Mach Learn (2010) 79: 261–282
DOI 10.1007/s10994-009-5144-4

A multivariate Bayesian scan statistic for early event detection and characterization

Daniel B. Neill · Gregory F. Cooper

Received: 27 November 2007 / Revised: 8 September 2008 / Accepted: 10 August 2009 /
Published online: 13 November 2009
Springer Science+Business Media, LLC 2009

Abstract We present the multivariate Bayesian scan statistic (MBSS), a general framework for event detection and characterization in multivariate spatial time series data. MBSS integrates prior information and observations from multiple data streams in a principled Bayesian framework, computing the posterior probability of each type of event in each space-time region. MBSS learns a multivariate Gamma-Poisson model from historical data, and models the effects of each event type on each stream using expert knowledge or labeled training examples. We evaluate MBSS on various disease surveillance tasks, detecting and characterizing outbreaks injected into three streams of Pennsylvania medication sales data. We demonstrate that MBSS can be used both as a "general" event detector, with high detection power across a variety of event types, and a "specific" detector that incorporates prior knowledge of an event's effects to achieve much higher detection power. MBSS has many other advantages over previous event detection approaches, including faster computation and easy interpretation and visualization of results, and allows faster and more accurate event detection by integrating information from the multiple streams. Most importantly, MBSS can model and differentiate between multiple event types, thus distinguishing between events requiring urgent responses and other, less relevant patterns in the data.

Keywords Event detection · Event characterization · Biosurveillance · Scan statistics

Editors: Dragos Margineantu, Denver Dash, and Weng-Keen Wong.

D.B. Neill (✉)
H.J. Heinz III College, School of Public Policy and Management, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
e-mail: neill@cs.cmu.edu

G.F. Cooper
Department of Biomedical Informatics, University of Pittsburgh, M-183 Parkvale Building,
200 Meyran Avenue, Pittsburgh, PA 15260, USA
e-mail: gfc@cbmi.pitt.edu

1 Introduction

Event surveillance tasks require continuous monitoring of massive quantities of multivariate data in order to detect and identify emerging patterns. For example, government agencies responsible for public health and safety must respond rapidly to a variety of potential threats, including wars, disease outbreaks, crime waves, natural disasters, and terrorist attacks. A timely and informed response to these events can substantially reduce the resulting human, financial, and societal costs, while a delayed or incorrect response may have catastrophic results. Automatic systems for event surveillance have the potential to improve the timeliness and correctness of response by rapidly identifying relevant patterns in the massive amount of data being monitored. In such systems, both *early detection* and *accurate characterization* of events are essential: relevant events must be quickly detected and correctly identified, enabling a timely and appropriate response, while false positives (due to irrelevant events or other patterns in the data) must be kept to a minimum. To achieve accurate detection and characterization of events, surveillance systems must integrate information from multiple streams of spatial and temporal data in order to obtain a coherent and complete situational awareness, identifying which events have occurred and which subsets of the data (e.g. spatial regions) have been affected by each event.

As a concrete example of the event surveillance problem, we focus on the motivating application of *disease surveillance*. In this application domain, we wish to develop systems that monitor electronically available public health data sources (such as hospital visits and medication sales) and automatically detect emerging outbreaks of disease. Both early detection and accurate characterization of events are important in this domain: major health threats such as pandemic avian influenza or a bioterrorist attack require rapid responses (such as treatment of potentially infected individuals, health advisories, travel restrictions, and quarantines) in order to control the spread of the outbreak and reduce its impact. However, taking appropriate actions often requires knowledge of the characteristics of the disease (e.g. source, method of transmission, and available treatments) and which areas have been affected. Similarly, serious outbreaks requiring urgent responses must be distinguished from less serious outbreaks (e.g. seasonal influenza) and from irrelevant patterns in the data (e.g. increases in over-the-counter medication sales due to store promotions). Thus disease surveillance systems must not only detect emerging outbreaks, but also determine the type of outbreak and its area of effect, to facilitate a timely and correct public health response.

Our disease surveillance system, described in Sabhnani et al. (2005), monitors daily data feeds from over 20,000 hospitals and pharmacies nationwide. Pharmacy data is made available through the National Retail Data Monitor (Wagner et al. 2004). We monitor two different data types: Emergency Department (ED) visits, classified by chief complaint type (e.g. respiratory, fever), and over-the-counter (OTC) medication sales, classified by product type (e.g. cough/cold, thermometers). Counts are aggregated at the zip code level, and each ED chief complaint type and each OTC product type is treated as a separate data stream. Thus we have a time series of daily counts for each data stream for each zip code, where each count represents the number of ED visits (or OTC sales) of a given type in that zip code on that day. Our current system monitors each data stream separately, using an *expectation-based scan statistic* (Neill et al. 2005b) to search for space-time regions where the recent counts for that stream are significantly higher than expected. This method first forecasts the expected counts for each data stream for each zip code using historical data, and then detects spatial clusters of zip codes with higher than expected counts.

While this system has been demonstrated to achieve early detection of real and simulated disease outbreaks (Neill 2006), several additional criteria must be met to improve the

timeliness of detection and the usefulness of the detected patterns. First, event surveillance systems should integrate information from multiple streams of spatial time series data, instead of monitoring each data stream separately, in order to achieve higher detection power for events that simultaneously affect multiple streams. Second, systems must be able to model and differentiate between multiple types of event, and to incorporate prior knowledge of the effects of each event type. These priors can either be specified by a domain expert, or learned from labeled training examples. As noted above, characterization of events is essential to distinguish patterns that are relevant to the user from those that are irrelevant, and to inform the user's response to any relevant events. Finally, the methods used for event detection and characterization must be computationally efficient, in order to detect patterns in large real-world datasets in near real time. Though no previous approach meets all of these criteria, some individual criteria have been met by previously proposed methods. In particular, our recently proposed Bayesian scan statistic (Neill et al. 2006) enables the incorporation of prior information into the event detection task, but only considers univariate data. As we demonstrate in this paper, we can achieve more useful characterization of events, as well as more timely and accurate event detection, by extending the Bayesian framework to multiple data streams and multiple event types.

Thus we develop a new methodology for multivariate event detection and characterization using spatial time series data, the "multivariate Bayesian scan statistic" (MBSS). The MBSS method integrates information from multiple data streams in a coherent and computationally efficient Bayesian framework, enabling faster and more accurate event detection. MBSS also incorporates prior information and incremental learning in order to model and distinguish between multiple event types, thus providing users with sufficient situational awareness to enable a rapid and informed response. In the following sections, we describe the MBSS framework in detail, and then evaluate the performance of MBSS on event detection and characterization tasks in the disease surveillance domain.

2 The multivariate Bayesian scan statistic framework

In the multivariate event surveillance problem, our main goal is to detect and characterize events (such as disease outbreaks) based on their effects on the monitored data sources. We typically monitor count data for multiple spatial locations, time steps, and data streams. For example, to detect an outbreak of avian influenza, we might monitor hospital Emergency Department (ED) visits, with each data stream representing the number of ED visits with a different chief complaint type (respiratory, fever, etc.), and over-the-counter (OTC) medication sales, with each stream representing the number of sales of a different product group (cough/cold medications, thermometers, etc.).

In the general case, we are given a dataset D consisting of multiple data streams D_m, for $m = 1 \ldots M$. Each data stream consists of spatial time series data collected at a set of spatial locations s_i, for $i = 1 \ldots I$. For each stream D_m and location s_i, we have a time series of counts $c_{i,m}^t$, where $t = 0$ represents the current time step and $t = 1 \ldots T$ represent the counts from 1 to T time steps ago respectively. For example, in disease surveillance, we typically have data collected on a daily basis, and aggregated at the zip code level due to data privacy concerns. Thus a given count $c_{i,m}^t$ might represent the number of respiratory ED visits, or the number of cough/cold drugs sold, for a given zip code on a given day.

As noted above, our goals in the MBSS framework are event detection and characterization: we wish to detect any relevant events occurring in the data, identify the type of event, and determine the event duration and affected locations. Thus we wish to compare

the set of alternative hypotheses $H_1(S, E_k)$, each representing the occurrence of some event of type E_k in some space-time region S, against the null hypothesis H_0 that no events have occurred. We assume that the set of event types $E = \{E_k\}$, for $k = 1 \ldots K$, is given, and that these events are mutually exclusive (i.e. at most one event occurs in the data). Moreover, each distinct hypothesis $H_1(S, E_k)$ assumes that the given event type E_k has affected all and only those locations $s_i \in S$, and thus all hypotheses $H_1(S, E_k)$ are mutually exclusive.

