than circles has been proposed and it makes sense to try such approaches for prospective multivariate space–time scan statistics as well.

The multivariate space–time scan statistic is considerably more computer intensive than the univariate version. Ideas presented by Neill and Moore [61] could potentially be very useful to improve speed if they can be generalized to the multivariate space–time setting.

DOI: 10.1002/sim

With the multivariate scan statistic, clusters may be detected in one or any combination of the data sets used. That is often what we want but in some cases we may only be interested in certain combinations of data sets. For example, while we may wish to group two different types of leukaemia, two different types of lymphomas, or all leukaemias with all lymphomas, we may not want to group one type of leukaemia with one type of lymphoma without including the other types. For that, one could combine the space–time scan statistic with the tree-based scan statistic [62] to create a space–time-tree dimensional scan statistic. How such a method would perform in practice is unknown.

In summary, the multivariate scan statistics shows promise as a useful tool for early disease outbreak detection when there are multiple data sources. There is still much to learn about its practical utility which includes (1) how to handle duplicate data, (2) the optimal use of temporal and spatial lags, (3) differential geographical data coverage effects, (4) irregularly shaped cluster issues, (5) computational algorithm efficiency and (6) how to combine the space–time and the tree-based scan statistics. It is in our interests to continue building upon this tool for exploiting multiple data source to enhance disease detection and rapid situational awareness. Much of the needed research, especially regarding surveillance data knowledge, will simultaneously serve interests for other methodologies that we will continue to explore and refine as we evolve our analytic opportunities in this field. The following concept represents another such opportunity.

A BAYESIAN APPROACH TO COMBINING MULTIPLE DATA STREAMS

In this section we describe a particular Bayesian approach to combining multiple data streams to perform biosurveillance for disease outbreaks. The approach is based on modelling a population as a set of the individuals (which are also modelled) in the population. We first describe this basic approach. Next we introduce the Bayesian network (BN), which is the representation we use in modelling a population as a set of individuals. We then describe a particular Bayesian-network-based biosurveillance system that we have implemented, and we explain how it combines multiple data streams in performing biosurveillance. We conclude with a set of research challenges.

Individual-based modelling

In individual-based modelling, each person in the population (being monitored for disease outbreaks) is represented as a sub-model. These sub-models are linked to each other through the outbreak diseases being modelled. Through inference, evidence about the individuals can influence belief about the presence of outbreak diseases in the population. Individual-based modelling allows us to represent what we know about each person that is relevant to disease outbreak detection. We may have more information about some people than others. For one person we may only have evidence about age, gender, and home zip code. For another person, we may also know that he came to the ED at a particular time with a particular primary problem (chief complaint).[¶] Beyond evidence about individuals, we also can represent evidence about the population by functionally modelling behaviour of individuals. As an example, the number of OTC thermometers sold in a

[¶] The data we use for biosurveillance are de-identified in the sense that appropriately we do not have information, such as name, address, or a social security number, that would uniquely identify a person, even if that person has visited in an ED, for example.

DOI: 10.1002/sim

given city on a given day can be reasonably modelled as a function of the purchases of those thermometers by individuals. By modelling (under uncertainty) thermometer purchases by individuals, we can derive a distribution over the number of thermometers sold on a given day in the region. Doing so is important, because typically we only know the aggregate sales volume, rather than the purchases of thermometers by specific individuals. In the next section, we describe the representation that we use to model individuals and populations.

Bayesian networks

A BN model represents the joint distribution of a set of variables of interest [63–65]. For example, over all possible joint states of the variables in evidence set E and hypothesis set H, a BN could represent $P(E, H)$. From such a joint distribution, we can derive any probability of interest, such as $P(H \mid E)$.

A BN has two parts: a structure and a set of parameters. The structure contains nodes, which represent model variables,[‖] as well as arcs, which connect nodes that are related probabilistically. The resulting graph is required to contain no directed cycles, meaning that it is not possible to start at some node and follow a path of arcs that leads back to that same node. The parents of a node (variable) X_i are those nodes that have arcs into X_i. A descendant of X_i is a node X_j for which there is a directed path from X_i to X_j. The following Markov condition specifies the independence that is expressed in a BN:

A node is independent of its non-descendants, given just the state of its parents

This is the most important key to understanding BNs. It tells us that if we want to predict the value of some variable in the BN, and we already know the values of all of its parents, then no variable (except possibly some of a node's descendents) could possibly give us any useful predictive information beyond that supplied by its parents. As an example, the BN structure in Figure 1 contains five nodes. The node *Season* could be modelled as having the values *spring*, *summer*, *fall* and *winter*. *Age* represents a patient's age, perhaps discretized into ranges of years. The nodes *Influenza*, *Cough* and *Fever* could be modelled as being *present* or *absent*.

As an example of the BN Markov condition, we see that the structure in Figure 1 specifies that *Cough* is independent of *Season* given the state of *Influenza*. If the designer of the network had decided that this independence assumption was unreasonable, she could have added an arc from *Season* to *Cough*.