Many other problem formulations are possible, e.g. assuming that each event type has an independent probability of occurrence, in which case we must deal with the added complexity that multiple events can affect the same location. Our simplifying assumption of mutually exclusive events enables an efficiently computable model and easily interpretable results, but has reduced power to detect multiple simultaneously occurring events, unless each compound event is modeled as an additional event type. We believe that this formulation is appropriate in cases where events are rare (e.g. in the disease surveillance domain), but less appropriate if events are very common.

Each event type can be thought of as a process that affects some subset of the data in some probabilistic manner. This leads to the key insight that we can both detect and characterize events by searching over subsets of the data, identifying subsets with high likelihood of some event type E_k. In the space-time event detection framework, we assume that the event causes an increase in counts (for some subset of data streams) in the affected area, and thus we search for space-time regions with higher than expected counts. In our disease surveillance example, the event types may be either specific illnesses (e.g. influenza, anthrax), non-specific syndromes (e.g. influenza-like illness, gastrointestinal illness), or other non-outbreak events that may result in patterns of increased counts, such as promotional sales of OTC medications, inclement weather, or tourism.

In addition to the set of event types, we are also given a set of space-time regions $\boldsymbol{\mathcal{S}}$ to search. Each region $S \in \boldsymbol{\mathcal{S}}$ contains some non-empty subset of the spatial locations s_i, and also has a time duration $W(S)$, indicating that these locations have been affected by an event during time steps $t = 0 \ldots W(S) - 1$. Note that a spatial location s_i may be contained in multiple distinct regions, and thus we typically search over a set of overlapping regions. While there are exponentially many possible regions S that could be considered, we generally do not want to evaluate all such regions, both due to computational and practical considerations. For example, we would not usually expect a disease outbreak to affect any arbitrary and spatially dispersed set of zip codes, but instead only sets of nearby zip codes.

When choosing the set of search regions $\boldsymbol{\mathcal{S}}$, we generally consider all spatial regions of a given shape (e.g. circles, rectangles) and varying sizes. For example, Kulldorff's original spatial scan statistic (Kulldorff 1997) searches over circular regions of continuously varying radius, centered at each location s_i. In this case, the number of distinct regions varies quadratically with the number of locations. We consider regions with time durations $W(S) = 1 \ldots W_{\max}$, for some constant $W_{\max}$. For example, if we are performing daily surveillance, $W_{\max} = 7$ would consider temporal windows up to 1 week in duration. This formulation assumes that we are performing prospective surveillance (each time series stretches back from the current time step, $t = 0$), and that we are interested only in events that are current (still affecting the data during the current time step $t = 0$) and recent (with time duration up to $W_{\max}$). Retrospective surveillance methods would instead allow both the start and end times of the temporal window to vary. We can also generalize the scan statistic to allow the set of locations affected to change over the duration of the event, in which case region S can denote a separate set of locations $\{s_i\}_t$ for each time step $t = 0 \ldots W(S) - 1$. Here we focus on the simpler case, where the set of affected locations remains constant over time.

Given the set of event types E, set of space-time regions $\boldsymbol{\mathcal{S}}$, and the multivariate dataset D, our goal is to compute the posterior probability $\Pr(H_1(S, E_k) \mid D)$ that each

event type E_k has affected each space-time region S, as well as the posterior probability $\Pr(H_0 \mid D)$ that no event has occurred. To do so, we must have the prior probability of each event type occurring in each space-time region, $\Pr(H_1(S, E_k))$, as well as the prior probability $\Pr(H_0)$ that no events have occurred. We must also be able to compute the likelihood of the multivariate data given each alternative hypothesis, $\Pr(D \mid H_1(S, E_k))$ for all S and E_k under consideration, and the likelihood of the data given the null hypothesis, $\Pr(D \mid H_0)$. We then apply Bayes' Theorem to compute the posterior probability of each hypothesis:

$$\Pr(H_1(S, E_k) \mid D) = \frac{\Pr(D \mid H_1(S, E_k))\Pr(H_1(S, E_k))}{\Pr(D)}$$

$$\Pr(H_0 \mid D) = \frac{\Pr(D \mid H_0)\Pr(H_0)}{\Pr(D)}$$

In this expression, the posterior probability of each hypothesis is normalized by the total probability of the data, $\Pr(D) = \Pr(D \mid H_0)\Pr(H_0) + \sum_{S,E_k} \Pr(D \mid H_1(S, E_k))\Pr(H_1(S, E_k))$. In the following sections, we consider how the priors $\Pr(H)$ and the likelihoods $\Pr(D \mid H)$ can be computed for each hypothesis under consideration.

3 Obtaining priors

As discussed above, our inputs into the MBSS method include the prior probability of each event type E_k occurring in each space-time region S, $\Pr(H_1(S, E_k))$, as well as the probability that no events occur, $\Pr(H_0)$. Each prior probability $\Pr(H_1(S, E_k))$ can be decomposed as the product of the prior probability of event type E_k and the conditional probability that subset S is affected by E_k: $\Pr(H_1(S, E_k)) = \Pr(E_k)\Pr(H_1(S, E_k) \mid E_k)$. In this expression, $\Pr(E_k)$ represents the overall prevalence of event type E_k, while $\Pr(H_1(S, E_k) \mid E_k)$ represents its distribution in space and time. As noted above, we assume that all event types are mutually exclusive, so that $\Pr(H_0) + \sum_k \Pr(E_k) = 1$. We also assume that each event only affects a single space-time region S, so that $\sum_S \Pr(H_1(S, E_k) \mid E_k) = 1$ for each event type E_k.

The prevalence of each event type can be obtained from expert knowledge, and can be allowed to vary over time (e.g. seasonal influenza peaks between December and March). Alternatively, we can *learn* these probabilities from labeled data: given a large training set of N days of data, where each day is labeled with an event type (E_k or H_0), we can use the smoothed maximum likelihood estimates $\Pr(E_k) = \frac{N_k + M_k}{N + M}$, where N_k and M_k are the observed count and prior weight of event type k respectively, and N and M are the total number of days observed and the total prior weight respectively (Neill 2007b). For the experiments described below, we assume a uniform prior over event types, with a fixed prior probability $\Pr(H_1) = 0.01$. In this case, we have $\Pr(H_0) = 0.99$, and $\Pr(E_k) = \frac{0.01}{K}$ for all $k = 1 \ldots K$.

Similarly, for the distribution of a given event type over regions S, we can either assume a uniform region prior $\Pr(H_1(S, E_k) \mid E_k) = \frac{1}{N_S}$, where N_S is the total number of space-time regions, obtain region priors from expert knowledge (e.g. the affected area for a waterborne illness can be predicted based on water distribution information), or use a smoothed maximum likelihood estimate from labeled data. Because the number of possible space-time regions is huge, we typically parameterize the prior based on the region size and shape, and learn each parameter separately (Neill 2007b). For the experiments described below, we assume a uniform region prior, and thus we have prior probabilities $\Pr(H_1(S, E_k)) = \frac{0.01}{K N_S}$ for all S and E_k under consideration. More complex methods of prior elicitation and prior learning will be addressed in future work.

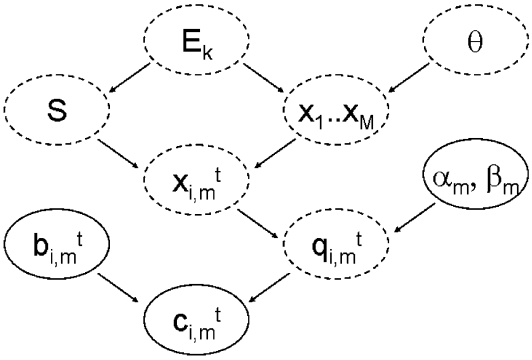

Fig. 1 Bayesian network representation of the MBSS method. *Solid ovals* represent observed quantities, and *dashed ovals* represent hidden quantities. The counts $c^t_{i,m}$ are directly observed, while the baselines $b^t_{i,m}$ and the parameter priors for each stream (α_m, β_m) are estimated from historical data

4 Computing likelihoods

To compute the likelihood of the data under the null hypothesis H_0 or an alternative hypothesis $H_1(S, E_k)$, we assume that counts have been generated from a hierarchical Gamma-Poisson model. A Bayesian network representation of this model is shown in Fig. 1, and we will now describe each aspect of the model in detail. The event type k (where $k = 0 \ldots K$) is drawn from a multinomial distribution: here we let $k = 0$ represent the null hypothesis H_0 of no events, with probability $\Pr(H_0)$, and $k = 1 \ldots K$ represent the occurrence of event type E_k, with probability $\Pr(E_k)$. The region of effect S is conditional on the event type k, with probabilities $\Pr(H_1(S, E_k) \mid E_k)$ as described above. Under the null hypothesis H_0, we assume that no locations are affected, i.e. $S = \emptyset$.