Someone who designs a BN first develops the structure. But more work is needed after that. The network must be populated with numerical parameters. For each node X_i, there is a probability distribution $P(X_i \mid \mathit{parents}(X_i))$, where $\mathit{parents}(X_i)$ are the parents of X_i in the BN. For example, we might have $P(\mathit{Cough} = \text{present} \mid \mathit{Influenza} = \text{present}) = 0.90$. If X_i contains no parents, then the probability $P(X_i)$ is specified. The BN Markov condition implies that we can factor the joint distribution of the variables in the model as follows [65]:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathit{parents}(X_i)) \tag{1}$$

The fewer the number of parents per node, the fewer the number of parameters needed to specify each conditional probability $P(X_i \mid \mathit{parents}(X_i))$, and thus, the fewer the number of parameters

[‖] We use the terms *node* and *variable* interchangeably in this section.

DOI: 10.1002/sim

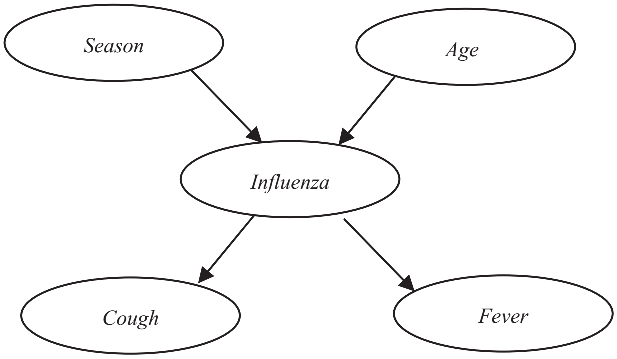

Figure 1. An example of a Bayesian network structure.

in the model. Therefore, a BN with relatively fewer arcs will require relatively fewer model parameters. A BN can thereby represent parsimoniously the joint distribution over n variables.

If the arcs are interpreted as being direct causal relationships (relative to the variables being modelled) then the model is called a *causal Bayesian network*. For a causal BN, the Markov condition states (in part) that effects are independent of distant causes, given the state of all the proximal causes. For example, *Fever* is independent of the distant influence of *Season* and *Age*, given that we know the state (present or absent) of *Influenza*, which is the proximal cause of *Fever* in the model in Figure 1.

Equation (1) specifies a complete joint probability distribution over all n variables in the BN model. From such a joint distribution we can derive the probability of any subset of nodes conditioned on the state of another subset of nodes. Thus, for the example BN, we could derive $P(\mathit{Influenza}\,|\,\mathit{Cough}=\text{present}, \mathit{Season}=\text{winter}, \mathit{Fever}=\text{present})$. Note that information about the patient's age is missing in this conditional probability; in general, we need only condition on a subset of the variables in the model.

Researchers have developed exact BN inference algorithms that take advantage of the independence among the variables that follows from the BN Markov condition when some arcs are missing. These algorithms are often able to derive conditional probabilities relatively efficiently [65]. When exact inference would require too much computation time, approximate algorithms are available [64, 65].

PANDA: a Bayesian-network approach to biosurveillance

Population-wide anomaly detection and assessment (PANDA) is a biosurveillance application that is based on a BN composed of sub-networks of individuals in the population. The current implementation focuses on representing non-contagious infectious diseases, such as inhalational anthrax. Figure 2 illustrates this representational focus schematically. *Population Risk Factors* represents factors that influence the current risk level, such as the national terrorist alert level. *Population disease factors* (PDF) denotes all the information about disease exposure that renders the individuals (persons) conditionally independent of one another, when we are not conditioning on *Population-Wide Evidence*. The *Population-Wide Evidence* designates evidence that is a function of the individuals and their behaviour, such as the total number of thermometer sales on a given day in the region.

DOI: 10.1002/sim

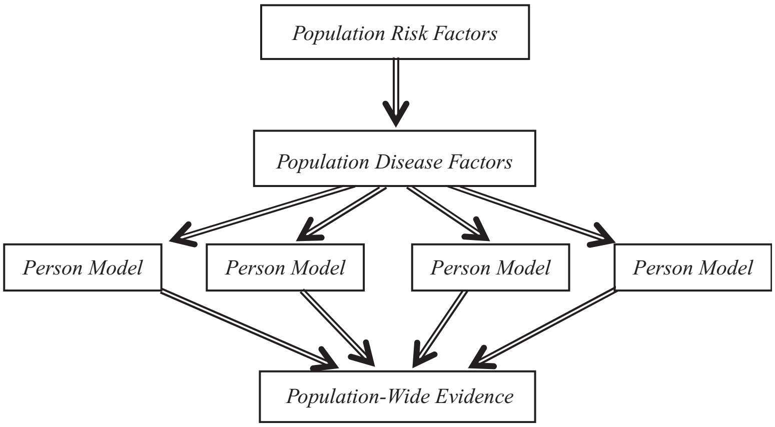

Figure 2. A schematic for representing how a non-contagious disease can influence individuals in the population, which in turn can affect population-wide evidence. Each rectangle in this diagram denotes a Bayesian sub-network and each arrow denotes one or more Bayesian-network arcs between the sub-networks.