The effects of an event $H_1(S, E_k)$ on the data are determined by a value $x^t_{i,m}$ for each location s_i, data stream D_m, and time step t. In this general formulation, the event's effects can vary spatially and temporally as well as varying across data streams. These effects are assumed to be multiplicative, increasing the expected value of each count $c^t_{i,m}$ by a factor of $x^t_{i,m}$, and thus a value $x^t_{i,m} = 1$ would signify no effect of the event for the given location, stream, and time step. For the null hypothesis H_0, no events have occurred, and thus we assume that $x^t_{i,m} = 1$ everywhere. For an event $H_1(S, E_k)$, we assume that only locations and time steps inside the space-time region S have been affected. Thus we assume $x^t_{i,m} = 1$ for all locations $s_i \notin S$ and for all time steps $t \geq W(S)$, where $W(S)$ is the time duration for region S. We often make the further simplifying assumption that the effects of an event $H_1(S, E_k)$ are constant for each data stream in the affected region S. In this case, we have a vector $x = (x_1 \ldots x_M)$ representing the effects of the event on each data stream D_m. Then $x^t_{i,m} = x_m$ for all $s_i \in S$ and $t < W(S)$, and $x^t_{i,m} = 1$ otherwise. In the simplified model discussed here, each event type E_k can have a different joint probability distribution over vectors $x = (x_1 \ldots x_M)$, while in the general case, each event type E_k can have a different joint distribution over all $x^t_{i,m}$.

We now consider how to compute the likelihood of the data under the null hypothesis H_0 or under an alternative hypothesis $H_1(S, E_k)$. We first derive an expression for the likelihood of the data given the values of $x^t_{i,m}$. The likelihood of the data given H_0 follows directly from this expression, since the null hypothesis assumes $x^t_{i,m} = 1$ everywhere. To calculate the marginal likelihood of the data given an alternative hypothesis $H_1(S, E_k)$, we must marginalize over the distribution of effects $x^t_{i,m}$, computing a weighted average of the data likelihoods given each effects vector $(x_1 \ldots x_M)$, weighted by the conditional probability of those effects given $H_1(S, E_k)$.

4.1 The Gamma-Poisson model

We now compute the likelihood of the dataset $D = \{c^t_{i,m}\}$ given the effects $x^t_{i,m}$ for each location s_i, data stream D_m, and time step t. As noted above, under the null hypothesis H_0 we have $x^t_{i,m} = 1$ everywhere, while the distribution of $x^t_{i,m}$ under an alternative hypothesis $H_1(S, E_k)$ will be discussed below. As is evident from the Bayesian network representation in Fig. 1, the counts $c^t_{i,m}$ are assumed to have been generated from a hierarchical model, with parameters dependent on the effects $x^t_{i,m}$ as well as various quantities estimated from historical data. These quantities include the baseline $b^t_{i,m}$ and the relative risk $q^t_{i,m}$ for each location, stream, and time step, as well as the parameter priors α_m and β_m for each data stream D_m. The baseline $b^t_{i,m}$ represents the expected value of the count $c^t_{i,m}$ assuming that no events are taking place, and is learned from time series analysis of historical data. Each count $c^t_{i,m}$ is assumed to be generated from a distribution with mean proportional to $b^t_{i,m}$ times the relative risk $q^t_{i,m}$, where the relative risk is a latent (unobserved) variable dependent on the effect $x^t_{i,m}$. If $x^t_{i,m} = 1$, the distribution of relative risks is determined by the parameter priors α_m and β_m, also learned from historical data. For $x^t_{i,m} \neq 1$, the expected value of the relative risk $q^t_{i,m}$, and thus the expected value of the count $c^t_{i,m}$, is multiplied by $x^t_{i,m}$. This model, where the relative risks can vary spatially and temporally as well as between streams, is more flexible than the previously proposed frequentist and Bayesian scan statistic models (Kulldorff 1997; Neill et al. 2006), in which the relative risk for each stream was assumed to be constant both inside and outside the affected region.

Within this general framework, we have much flexibility in defining the generative model for the counts $c^t_{i,m}$. Here we assume a hierarchical Gamma-Poisson model, since these models are commonly used to represent the distribution of counts in the disease surveillance domain. Gamma-Poisson models have been used successfully for disease surveillance by Clayton and Kaldor (1987), Mollié (1999), and others, though most of these models focus on inferring the relative risks rather than detecting clusters of increased counts.

At the bottom level of our Gamma-Poisson model, each count $c^t_{i,m}$ is assumed to have been drawn from a Poisson distribution with mean proportional to the product of the expected count $b^t_{i,m}$ and the relative risk $q^t_{i,m}$: $c^t_{i,m} \sim \text{Poisson}(q^t_{i,m} b^t_{i,m})$.

As noted above, the baselines $b^t_{i,m}$ are learned from the historical data by time series analysis. Here we use the method suggested by Kulldorff et al. (2005), in which the expected count for a given location on a given day is equal to the total count for that day multiplied by the proportion of all counts corresponding to that location:

$$b^t_{i,m} = \frac{\sum_i c^t_{i,m} \sum_t c^t_{i,m}}{\sum_i \sum_t c^t_{i,m}}$$

This baseline method has the advantage of adjusting for seasonal and day of week trends, reducing the impact of misestimation of baselines on the detection methods. On the other hand, it has reduced power to detect spatially dispersed events, since it conditions on the total count of the current day. In application domains where events may simultaneously affect a large fraction of the monitored locations, other time series analysis methods (such as a moving average, adjusted for day of week and seasonality) should be used. Also, we note that only counts from the given location s_i and data stream D_m are used to compute the expected counts $b^t_{i,m}$. We could also share information across multiple locations and multiple streams when computing expected counts.

For given values of the parameter priors (α_m, β_m) and the effect $x^t_{i,m}$, the relative risk $q^t_{i,m}$ is assumed to have been drawn from a Gamma distribution with parameters $\alpha = x^t_{i,m}\alpha_m$

and $\beta = \beta_m$: $q^t_{i,m} \sim \text{Gamma}(x^t_{i,m}\alpha_m, \beta_m)$. Thus we have $q^t_{i,m} \sim \text{Gamma}(\alpha_m, \beta_m)$ under the null hypothesis H_0, i.e. all relative risks for a given data stream are drawn from the same distribution. If some event is taking place and has effect $x^t_{i,m} \neq 1$, the mean (and variance) of the relative risk distribution are multiplied by $x^t_{i,m}$, thus multiplying the expected value of the count $c^t_{i,m}$ by a factor of $x^t_{i,m}$.

We now compute the marginal likelihood of each observed count $c^t_{i,m}$, given the effect $x^t_{i,m}$, the baseline $b^t_{i,m}$, and the parameter priors α_m and β_m. To do so, we must integrate over all possible values of the relative risk $q^t_{i,m}$, weighted by their respective probabilities. But since we have a conjugate prior, we can obtain a closed form solution for the marginal likelihood, as given below. For simplicity of notation, we drop the sub- and superscripts in our derivation, and simply write $c = c^t_{i,m}$, $b = b^t_{i,m}$, $q = q^t_{i,m}$, $x = x^t_{i,m}$, $\alpha = \alpha_m$, and $\beta = \beta_m$:

$$\begin{aligned}
\Pr(c \mid b, x, \alpha, \beta) &= \int \Pr(q \sim \text{Gamma}(x\alpha, \beta))\Pr(c \sim \text{Poisson}(qb))\, dq \\
&= \int \frac{\beta^{x\alpha}}{\Gamma(x\alpha)} q^{x\alpha-1} e^{-\beta q} \frac{(qb)^c e^{-qb}}{c!}\, dq \\
&= \frac{\beta^{x\alpha} b^c}{\Gamma(x\alpha)c!} \int q^{x\alpha-1} e^{-\beta q} q^c e^{-qb}\, dq \\
&= \frac{\beta^{x\alpha} b^c}{\Gamma(x\alpha)c!} \int q^{x\alpha+c-1} e^{-(\beta+b)q}\, dq = \frac{\beta^{x\alpha} b^c \Gamma(x\alpha + c)}{(\beta + b)^{x\alpha+c}\Gamma(x\alpha)c!}
\end{aligned}$$

Thus each count $c^t_{i,m}$ follows a negative binomial distribution with parameters $x^t_{i,m}\alpha_m$ and $\frac{\beta_m}{\beta_m + b^t_{i,m}}$. The mean of this distribution is $x^t_{i,m}\frac{\alpha_m}{\beta_m} b^t_{i,m}$, and the variance is $x^t_{i,m}(\frac{\alpha_m}{\beta_m^2}(b^t_{i,m})^2 + \frac{\alpha_m}{\beta_m} b^t_{i,m})$.

Since the counts are conditionally independent given the values of $b^t_{i,m}$, $x^t_{i,m}$, α_m, and β_m, the likelihood of the entire dataset $D = \{c^t_{i,m}\}$ for a given set of effects $X = \{x^t_{i,m}\}$ is the product of these conditional probabilities:

$$\begin{aligned}
\Pr(D \mid X) &= \prod_{i,m,t} \Pr(c^t_{i,m} \mid b^t_{i,m}, x^t_{i,m}, \alpha_m, \beta_m) \\
&\propto \prod_{i,m,t} \left(\frac{\beta_m}{\beta_m + b^t_{i,m}}\right)^{x^t_{i,m}\alpha_m} \frac{\Gamma(x^t_{i,m}\alpha_m + c^t_{i,m})}{\Gamma(x^t_{i,m}\alpha_m)}
\end{aligned}$$

In this expression, terms not dependent on the $x^t_{i,m}$ have been removed, since these are constant for all hypotheses under consideration. For the null hypothesis H_0, we have $x^t_{i,m} = 1$ everywhere:

$$\Pr(D \mid H_0) \propto \prod_{i,m,t} \left(\frac{\beta_m}{\beta_m + b^t_{i,m}}\right)^{\alpha_m} \frac{\Gamma(\alpha_m + c^t_{i,m})}{\Gamma(\alpha_m)}$$

For the alternative hypothesis $H_1(S, E_k)$, we must marginalize over the values of $x^t_{i,m}$:

$$\Pr(D \mid H_1(S, E_k)) = \sum_X \Pr(D \mid X)\Pr(X \mid H_1(S, E_k))$$

The distribution of the $x^t_{i,m}$ is conditional on the event type E_k and affected region S, as discussed below. We note that, even though the counts are assumed to be conditionally

independent given the baselines, effects, and parameter priors, we are still able to model correlations between counts both under the null and alternative hypotheses. Since the values of the baselines $b^t_{i,m}$ are learned by time series analysis of historical data, accounting for day-of-week and seasonal variation (and any other relevant covariates), these values introduce correlations between counts under the null hypothesis. The relative risks $q^t_{i,m}$ are assumed to be conditionally independent under the null hypothesis, given the values of α_m and β_m. Under the alternative hypothesis $H_1(S, E_k)$, the relative risks in region S are correlated by the dependence on the effects $x^t_{i,m}$, which are dependent not only on the event type E_k but also the magnitude of the event (as discussed below). While the conditional independence assumptions made by our model may not be valid in all cases (e.g. if relative risks $q^t_{i,m}$ are correlated under the null hypothesis), we believe that this model is substantially more realistic than the typical multivariate scan statistic assumption of independent streams.

4.2 Computing the parameter priors

We can obtain estimated values of the parameter priors α_m and β_m for each data stream using a "parametric empirical Bayes" procedure, matching the first two moments (mean and variance) of the Gamma-Poisson model to their observed values from historical data. For this computation, we assume that no events have taken place in the historical data; any counts and baselines corresponding to known events should be removed. For each data stream D_m, we can compute the values of α_m and β_m using only the historical counts $c^t_{i,m}$ and computed baselines $b^t_{i,m}$ for that stream. From above, we know that the marginal distribution of each count $c^t_{i,m}$ under the null is negative binomial, with mean $\frac{\alpha_m}{\beta_m} b^t_{i,m}$ and variance $\frac{\alpha_m}{\beta_m^2}(b^t_{i,m})^2 + \frac{\alpha_m}{\beta_m} b^t_{i,m}$. We can then consider the ratios $r^t_{i,m} = \frac{c^t_{i,m}}{b^t_{i,m}}$, each of which will be distributed with mean $\frac{\alpha_m}{\beta_m}$ and variance $\frac{\alpha_m}{\beta_m^2} + \frac{1}{b^t_{i,m}} \frac{\alpha_m}{\beta_m}$. This means that the expected value of the sample mean of the $r^t_{i,m}$, $\bar{r}_m$, is equal to $\frac{\alpha_m}{\beta_m}$. Thus we can use the observed value of $\bar{r}_m$ as an estimate for $\frac{\alpha_m}{\beta_m}$. Similarly, the expected value of the sample variance of the $r^t_{i,m}$, $s^2_{r_m}$, is equal to $\frac{\alpha_m}{\beta_m^2} + \mathrm{E}[\frac{1}{b^t_{i,m}}]\frac{\alpha_m}{\beta_m}$, where $\mathrm{E}[\frac{1}{b^t_{i,m}}]$ is the sample mean of the values $\frac{1}{b^t_{i,m}}$. Thus we solve for α_m and β_m, in terms of the sample means $\bar{r}_m$ and $\mathrm{E}[\frac{1}{b^t_{i,m}}]$ and sample variance $s^2_{r_m}$:

$$\begin{aligned} \bar{r}_m &= \frac{\alpha_m}{\beta_m} & \qquad \alpha_m &= \frac{\bar{r}_m^2}{s^2_{r_m} - \bar{r}_m \mathrm{E}[\frac{1}{b^t_{i,m}}]} \\ & & \Longrightarrow & \\ s^2_{r_m} &= \frac{\alpha_m}{\beta_m^2} + \mathrm{E}\big[\tfrac{1}{b^t_{i,m}}\big]\frac{\alpha_m}{\beta_m} & \qquad \beta_m &= \frac{\bar{r}_m}{s^2_{r_m} - \bar{r}_m \mathrm{E}[\frac{1}{b^t_{i,m}}]} \end{aligned}$$

4.3 Event models

As noted above, each event type E_k is assumed to have a different joint probability distribution over $x^t_{i,m}$; we call this the "event model" corresponding to event type E_k. Here we assume a simplified event model, in which the effect on each data stream D_m is some constant x_m. In this case, we have $x^t_{i,m} = x_m$ for affected locations and time steps $(s_i \in S, t < W(S))$ and $x^t_{i,m} = 1$ otherwise. We further simplify the event model by parameterizing this specification in terms of the average effects $x_{km,avg}$ of each event type E_k on each data stream D_m and the event magnitude θ. For a given event type E_k with average effects $x_{km,avg}$ on each stream, and for a given value of θ, we set each x_m equal to $1 + \theta(x_{km,avg} - 1)$. For example, if a given event type E_k has average effects (1.5, 1, 1.2) on three data streams, this would mean that it increases counts for streams 1 and 3 by an average of 50% and

20% respectively in the affected region, and has no effect on counts for stream 2. Then an event of this type with magnitude $\theta = 1$ would have $x_1 = 1.5$, $x_2 = 1$, and $x_3 = 1.2$, while an event with $\theta = 2$ would have $x_1 = 2$, $x_2 = 1$, and $x_3 = 1.4$. We assume a fixed, discrete distribution for θ, mixing uniformly over $\theta \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1, \frac{3}{2}, 2, 3, 4\}$. The average effects of each event type on each stream, $x_{km,avg}$, can either be specified by a domain expert or learned from training data. One simple way to learn these average effects is by *maximum likelihood*: given multiple days of data, each with a labeled event of type E_k in some region S, each value $x_{km,avg}$ can be computed as the average ratio of total count $\sum_{i,t} c^t_{i,m}$ to total baseline $\sum_{i,t} b^t_{i,m}$ for stream m in regions affected by event type E_k. This method of maximum likelihood learning is used in our "MBSS-FIT" method described below.