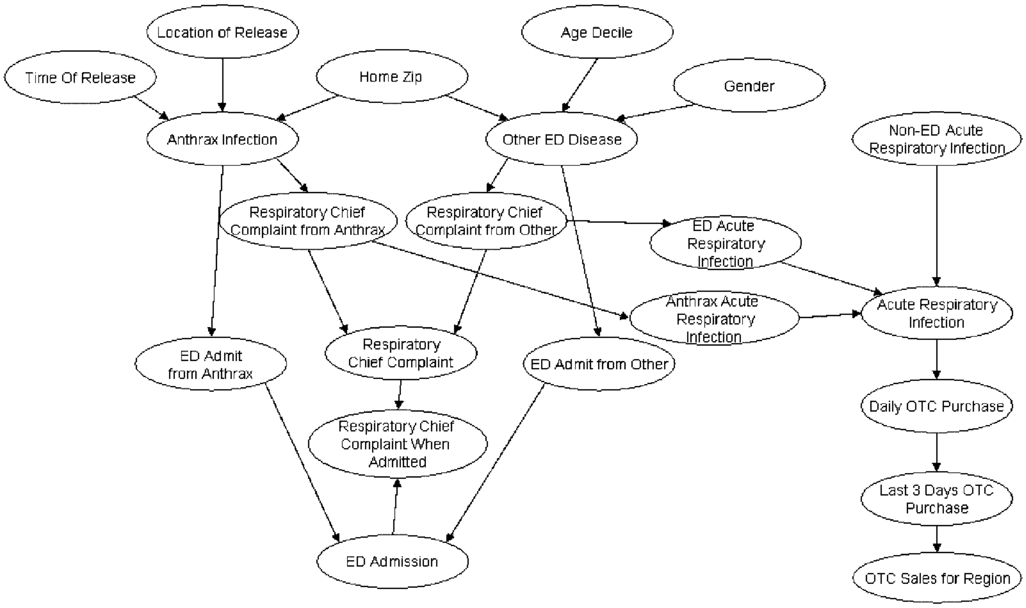
Location of Release
Age Decile
Time Of Release
Home Zip
Gender
Anthrax Infection
Other ED Disease
Non-ED Acute Respiratory Infection
Respiratory Chief Complaint from Anthrax
Respiratory Chief Complaint from Other
ED Acute Respiratory Infection
Anthrax Acute Respiratory Infection
Acute Respiratory Infection
ED Admit from Anthrax
Respiratory Chief Complaint
ED Admit from Other
Daily OTC Purchase
Respiratory Chief Complaint When Admitted
Last 3 Days OTC Purchase
ED Admission
OTC Sales for Region

Figure 3. The person sub-network used by PANDA to model disease that is due to inhalational anthrax. This network corresponds to one of the *Person Models* in Figure 2.

Figure 3 shows the person sub-network that PANDA currently uses to model the effect of inhalational anthrax disease on an individual. A more detailed discussion of this model is in Wong *et al.* [66]. We provide here a selected summary of points that are needed in the current report.

DOI: 10.1002/sim

Space constraints permit us to discuss selected structural aspects of the model, but not any details about the parameters used in the model.

At the top left, *Time of Release* and *Location of Release* are PDF nodes. *Time of Release* ranges over the values *today*, *yesterday* and *day-before-yesterday*, which model when anthrax spores are hypothesized to have been released out-of-doors within the region, plus the value *never*, which denotes no release of inhalational anthrax. *Location of Release* has values corresponding to one of approximately 100 ZIP codes in the county-wide region being modelled, plus the value *nowhere*.

The nodes in the middle portion of Figure 3 represent that respiratory disease, as well as a possible ED admission for it, may be due to either inhalational anthrax infection or some other disease (or both). The chance that a person has an anthrax infection is influenced by when and where the release occurred (if ever) and where the person lives (*Home Zip*).

The nodes at the far right indicate that the status of *Acute Respiratory Infection* in the person influences whether she makes (or has made for her) a *Daily OTC Purchase* of cough medication for each of the previous three days, which in turn influences whether such an OTC purchase was made on any of the previous three days (*Last 3 Days OTC Purchase*). The *OTC Sales for Region* node, which represents a count, is part of the population-wide evidence layer. Although in Figure 3 this node is shown with input only from a single person model, it actually has arcs from all person models, in the same way that the *Population-Wide Evidence* node in Figure 2 has arcs from all the person models.

Combining multiple data streams

We use a specific example to illustrate a general approach to combining multiple data streams in performing biosurveillance. In particular, we show how PANDA combines OTC and ED data in deriving a posterior probability for inhalational anthrax disease in the population. For simplicity, we assume a model in which the population risk factor (see Figure 2) is just the presence of the outbreak disease itself in the population, which we represent with the binary variable T for the *target* node. The prior probability distribution over T represents the risk of a release of outdoor inhalational anthrax in the region being monitored. In the BN model, T has arcs into the *Time of Release* and the *Location of Release* nodes, which constitute the PDFs (Figure 2).[**] Let $\mathbf{e}$ represent the union of the evidence about specific individuals in the population. For those people who have recently visited the ED, we assume we know the states of the following nodes in Figure 3: *Age Decile*, *Gender*, *Home Zip* code, *ED Admission* (= yes), and *Respiratory Chief Complaint When Admitted* (= yes or no). For people who have not recently visited the ED, we only know the states for the nodes *Age Decile*, *Gender*, and *Home Zip* from U.S. Census data. Let $\mathbf{o}$ represent the node *OTC Sales for Region*, which denotes the total number of OTC sales of cough medications during the three previous days in the entire geographic region being modelled. We assume that OTC sales in the region are due to (and only due to) purchases by people in that region.

Equation (2) shows the posterior probability of the target node on the left and its derivation *via* Bayes rule on the right

$$P(T \mid \mathbf{o}, \mathbf{e}) = \frac{P(\mathbf{o}, \mathbf{e} \mid T) \cdot P(T)}{\sum_T P(\mathbf{o}, \mathbf{e} \mid T) \cdot P(T)} \tag{2}$$

[**] In an extended model, other factors could profitably be included, such as the amount of the release and the weather conditions at the time of the release.

DOI: 10.1002/sim

A key term in equation (2) is derived in equation (3) by summing over all the joint values of the two variables in the PDF set, which for brevity we denote as **I** in equation (3). We can derive the term $P(\mathbf{I}\mid T)$ in equation (3) using BN inference

$$P(\mathbf{o},\mathbf{e}\mid T)=\sum_{\mathbf{I}} P(\mathbf{o},\mathbf{e}\mid \mathbf{I})\cdot P(\mathbf{I}\mid T) \tag{3}$$

We can derive the remaining term in equation (3) using the following equation:

$$P(\mathbf{o},\mathbf{e}\mid \mathbf{I})=P(\mathbf{o}\mid \mathbf{e},\mathbf{I})\cdot P(\mathbf{e}\mid \mathbf{I}) \tag{4}$$

We can derive the term $P(\mathbf{e}\mid\mathbf{I})$ in equation (4) using BN inference, as explained in [67]. In the remainder of this section, we describe a method to derive the term $P(\mathbf{o}\mid\mathbf{e},\mathbf{I})$ in equation (4). Consider a set of individuals who share a common set of values for the nodes represented by **e**. Let Q_j designate an arbitrary such set, let n_j designate the number of individuals in Q_j, let e_j designate their shared set of evidential values, let p_j designate the probability distribution that an individual in the set has made an OTC cough medication purchase within the previous 3 days, and let o_j be a random variable that represents the number of OTC purchases by the individuals in Q_j. Note that p_j is a function of both e_j and the state of **I**, as given by the network in Figure 3. We will call Q_j an equivalence class. Let Ω denote all such equivalence classes. The distribution of **o** is just the distribution of the sum of the o_j, each of which is a binomial variate. Since there is no efficient way directly to derive the distribution over a sum of binomial variates, we approximate the binomial distribution of each o_j as a normal distribution with mean $n_j p_j$ and variance $n_j p_j(1-p_j)$ [68]. The distribution of **o** is then a normal distribution with the following mean and variance:

$$\mu=\sum_{Q_j\in\Omega} n_j p_j \quad\text{and}\quad \sigma^2=\sum_{Q_j\in\Omega} n_j p_j(1-p_j)$$

Finally, the probability of observing a given value (count) for **o** is approximated as follows:

$$P(\mathbf{o}\mid\mathbf{e},\mathbf{I})=\int_{\mathbf{o}-0.5}^{\mathbf{o}+0.5}\mathrm{N}(x;\mu,\sigma^2)\,\mathrm{d}x$$

In our implementation, we took the single-region approach described above and generalized it to model approximately 100 Zip code regions. We then performed a preliminary evaluation of the computational run time when processing semi-synthetic data from a 1.4 million person region (see [66] for details). The system required 210 s to initialize, and then it used about 94 s of CPU time in monitoring 24 h worth of ED and OTC data. Thus, the system's run time was significantly faster than real time.

We can generalize the above technique by using multinomial distributions to represent the probability distribution that each person in an equivalence class will make an OTC purchase in each of the possible Zip codes. We also can generalize the above approximation by using multivariate normal distributions to approximate multinomial distributions [69, p. 87]. In addition, it is possible to represent more than one type of OTC purchase.