5 Computational considerations

In practice, rather than explicitly computing the data likelihoods $\Pr(D \mid H)$ for each hypothesis, we compute the likelihood ratios $\frac{\Pr(D|H_1(S,E_k))}{\Pr(D|H_0)}$ for each alternative hypothesis $H_1(S, E_k)$. These likelihood ratios are faster to compute since they depend only on the counts inside region S, and (along with the prior probabilities of each hypothesis) are sufficient to compute the posterior probabilities. For the simplified event model described above, we have the following expression for the likelihood ratio:

$$\begin{aligned}
&\frac{\Pr(D \mid H_1(S, E_k))}{\Pr(D \mid H_0)} \\
&= \sum_{x_1 \ldots x_M} \Pr(x_1 \ldots x_M \mid E_k) \prod_{i,m,t \in S} \frac{\Pr(c^t_{i,m} \mid b^t_{i,m}, x_m \alpha_m, \beta_m)}{\Pr(c^t_{i,m} \mid b^t_{i,m}, \alpha_m, \beta_m)} \\
&= \sum_{x_1 \ldots x_M} \Pr(x_1 \ldots x_M \mid E_k) \prod_{i,m,t \in S} \left(\frac{\beta_m}{\beta_m + b^t_{i,m}}\right)^{(x_m - 1)\alpha_m} \frac{\Gamma(\alpha_m)\Gamma(x_m \alpha_m + c^t_{i,m})}{\Gamma(x_m \alpha_m)\Gamma(\alpha_m + c^t_{i,m})} \\
&= \sum_{x_1 \ldots x_M} \Pr(x_1 \ldots x_M \mid E_k) \prod_{i,t \in S} \prod_m (LR^t_{i,m} \mid x_m)
\end{aligned}$$

For a given vector of effects $(x_1 \ldots x_M)$, we can precompute the log-likelihood ratios $LLR^t_i = \sum_m \log(LR^t_{i,m} \mid x_m)$ for each spatial location s_i and each time step $t < W_{\max}$. Then to compute the log-likelihood ratio for a given spatial region S and vector of effects, we need only to sum the log-likelihood ratios (given $x_1 \ldots x_M$) for all locations s_i and time steps t in S. This formulation has the added benefit that the expensive likelihood ratio computations are only performed a number of times proportional to the number of locations, rather than the (much larger) number of regions.

Thus the multivariate Bayesian scan statistic can be computed in five steps. First, we load the counts $c^t_{i,m}$ for each spatial location s_i, data stream D_m, and time step t. This step requires time $O(IMT)$, where I, M, and T are the numbers of locations, streams, and time steps respectively. Second, we compute the baselines $b^t_{i,m}$ for each location, stream, and time step. A separate set of baseline computations must be performed for each of the I locations and M data streams. For a given location s_i and stream D_m, computation of the baselines $b^t_{i,m}$ (for all $t = 1 \ldots T$) requires $O(T)$ time, giving a total time complexity of $O(IMT)$ for this step. Third, we compute the parameter priors (α_m, β_m) for each of the M streams. Each such computation requires us to compute the mean and variance of the ratios $\frac{c^t_{i,m}}{b^t_{i,m}}$ for the I

locations and T time steps for that stream, giving a total time complexity of $O(IMT)$ for this step as well. Thus the first three steps have complexity proportional to the total size of the dataset, $O(IMT)$, but independent of the number of event models K and the number of search regions N_S.

The fourth step is to compute the log-likelihood ratios LLR_i^t for each location s_i and time step t, given each hypothesis E_k and corresponding vector of effects $(x_1 \dots x_M)$. This computation only needs to be performed for the most recent $W_{\max}$ time steps, where $W_{\max}$ is the maximum temporal window size, rather than for all T time steps. However, a separate computation must be done for each of the K event models and N_θ vectors of effects for each event model, and each computation requires us to compute the likelihood ratios $LR_{i,m}^t$ for the M data streams. This gives a total time complexity of $O(IMW_{\max}KN_\theta)$ for the fourth step. The fifth and final step is to compute the posterior probability for each hypothesis $H_1(S, E_k)$. Each likelihood $\Pr(D \mid H_1(S, E_k))$ can be computed by averaging over the N_θ possible parameter vectors $(x_1 \dots x_M)$ for that hypothesis, adding the corresponding log-likelihood ratios LLR_i^t for all $s_i \in S$ and all $t = 0 \dots W(S) - 1$. For each hypothesis $H_1(S, E_k)$ and parameter vector $(x_1 \dots x_M)$, we can compute the likelihood in time proportional to the number of locations in S. However, for many common region shapes (e.g. circles, rectangles) we can reduce this to an amortized $O(1)$ per region (Neill 2006). Once we have the likelihoods and priors for each hypothesis, the posteriors can be computed using Bayes' Theorem in $O(1)$ per hypothesis, or $O(N_S K)$ total time. This gives us a total time complexity of $O(N_S K N_\theta)$ for this step, and a total time complexity of $O(IMT + IMW_{\max}KN_\theta + N_S K N_\theta)$ for the entire algorithm. In practice, any of these steps can dominate the run time, since $T \gg W_{\max}$, $N_S \gg I$, and the fourth step has the most expensive computations (i.e. computing the log-likelihoods). To give an idea of the relative sizes of these quantities, our set of experiments described below used $I = 58$, $M \leq 3$, $T = 56$, $W_{\max} = 1$, $K \leq 7$, $N_\theta = 9$, and $N_S = 1292$.

6 Related work

The multivariate Bayesian scan statistic builds most directly on the univariate Bayesian scan statistic (Neill et al. 2006), extending the Bayesian event detection framework to multiple data streams and multiple event types. This extension allows us to achieve higher detection power by integrating information from the multiple streams, and also allows event characterization (by modeling and distinguishing between multiple types of event). The Bayesian scan statistic framework is a variant of the more traditional, hypothesis testing approach to spatial scan statistics, first developed by Kulldorff and Nagarwalla (1995) and Kulldorff (1997), and incorporated into a general cluster detection framework by Neill and Moore (2005). As we demonstrated in Neill et al. (2006), the Bayesian scan statistic approach has several advantages over the frequentist methods, including higher detection power, fast computation, easy interpretability of results, and ability to incorporate prior knowledge; all of these advantages also apply to our current work. While Kulldorff's original spatial scan statistic (Kulldorff 1997) did not take the time dimension into account, later work generalized this method to the "space-time scan statistic" by scanning over variable size temporal windows (Kulldorff et al. 1998; Kulldorff 2001). Recent extensions such as the expectation-based scan statistic (Neill et al. 2005b) and model-based scan statistic (Kleinman et al. 2005) also take the time dimension into account by using historical data to model the expected distribution of counts in each spatial location.

Many other variants of the spatial and space-time scan statistics have been proposed, differing in both the set of regions to be searched and the underlying statistical models. While Kulldorff's original method (Kulldorff 1997) assumed circular search regions,

other methods have searched over rectangles (Neill et al. 2005a), ellipses (Kulldorff et al. 2006), and various sets of irregularly shaped regions (Duczmal and Assuncao 2004; Patil and Taillie 2004; Tango and Takahashi 2005). In past work, we demonstrated that the "fast spatial scan" method can be used to speed up scan statistic calculations by 2–3 orders of magnitude when searching over rectangular regions (Neill and Moore 2004). Various statistical models have been proposed for the spatial scan, ranging from simple Poisson and Gaussian statistics (Neill et al. 2005b; Neill 2006) to robust and nonparametric models (Neill and Sabhnani 2007; Neill and Lingwall 2007). In Neill (2007a), we compared 25 different variants of the frequentist spatial scan statistic (Kulldorff 1997) on various disease surveillance tasks, and demonstrated that the expectation-based Poisson statistic (Neill et al. 2005b) outperformed Kulldorff's original statistic across a variety of datasets and outbreak sizes. Our current approach builds on the expectation-based Poisson model by incorporating it into a hierarchical Gamma-Poisson model and computing marginal likelihoods in a Bayesian framework.

Additionally, two multivariate extensions of the frequentist spatial scan have recently been proposed: Kulldorff's parametric scan (Kulldorff et al. 2007), which directly extends the original spatial scan statistic to multiple data streams by assuming that all data streams are independent, and the nonparametric scan (Neill and Lingwall 2007), which combines empirical p-values from multiple data streams without relying on an underlying parametric model. As we demonstrate in Neill and Lingwall (2007), the nonparametric scan can accurately characterize events by identifying which data streams have been affected, since it scans over subsets of the data streams as well as over space and time. However, neither of these two methods can differentiate between multiple types of event that might result in space-time clusters. We compare the detection power of the MBSS method to Kulldorff's multivariate scan in several outbreak detection scenarios below.

Several other multivariate surveillance methods have been proposed and applied to the disease surveillance domain, including multivariate extensions of traditional time series analysis methods (Burkom 2003; Burkom et al. 2005) and network-based methods (Reis et al. 2007) that detect anomalous ratios of counts between streams. These "purely temporal" detection methods do not take spatial information into account: they may be used to detect anomalous increases in the aggregate time series of the entire area being monitored, rather than detecting and pinpointing a spatial cluster of affected locations. Additionally, these methods cannot model and differentiate between multiple event types. Finally, the PANDA system (Cooper et al. 2004, 2007) uses Bayesian network models to differentiate between multiple outbreak types (e.g. the CDC Category A diseases), assuming an underlying entity-based model of Emergency Department visits. We have recently developed a multivariate model that incorporates spatial information into PANDA, using Emergency Department chief complaint data as evidence (Jiang et al. 2008).