Summary and challenges

The Bayesian method described in this section appears promising as an approach to combining some of the multiple data streams that appear useful for biosurveillance. A preliminary evaluation suggests that the approach is computationally feasible when given realistic amounts of data to monitor. Nevertheless, a number of challenges remain. In general, many potential outbreak diseases of concern (e.g. inhalational anthrax and smallpox) are rare. Thus, it is challenging to model how the diseases present in individuals and how those individuals will behave (e.g. when and what OTC medications they will purchase and when they will seek medical care). A Bayesian approach does at least provide us with the capability of modelling such events under uncertainty. Thus, our detection models can in principle be as informed as possible, relative to the best existing knowledge from experts and the literature.

We need to investigate more thoroughly the quality of the normal distribution approximations of binomial distributions for the current application, and by extension the multivariate normal approximations of multinomial distributions. Since the counts in the equivalence classes can be small, the quality of the approximations may suffer.

Another challenge is to develop models of contagious diseases (e.g. Ebola and Smallpox) and tractable inference algorithms for applying those models for biosurveillance.

Finally, a general challenge for all biosurveillance research 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.

CONCLUSIONS AND RESEARCH AGENDA DISCUSSION

We can characterize the core challenges in public health surveillance for BT as (1) quick access to useful information from data when it is needed for a yet unknown reason (i.e. situational awareness) and (2) anomaly detection in heterogeneous, multivariate, spatially referenced, time series (i.e. event detection). Reports continue to appear in the Statistics literature concerning change-point detection in *univariate* time series. Methodology for the multivariate setting remains incomplete at best and many research questions present themselves. Here we consider some particular questions that we believe are of importance.

Visualization

Analytic surveillance data monitors need effective visualization tools for multivariate heterogeneous time series. The spatial aspect of the data makes problem even harder and much work remains to be done. The need to include all available evidence with appropriate weighting (clinical *versus* non-clinical, diagnostic *versus* syndromic) requires data reduction so that health monitors will not be overwhelmed by tables and plots. Algorithms tracking all possible combinations of data streams and covariates would be prohibitively resource-intensive; decisions regarding which data streams and combinations to track and how to weight them should have a solid statistical basis informed by surveillance data analysis and epidemiological considerations.

Data

Successful research on analytic methods for multivariate surveillance requires ready access to large-scale 'real' data. The current absence of such publicly available data represents a key barrier to progress. For the immediate future, surveillance systems that combine univariate analyses provide the most practical solution. Nonetheless, even within the univariate setting, several research challenges remain open including anomaly detection with non-standard data types (e.g. emergency

DOI: 10.1002/sim

room chief complaints) and development of a (univariate) decision-theoretic approach that considers the costs of various errors. We described two combination approaches above—Bonferroni adjustment for multiple tests and multivariate scan statistics that sum log-likelihoods. Further investigation is certainly warranted. Practical algorithms to address the alignment problem we described above are also needed. In the longer term, model-based approaches offer considerable promise. The general framework can harness prior knowledge in a natural fashion and combine this knowledge with data. Furthermore, decision-theoretic extensions are, at least conceptually, straightforward. With respect to either timeframe, rapid progress towards these research goals will require public release of data to the research community.

Alignment

Following a BT event such as a deliberate anthrax release, signals will present themselves at different times in different data sources. For example, OTC medication sales will probably rise before emergency room visit numbers. Aligning these time series to detect the single signal source represents a non-trivial statistical and algorithmic challenge.

Multiple testing

One approach to multivariate anomaly detection conducts statistical tests in a univariate setting and combines these tests with some appropriate multiplicity adjustment. This topic has a rich history in Statistics. The 'FDR' approach of Benjamini and Hochberg [18] has attracted particular attention in recent years and a number of generalizations and specializations now exist [70]. In the context of public health surveillance for BT, classical approaches such as Bonferroni may prove adequate, but characterization of optimal approaches with respect to various criteria would certainly be of value.

Record linkage

Unlinked data sources provide a further challenge, especially to surveillance approaches based on statistical tests. For example, a particular geographic region may present 100 emergency room visits for a respiratory syndrome and, in the same time period, 100 sales of OTC cough medication. Absent any form of record linkage, this could represent anywhere from 100 to 200 individuals. Since privacy concerns essentially rule out accurate linkage, statistical methods need to acknowledge and account for the added uncertainty that this problem introduces.

Model-based approaches

The model-based approach to surveillance typically attempts to model the data generating mechanism under normal circumstances in the hope of being able to recognize a future abnormality. Some models, such as Cooper *et al.*'s PANDA model, also include components that describe the expected impact of a BT event. The available data can inform the part of the model that describes normal circumstances, but the part that concerns effects of a BT event necessarily draws only on prior knowledge. Several research challenges present themselves:

- While much progress has been made in recent years, modelling the complex spatio-temporal variability of routine public health data under normal circumstances remains challenging.

- Models such as PANDA that include individual-level sub-models can present significant computational challenges, especially in the presence of contagions.
- Absent a BT event, model checking and critiquing remains a subjective undertaking; establishing the operational characteristics of models and algorithms relies on high-fidelity simulations that are hard to develop and hard to check.
- Computer-based explanations of alerts represent an important area for future research. Once an alert is raised, it would be helpful to explain to the public-health officials in terms that are meaningful to their responsibilities the reason *why* the surveillance system raised it.
- The hidden Markov model framework may prove useful in surveillance applications and avoid some of the complexities of individual-level models [71]. However, the operational characteristics of the approach remain unknown.

Decision theoretic surveillance

Extension of alerting systems to decision-analysis-based tools is another important area for future research. Such tools could recommend actions to take, including when/where to gather additional information, who/where/how to treat subpopulations, and who/where to quarantine. Again, explanation for why these recommendations are being made would likely be critical, including making clear the assumptions underlying the recommendations, such as the cost functions involved. Ultimately, a sequential decision-making approach will be appropriate where decision-theoretic considerations guide the information gathering sequence.

Theory

The mathematical study of traditional one-dimensional surveillance methods such as the Shewhart method or CUSUM has reached a high level of sophistication [72]. Much work remains to better understand more complex kinds of spatio-temporal surveillance algorithms.

Dual-use

The US is making significant investments in large-scale public health surveillance systems in the hope of detecting a future BT attack and providing situational awareness to minimize adverse public health effects. Consideration of the likely nature of such an attack informs much of the current work in a very dissociative fashion and indeed much of the current activity seems to focus at least implicitly on anthrax. Characterizing the types of BT events that a particular surveillance system could have a reasonable chance of detecting is an important research question. Security questions also arise; if the range of detectable BT events is public knowledge, how does this affect the probability of such an event occurring? Furthermore, we all hope that a BT event never occurs and in fact some public health practitioners contend that the systems now being built may never detect a BT event. However, the impact of surveillance systems on our understanding of public health is potentially significant, important, and certainly worthy of study. The usage of surveillance systems for monitoring both routine health vents and unlikely bioterrorist attacks brings the dilemma of whether to treat customary seasonal outbreaks as signal or noise. It seems unlikely that public health monitors will continue to use a system designed to detect only scenario-based point-source outbreaks. A viable dual-use system must include multiple algorithms and/or data models to detect both natural and deliberate events and must support the user's ability to discriminate between them.

DOI: 10.1002/sim

Consideration of the broader objectives of public health surveillance must ultimately guide research and development of analytic methods for surveillance. Traditional public health concerns as well as security concerns and the need for situational awareness and early detection of potential BT threats provide a complex backdrop. We believe this is an exciting area for statistical research.

ACKNOWLEDGEMENTS

We would like to thank the National Science Foundation for supporting The Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) which enabled this collaborative report as well as grant IIS-0325581. We additionally acknowledge supporting grants F30602-01-2-0550 from the Department of Homeland Security, and ME-01-737 from the Pennsylvania Department of Health.

REFERENCES

1. Fienberg SE, Shmueli G. Statistical issues and challenges associated with rapid detection of bio-terrorist attacks. *Statistics in Medicine* 2005; **24**:513–529. DOI: 10.1002/sim.2032
2. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Sciences of the United States of America* 2002; **99**(8):5237–5240.
3. Begier EM, Sockwell D, Branch LM, Davies-Cole JO, Jones LH, Edwards L, Casani JA, Blythe D. The National Capitol Region's emergency department syndromic surveillance system: do chief complaint and discharge diagnosis yield different results? *Emerging Infectious Disease* 2003; **9**(3):393–396.
4. Greenko J, Mostashari F, Fine A, Layton M. Clinical evaluation of the Emergency Medical Services (EMS) ambulance dispatch-based syndromic surveillance system, New York City. *Journal of Urban Health* 2003; **80**(Suppl 1):I50–I56.
5. Magruder SF, Lewis SH, Najmi A, Florio E. Progress in understanding and using over-the-counter pharmaceuticals for syndromic surveillance. *Syndromic Surveillance*: *Reports from a National Conference*, 2003; *Morbidity and Mortality Weekly Report* 2004; **53**(Suppl):117–122.
6. Wagner MM, Aryel R, Dato VM, Krenzelok E, Fapohunda A, Sharma R. Availability and comparative value of data elements required for an effective bioterrorism detection system. Report commissioned by AHRQ. Delivered 28 November 2001; 1–184 (http://rods.health.pitt.edu/LIBRARY/AHRQInterimRpt112801.pdf).
7. Edge VL, Pollari F, Lim G, Aramini J, Sockett P, Martin SW, Wilson J, Ellis A. Syndromic surveillance of gastrointestinal illness using pharmacy over-the-counter sales. A retrospective study of waterborne outbreaks in Saskatchewan and Ontario. *Canadian Journal of Public Health* 2004; **95**(6):446–450.
8. Sokolow LZ, Grady N, Rolka H, Walker D, McMurray P, English R, Loonsk J. Deciphering data anomalies in BioSense. *Morbidity and Mortality Weekly Report* 2005; **54**(Suppl):133–139.
9. Hurt-Mullen KJ, Holtry RS, Babin SM, Coberly JS. Syndromic surveillance on the epidemiologist's desktop: making sense of much data. *Morbidity and Mortality Weekly Report* 2005; **54**(Suppl):141–146.
10. Balter S. Three years of emergency department gastrointestinal syndromic surveillance in NYC: what have we found? *Morbidity and Mortality Weekly Report* 2005; **54**(Suppl):175–180.
11. Burkom HS, Murphy SP, Coberly JS, Hurt-Mullen KJ. Public health monitoring tools for multiple data streams. *Morbidity and Mortality Weekly Report* 2005; **54**(Suppl):55–62.
12. Miller CJ, Genovese C, Nichol RC, Wasserman L, Connolly A, Reichart D, Hopkins A, Schneider J, Moore A. Controlling the false discovery rate in astrophysical data analysis. *Astronomical Journal* 2001; **122**:3492–3505.
13. Ryan TP. *Statistical Methods for Quality Improvement*. Wiley: New York, 1989.
14. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988; **75**:383–386.
15. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**:751–754.
16. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**:800–802.
17. Sarkar SK, Chang CK. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Medical Association* 1997; **92**(440):1601–1608.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 1995; **57**:289–300.