7 Evaluation

We evaluated the event detection and characterization performance of the MBSS method, as compared to several previously proposed detection methods, on two sets of prospective disease surveillance tasks. The first set of experiments focused on outbreak detection, while the second set of experiments focused on outbreak characterization (discriminating between multiple types of outbreak). In all of our experiments, multiple streams of over-the-counter medication sales data (from Allegheny County, Pennsylvania) were monitored on a daily basis. The goal of this surveillance was to achieve timely detection and accurate characterization of emerging outbreaks of disease, while keeping the number of false positives low.

Table 1 Minimum, maximum, mean, and standard deviation of daily counts for each of the three public health datasets

Dataset	Minimum	Maximum	Mean	Standard deviation
CC	338	5474	2428.46	923.47
AF	83	2875	1321.70	279.88
TH	4	99	41.44	17.96

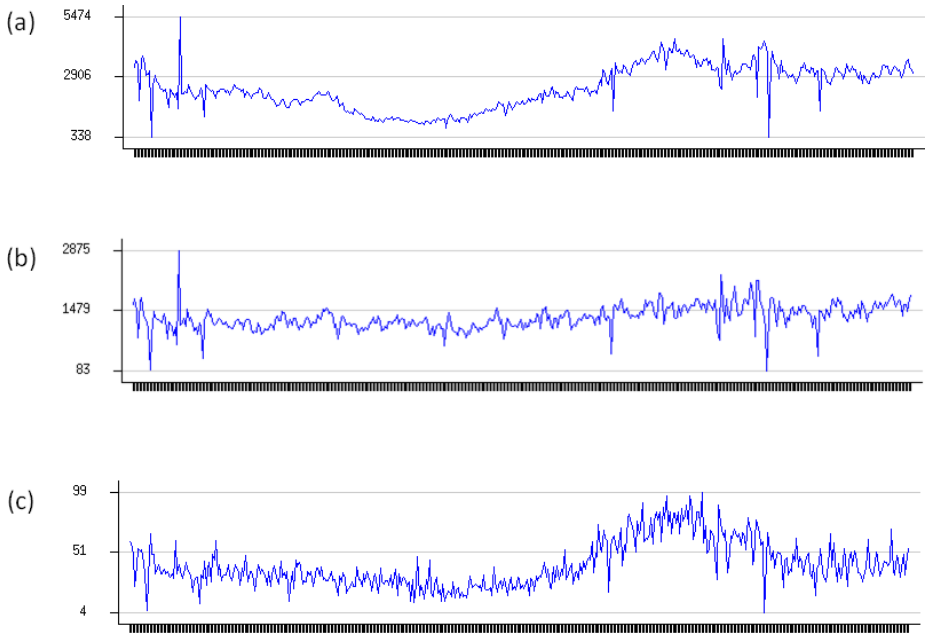

Fig. 2 Time series of daily counts for three public health datasets from October 1, 2004 to January 4, 2006. (**a**)–(**c**) represent the CC, AF, and TH datasets respectively

We used a semi-synthetic testing framework, injecting simulated disease outbreaks into real-world datasets, and analyzing the proportion of outbreaks detected and the time to detection. We obtained daily counts for three categories of OTC sales (cough/cold, anti-fever, and thermometers) for 58 Allegheny County zip codes, from October 1, 2004 to January 4, 2006. We denote these three datasets by CC, AF, and TH respectively. The first 84 days of data were used for baseline calculations only, giving 377 days of data for testing. Information about each dataset's daily counts (minimum, maximum, mean, and standard deviation) is given in Table 1. From this table, it is evident that the CC and AF datasets have much larger average counts than the TH dataset, and that all three datasets are overdispersed. The datasets also demonstrate significant day-of-week and seasonal trends; the time series of daily counts for each dataset is shown in Fig. 2.

We considered a simple class of circular outbreaks with a linear increase in the expected number of cases over the duration of the outbreak. More precisely, our outbreak simulator takes four parameters: the outbreak duration T, the outbreak severity Δ_m for each stream D_m, and the minimum and maximum number of zip codes affected, $k_{\min}$ and $k_{\max}$. Then for each injected outbreak, the outbreak simulator chooses the start date of the outbreak

t_{start}, the number of zip codes affected k, and the center zip code s_{center} uniformly at random. The outbreak is assumed to affect zip code s_{center} and its $k-1$ nearest neighbors, as measured by distance between the zip code centroids. On each day t of the outbreak, $t = 1 \ldots T$, the outbreak simulator injects Poisson$(t w_{i,m} \Delta_m)$ cases into each stream of each affected zip code, where $w_{i,m}$ is the "weight" of the zip code for that stream, $w_{i,m} = \frac{\sum_t c_{i,m}^t}{\sum_i \sum_t c_{i,m}^t}$. We used a constant value of $T = 7$ for all outbreaks, and thus all outbreaks were assumed to be one week in duration; outbreak size was allowed to vary between $k_{\min} = 5$ and $k_{\max} = 35$ affected zip codes.

We note that simulation of outbreaks is an active area of ongoing research in biosurveillance. The creation of realistic outbreak scenarios is important because of the difficulty of obtaining sufficient labeled data from real outbreaks, but is also very challenging. State-of-the-art outbreak simulations such as those of Buckeridge et al. (2004) and Wallstrom et al. (2005) combine disease trends observed from past outbreaks with information about the current background data into which the outbreak is being injected, as well as allowing the user to adjust parameters such as outbreak duration and severity. While the simple linear outbreak model that we use here is not a realistic model of the temporal progression of an outbreak, it is sufficient for testing spatial detection methods, with the idea that we gradually ramp up the amount of increase until the outbreak is detected.

For all of the methods under consideration, we scanned over the same predetermined set of search regions $\mathcal{S}$. This set of regions was formed by mapping the Allegheny County zip codes to a 16×16 grid, and searching over all rectangular regions on the grid with size up to 8×8. Note that this set of search regions is substantially different than the set of inject regions used by our outbreak simulator: this is typical of real-world outbreak detection scenarios, where the size and shape of potential outbreaks is not known in advance. In general, searching over rectangles has the advantages of computational efficiency as well as high power to detect both compact and elongated clusters (Neill and Moore 2004; Neill et al. 2005a). Of course, if we had assumed some prior knowledge of the set of inject regions, this knowledge could be used to refine our search accordingly. We used $W_{\max} = 1$, and thus only searched over regions of 1-day duration; a larger value of the maximum temporal window size would be useful for more slowly growing outbreaks. Choosing a different set of search regions would most likely affect the detection power of our methods, however, we expect that the relative performance of different methods will remain approximately the same. The question of choosing an optimal set of search regions (in order to maintain high detection power over a wide range of outbreak shapes and sizes) is orthogonal to our question of choosing the correct statistical method, and has been investigated in detail by Duczmal and Assuncao (2004), Patil and Taillie (2004), Tango and Takahashi (2005), and many others.

7.1 Evaluation of event detection

In our first set of experiments, we ran 26 outbreak simulations in order to compare the detection power of different methods across a variety of outbreak detection scenarios. Each simulation was characterized by a different set of parameters (Δ_{CC}, Δ_{AF}, and Δ_{TH}) representing the effects on each data stream, as described above. The 26 simulations included combinations of $\Delta_{CC} \in \{100, 50, 0\}$, $\Delta_{AF} \in \{60, 30, 0\}$, and $\Delta_{TH} \in \{4, 2, 0\}$. The "base value" for each data stream ($\Delta_{CC} = 100$, $\Delta_{AF} = 60$, $\Delta_{TH} = 4$) was chosen proportional to its total count; hence we inject far fewer cases into the TH stream as compared to CC or AF. We then allowed each Δ_m to be equal to the base value, half the base value, or zero (in which case the outbreak has no effect on that data stream), but excluded the case where no cases are

injected into any of the streams. For each of the 26 simulations, we averaged results over 250 different, randomly generated outbreaks, giving a total of 6,500 outbreaks for evaluation.

We compared a total of seven methods for this evaluation: three variants of the MBSS method (MBSS-EQ, MBSS-FIT, MBSS-7M), three univariate Bayesian detectors (BSS-CC, BSS-AF, BSS-TH), and Kulldorff's multivariate spatial scan (Kulldorff et al. 2007) (KULL). All methods assumed the same set of search regions (as discussed above), and the Bayesian methods assumed a uniform prior over search regions. MBSS-EQ used a single event model $x_{km,avg} = (1.5, 1.5, 1.5)$, which assumed equal average effects on the three data streams. MBSS-7M used 7 event models, each assuming that a different subset of streams has been affected: {CC-AF-TH, CC-AF, CC-TH, AF-TH, CC, AF, TH}. Each model assumes equal relative effects on the affected data streams and no effects on the other streams; for example, the CC-AF-TH model has average effects $x_{km,avg} = (1.5, 1.5, 1.5)$, and the AF-TH model has average effects $x_{km,avg} = (1, 1.5, 1.5)$ on the CC, AF, and TH streams respectively.

As opposed to the more general models of MBSS-EQ and MBSS-7M, the MBSS-FIT method assumes a single, specific event model which is fitted to the event type under consideration. We learned the vector of "average effects" on each stream by maximum likelihood estimation, as described above. Learning was performed incrementally, using each labeled outbreak (S, E_k) to update the average effects of the given outbreak type E_k on each stream. The vector of average effects converged quickly (within 10–15 training examples), and thus we focus on the performance of the fitted model rather than the dynamics of the learning process. BSS-CC, BSS-AF, and BSS-TH are univariate versions of the MBSS method that each only monitor one data stream, as in our previous univariate Bayesian scan statistics work (Neill et al. 2006). The average effect on that data stream was assumed to be $x_{km,avg} = 1.5$. Finally, KULL is the recently proposed multivariate version of Kulldorff's spatial scan statistic, as described in Kulldorff et al. (2007). This method extends the original spatial scan statistic by assuming that streams are independent, giving a total region score equal to the sum of the region's log-likelihood ratio scores for each individual data stream.

We computed each method's proportion of outbreaks detected and average number of days to detect, on each of the 26 simulations, as a function of the allowable false positive rate. To do this, we first computed the total posterior probability of an outbreak, $\Pr(H_1 \mid D) = \sum_{S,E_k} \Pr(H_1(S, E_k) \mid D)$, for each day of the original dataset with no outbreaks injected (as noted above, the first three months of data are excluded, since these are used to calculate baselines for our methods). Then for each injected outbreak, we computed the total posterior probability of an outbreak for each day of the outbreak, and determined what proportion of the days for the original dataset have higher outbreak probabilities. Assuming that the original dataset contains no outbreaks, this is the proportion of false positives that we would have to accept in order to have detected the outbreak on day t. For a fixed false positive rate r, the "days to detect" for a given simulated outbreak is computed as the first outbreak day ($t = 1 \ldots 7$) with proportion of false positives less than r. If no day of the outbreak has a proportion of false positives less than r, the method has failed to detect that outbreak. As a useful summary measure, we can consider the average "adjusted days to detect" for each method, at a fixed false positive rate of 1/month. For this measure, any missed outbreaks are penalized by the entire duration of the outbreak, and thus counted as requiring $t = 14$ days to detect. The detection performance (average adjusted days to detect) for each method, for each of the 26 simulations, is presented in Table 2. The mean performance (adjusted days to detect) was also computed for each method, giving an aggregate measure of each method's ability to detect a variety of different outbreak types.

As expected, the three univariate detectors achieved timely detection when an outbreak type had high impact on the monitored data stream, and performed poorly otherwise; thus

Table 2 Average adjusted days to detection at 1 false positive per month, for each of the 26 simulations. Methods in bold are not significantly different (at $\alpha = .05$) from the best-performing method

$(\Delta_{CC}, \Delta_{AF}, \Delta_{TH})$	MBSS-EQ	MBSS-7M	MBSS-FIT	BSS-CC	BSS-AF	BSS-TH	KULL
(100, 60, 4)	**2.512**	**2.672**	**2.492**	2.836	3.228	3.648	**2.660**
(100, 60, 2)	**2.640**	**2.756**	**2.604**	**2.880**	3.272	6.340	**2.680**
(100, 60, 0)	**2.768**	**2.860**	**2.720**	3.064	3.364	12.228	**2.572**
(100, 30, 4)	3.168	3.204	**2.832**	**2.780**	6.016	3.496	**2.796**
(100, 30, 2)	3.448	3.512	**3.160**	**2.980**	6.420	6.568	**2.916**
(100, 30, 0)	3.548	3.584	3.456	**3.224**	6.648	11.684	**3.136**
(100, 0, 4)	4.608	3.744	**2.632**	**2.896**	12.088	3.604	3.224
(100, 0, 2)	5.416	4.296	**3.284**	**3.084**	12.052	6.708	**3.340**
(100, 0, 0)	4.872	3.840	3.080	**2.728**	12.056	12.000	3.088
(50, 60, 4)	**3.112**	3.216	**2.836**	5.328	3.140	3.464	3.188
(50, 60, 2)	3.344	3.352	**2.980**	5.692	**3.088**	6.204	3.280
(50, 60, 0)	3.552	3.528	**3.344**	5.648	**3.060**	12.104	3.396
(50, 30, 4)	4.428	4.732	3.944	5.244	6.572	**3.528**	4.660
(50, 30, 2)	**5.184**	5.384	**4.976**	5.884	7.112	6.728	5.276
(50, 30, 0)	5.392	5.372	**5.208**	5.812	6.940	11.868	**5.004**
(50, 0, 4)	7.392	6.612	**3.720**	5.728	11.468	**3.744**	5.792
(50, 0, 2)	8.228	7.212	**5.096**	5.656	12.008	6.732	6.228
(50, 0, 0)	9.256	8.204	6.472	**5.972**	11.648	11.740	6.804
(0, 60, 4)	5.288	3.880	**2.536**	12.264	3.196	3.468	4.036
(0, 60, 2)	5.956	4.272	**2.700**	11.772	3.032	6.056	4.004
(0, 60, 0)	6.316	4.364	**2.860**	12.664	3.248	12.228	4.764
(0, 30, 4)	7.772	6.476	**3.212**	11.700	6.172	3.604	7.212
(0, 30, 2)	9.120	8.296	**4.648**	12.252	7.040	6.600	8.744
(0, 30, 0)	10.068	8.288	**5.612**	12.120	6.596	12.044	8.876
(0, 0, 4)	11.384	11.112	**3.500**	12.104	12.120	**3.764**	12.148
(0, 0, 2)	11.584	11.696	7.184	11.992	11.648	**6.584**	12.068
Mean	5.783	5.249	**3.734**	6.704	7.047	7.182	5.073

we would expect that higher average performance could be achieved by integrating information from multiple data streams. The results of Table 2 confirm our expectations: all four of the multivariate detectors were able to detect a full day faster (on average) than any of the univariate detectors, though the univariate detectors performed significantly better than the MBSS-EQ, MBSS-7M, and KULL methods when the outbreak only affected the corresponding single data stream. This effect was particularly evident for the thermometers (TH) data stream: since TH had much smaller average counts than the other streams, signals in this stream were overwhelmed by noise from the other streams, and the MBSS-EQ, MBSS-7M, and KULL methods had low detection power for these outbreaks.

On the other hand, the MBSS-FIT method was able to achieve timely detection of all outbreak types by learning which data streams were most affected, and fitting a model of the average effects on each stream. As a result, MBSS-FIT was able to detect outbreaks an average of 1.3 days faster than any of the other methods. MBSS-FIT achieved performance comparable to the best univariate detector (4.78 vs. 4.82 adjusted days to detect) for

outbreaks affecting only a single data stream, and significantly better performance than the best univariate detector (3.42 vs. 3.70 adjusted days to detect) for outbreaks affecting multiple data streams. The high performance of MBSS-FIT demonstrates one major advantage of MBSS over other detection methods: much higher detection power can be achieved by incorporating information about an event's effects on the different data streams. This knowledge can either be pre-specified by a domain expert, or learned from labeled training examples, creating a specific detector with high power to detect the given event type.

However, the performance of the MBSS-EQ method demonstrates the dangers of an incorrectly specified event model: MBSS-EQ achieved high detection performance for outbreak types corresponding to its model (i.e. equal effects on the three data streams) but performed substantially worse than the two more general multivariate detectors (MBSS-7M and KULL) when its models were incorrect. As a result, both MBSS-7M and KULL were able to detect outbreaks over half a day faster (on average) than MBSS-EQ. These results demonstrate that a single, specific event model should only be used if we have prior knowledge of the event's effects. If we have no prior knowledge, and want to maintain high detection power across a variety of event types, a more general detector such as MBSS-7M or KULL should be used. The MBSS-7M and KULL methods achieved comparable average performance across outbreak types, demonstrating that MBSS can also be used for general event detection by including multiple event models. The KULL method tended to place more weight on the cough/cold data stream, outperforming MBSS-7M (2.93 vs. 3.39 adjusted days to detect) when this stream was highly affected by an outbreak and performing worse (7.73 vs. 7.30 adjusted days to detect) when cough/cold sales were unaffected, but the results were otherwise very similar for these two methods.