19. Marshall C, Best N, Bottle A, Aylin P. Statistical issues in prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society* 2004; **167**(Pt. 3):541–559.
20. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
21. Edgington ES. An additive method for combining probability values from independent experiments. *Journal of Psychology* 1972; **80**:351–363.
22. Edgington ES. A normal curve method for combining probability values from independent experiments. *Journal of Psychology* 1972; **82**:85–89.
23. Williamson GD, Weatherby-Hudson G. A monitoring system for detecting aberrations in public health surveillance reports. *Statistics in Medicine* 1999; **18**:3283–3298.
24. Rogerson PA, Yamada I. Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine* 2004; **23**(14):2195–2214.
25. Benneyan JC. Statistical quality control methods in infection control and hospital epidemiology, part I: introduction and basic theory. *Infection Control and Hospital Epidemiology* 1998; **19**(4):194–214.
26. Ye N, Cheng Q, Emran S, Vilbert S. Hotelling's T^2 multivariate profiling for anomaly detection. *Proceedings 2000 IEEE Workshop on Information Assurance and Security*, West Point, New York, June 2002.
27. Lowry CA, Woodall WH. A multivariate exponentially weighted moving average control chart. *Technometrics* 1992; **34**(1):46–53.
28. Hong B, Hardin M. A study of the properties of the multivariate forecast-based processing scheme. *Presented at Joint Statistical Meetings*, Toronto, August 2004.
29. Pignatiello JJ, Runger GC. Comparisons of multivariate CUSUM charts. *Journal of Quality Technology* 1990; **22**(3):173–186.
30. Hawkins D. Multivariate quality control based on regression-adjusted variables. *Technometrics* 1991; **33**(1):61–75.
31. Crosier RB. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics* 1988; **30**(3):291–303.
32. Burkom H, Hutwagner L, Rodriguez R. Using point-source epidemic curves to evaluate alerting algorithms for biosurveillance. *2004 Proceedings of the American Statistical Association* (Statistics in Government Section (CD-ROM)). American Statistical Association: Toronto, 2005.
33. Sartwell PE. The distribution of incubation periods of infectious disease. *American Journal of Hygiene* 1950; **51**:310–318; reprinted in *American Journal of Epidemiology* 1995; **141**(5).
34. USAMRIID's Medical Management for Biological Casualties Handbook (4th edn), U.S. Army Medical Research Institute of Infectious Diseases, February 2001, Ft. Detrick, MD (http://ehs.ucdavis.edu/ucbso/ReferenceDoc/USAMRIID_Blue_book.pdf).
35. Lombardo J, Burkom H, Pavlin J. ESSENCE II and the framework for evaluating syndromic surveillance systems. *Morbidity and Mortality Weekly Report* 2004; **53**(Suppl):159–165.
36. Bradley CA, Rolka HR, Walker DW, Loonsk J. BioSense: implementation of a national early event detection and situational awareness system. *Morbidity and Mortality Weekly Report* 2005; **54**(Suppl):11–19.
37. Naus J. Clustering of random points in two dimensions. *Biometrika* 1965; **52**:263–267.
38. Glaz J, Balakrishnan N (eds). *Scan Statistics and Applications*. Birkhauser: Basel, 1999.
39. Glaz J, Naus J, Wallenstein S. *Scan Statistics*. Springer: Berlin, 2001.
40. Weinstock MA. A generalized scan statistic test for the detection of clusters. *International Journal of Epidemiology* 1981; **10**:289–293.
41. Wallenstein S. A test for detection of clustering over time. *American Journal of Epidemiology* 1980; **111**:367–372.
42. Kulldorff M. A spatial scan statistic. *Communications in Statistics*: *Theory and Methods* 1997; **26**:1481–1496.
43. Hsu CE, Jacobson HE, Soto Mas F. Evaluating the disparity of female breast cancer mortality among racial groups—a spatiotemporal analysis. *International Journal of Health Geographics* 2004; **3**:4.
44. Klassen A, Kulldorff M, Curriero F. Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *International Journal of Health Geographics* 2005; **4**:1.
45. Washington CH, Radday J, Streit TG, Boyd HA, Beach MJ, Addiss DG, Lovince R, Lovegrove MC, Lafontant JG, Lammie PJ, Hightower AW. Spatial clustering of filarial transmission before and after a mass drug administration in a setting of low infection prevalence. *Filaria Journal* 2004; **3**:3.
46. Odoi A, Martin SW, Michel P, Middleton D, Holt J, Wilson J. Investigation of clusters of giardiasis using GIS and a spatial scan statistic. *International Journal of Health Geographics* 2004; **3**:11.
47. Andrade AL, Silva SA, Martelli CM, Oliveira RM, Morais Neto OL, Siqueira Jr JB, Melo LK, Di Fabio JL. Population-based surveillance of pediatric pneumonia: use of spatial analysis in an urban area of Central Brazil. *Cadernos de Saúde Pública* 2004; **20**:411–421.

DOI: 10.1002/sim

48. Ali M, Asefaw T, Byass P, Beyene H, Pedersen FK. Helping northern Ethiopian communities reduce childhood mortality: population-based intervention trial. *Bulletin of the World Health Organization* 2005; **83**(1):27–33.
49. Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society* 2001; **A164**:61–72.
50. Kulldorff M, Heffernan R, Hartman J, Assunção R, Mostashari F. A space–time permutation scan statistic for the early detection of disease outbreaks. *Public Library of Science* (*Medicine*) 2005; **2**:e59.
51. Mostashari F, Kulldorff M, Hartman JJ, Miller JR, Kulasekera V. Dead bird clustering: a potential early warning system for West Nile virus activity. *Emerging Infectious Diseases* 2003; **9**:641–646.
52. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice: The New York City emergency department system. *Emerging Infectious Diseases* 2004; **10**:858–864.
53. Dwass M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 1957; **28**:181–187.
54. Burkom H. Biosurveillance applying scan statistics with multiple disparate data sources. *Journal of Urban Health* 2003; **80**:57–i65.
55. Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, Platt R. Multivariate spatial scan statistics for disease surveillance. *Statistics in Medicine*, submitted.
56. Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC Public Health* 2001; **1**(1):9.
57. Patil GP, Taillie C. Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science* 2003; **18**:457–465.
58. Duczmal L, Assuncao R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* 2004; **45**:269–286.
59. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 2005; **4**:11.
60. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine* 2006; **25**(22):3929–3943.
61. Neill D, Moore A. A fast multi-resolution method for detection of significant spatial disease clusters. *Advances in Neural Information Processing Systems* 2003; **16**.
62. Kulldorff M, Fang Z, Walsh S. A tree-based scan statistic for database disease surveillance. *Biometrics* 2003; **9**:641–646.
63. Charniak E. Bayesian networks without tears. *AI Magazine* 1991; **Winter**:50–63.
64. Murphy K. A Brief Introduction to Graphical Models and Bayesian Networks 1998 (available at http://www.cs.ubc.ca/∼murphyk/Bayes/bayes.html).
65. Neapolitan RE. *Learning Bayesian Networks*. Prentice-Hall: Upper Saddle River, NJ, 2004.
66. Wong W, Cooper G, Dash D, Levander J, Dowling J, Hogan W, Wagner M. Bayesian biosurveillance using multiple data streams. *Morbidity and Mortality Weekly Report Supplement* 2005 (available at http://www.cbmi.pitt.edu/panda/publications.html).
67. Cooper G, Dash D, Levander J, Wong W, Hogan W, Wagner M. Bayesian biosurveillance of disease outbreaks. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2004; 94–103 (available at http://www.cbmi.pitt.edu/panda/publications.html).
68. DeGroot MH. *Probability and Statistics* (2nd edn). Addison-Wesley: Reading, MA, 1989.
69. Timm NH. *Applied Multivariate Analysis*. Springer: New York, NY, 2002.
70. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* (*Series B*) 1995; **57**:289–300.
71. Madigan D. Bayesian data mining for surveillance. In *Spatial and Syndromic Surveillance for Public Health*, Lawson A, Kleinman K (eds). Wiley: West Sussex, England, 2005; 203–221.
72. Frisen M. Statistical surveillance: optimality and methods. *International Statistical Review* 2003; **71**:403–434.