7.2 Evaluation of event characterization

In our second set of experiments, we examined the ability of the MBSS method to characterize events in the disease surveillance domain, by differentiating between multiple types of disease outbreak. For each of these experiments, we injected several types of outbreak into the cough/cold and anti-fever sales for Allegheny County; each outbreak type had different effects on these two streams. In our first experiment, we injected two outbreak types: one type with larger effects on the CC stream ($\Delta_{CC} = 100$, $\Delta_{AF} = 30$) and one type with larger effects on the AF stream ($\Delta_{CC} = 50$, $\Delta_{AF} = 60$). In our second experiment, we injected three outbreak types: one type only affecting the CC stream ($\Delta_{CC} = 100$, $\Delta_{AF} = 0$), one type only affecting the AF stream ($\Delta_{CC} = 0$, $\Delta_{AF} = 60$), and one type affecting both streams ($\Delta_{CC} = 50$, $\Delta_{AF} = 30$). In each experiment, we used a total of 500 outbreaks for evaluation, alternating examples of each outbreak type. We used the MBSS-FIT method to learn a different event model for each outbreak type: two models were learned for the first experiment, and three models were learned for the second experiment. As in our first set of experiments, learning was performed incrementally from each outbreak (labeled with the appropriate outbreak type), and the vector of average effects for each event model converged quickly (within 10–15 examples of that type) to a fixed value.

To evaluate the event characterization ability of the MBSS-FIT method, we recorded the posterior probability of each outbreak type, $\Pr(E_k \mid D) = \sum_S \Pr(H_1(S, E_k) \mid D)$, and the posterior probability of the null hypothesis, $\Pr(H_0 \mid D)$, on each outbreak day. We then computed the average posterior probabilities of the correct outbreak type, the incorrect outbreak type(s), and the null hypothesis as a function of the number of days since the start of the outbreak. Figure 3 shows the results for the first experiment, using two event models. From this figure, it is evident that MBSS is able to accurately characterize outbreaks by

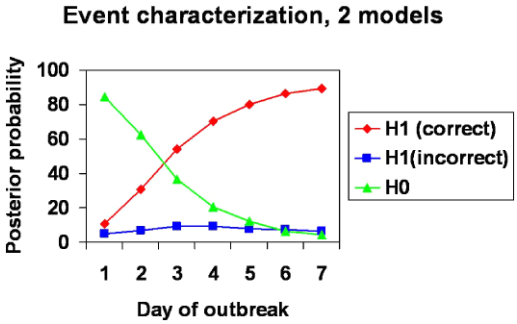

Fig. 3 Event characterization, 2 models. Average posterior probability (percent) of correct outbreak type, incorrect outbreak type, and null hypothesis on each outbreak day

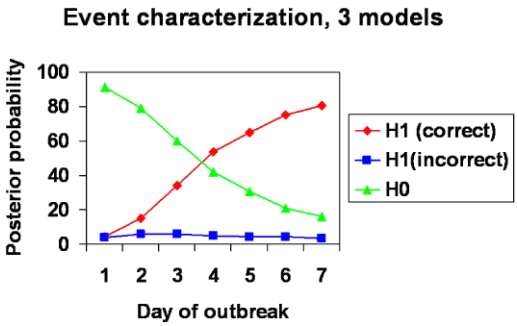

Fig. 4 Event characterization, 3 models. Average posterior probability (percent) of correct outbreak type, incorrect outbreak type, and null hypothesis on each outbreak day

the second outbreak day: the average posterior probability of the correct outbreak type was 31% on day 2, while the posterior probability of the incorrect outbreak type was only 7%. By the midpoint of the outbreak (day 4), the average posterior probability of the correct model increased to 70%, while the posterior probability of the incorrect model remained approximately constant. We see similar results for the second experiment, using three event models, in Fig. 4. By the third outbreak day, the average posterior probability of the correct outbreak type was 34%, while the posterior probability of the two incorrect outbreak types was only 6%. By the fifth outbreak day, the average posterior probability of the correct outbreak type increased to 65%, while the posterior probability of the incorrect outbreak types decreased slightly to 4%. Thus the MBSS method is able to accurately characterize and distinguish between different types of outbreak, using a learned model of the effects of each outbreak type.

We also compared the detection power of the MBSS-FIT method when learning a separate event model for each outbreak type to its detection power when learning only a single event model. As in our first set of experiments, we compared the average adjusted days to detect for each method, at a fixed false positive rate of 1/month. In the first experiment, learning two separate event models only achieved a slight (but not statistically significant) increase in performance as compared to learning a single model, reducing the average adjusted days to detect from 3.312 to 3.232. This result demonstrates that when the models for different event types are very similar, a single model is sufficient for good detection performance; however, having multiple models enables MBSS to accurately differentiate between the event types. In the second experiment, learning three separate models for the three outbreak types led to a significant improvement in performance, reducing the average adjusted days to detect from 5.404 to 4.272. This result demonstrates that when the different event types have very different effects on the data streams, multiple models are necessary to

achieve high detection power, as well as enabling accurate characterization of which type of event has occurred.

8 Discussion

Our evaluation results demonstrate three major advantages of the MBSS method for event detection and surveillance. As in two other recently proposed multivariate detection methods (Kulldorff et al. 2007; Neill and Lingwall 2007), MBSS achieves high detection power by combining information from multiple data streams, spatial locations, and time steps. This integration of information is essential for detecting emerging patterns (e.g. the early stages of an emerging outbreak of disease) that would not be visible from monitoring only a single data stream, spatial location, or time step. As demonstrated by our first set of experiments, MBSS can be used as a *general* event detector (by including multiple event models and using uninformative priors), and can achieve detection power comparable to the current state of the art (Kulldorff et al. 2007) across a wide range of event types.

However, MBSS improves substantially on the current state of the art in two other respects. First, we can incorporate informative priors into the MBSS models, and thus use MBSS as a *specific* event detector with much higher detection power for the specified event types. Our first set of experiments demonstrated that the effects of each event type on the multiple data streams can be learned from a small number of labeled training examples, and that the fitted models gained a large improvement (average of 1.3 days faster detection) as compared to the general multivariate detectors. In the disease surveillance domain, the event models for common outbreak types (such as seasonal influenza and rotavirus) or general outbreak classes (e.g. influenza-like illness) can be learned from real outbreaks labeled by domain experts, while models of rare events (e.g. terrorist bio-attacks) would need to be pre-specified (or learned from sufficiently realistic simulations) due to the sparsity of real-world examples. Other aspects of the MBSS models, including the prevalence, size, shape, and duration of events, can also be specified or learned from data, and these priors can also be used to improve detection performance.

Our second set of experiments demonstrates perhaps the most important advantage of MBSS: the ability to *characterize* events by specifying models for multiple event types and computing the probability that each type of event has occurred. As noted above, event characterization is necessary for several reasons: different types of event may be more or less relevant to the user (some events may require urgent responses, while some can safely be ignored), and different event types may also require the user to take different courses of action. By characterizing events as well as detecting them, the MBSS system can not only alert the user when relevant events are taking place (avoiding false positives due to irrelevant events), but also provide the user with a complete situational awareness including the type of event and which subset of the data (spatial region and time duration) has been affected.

Thus MBSS can detect faster and more accurately by integrating multiple data streams, and can model and differentiate between multiple event types. We now briefly consider several other advantages of using our Bayesian event detection framework, as compared to the standard frequentist hypothesis testing approach originated by Kulldorff's spatial scan statistic (Kulldorff 1997). First, unlike the frequentist approach, randomization testing is not necessary in the Bayesian framework. Since 999 or more Monte Carlo replications must typically be performed to obtain accurate p-values for the frequentist approach, we can obtain a $1000\times$ speedup by avoiding the need for randomization. Computation is fast in the Bayesian framework for several reasons: the marginal likelihoods can be computed in

closed form (thanks to the use of conjugate priors, with empirical Bayes estimates of the hyperpriors), and the expensive log-likelihood computations need only be done a number of times proportional to the number of locations, not the number of regions. As a result, computation of the multivariate Bayesian scan statistic (for a single event model) can be performed in 0.84 seconds per day of data on our systems, approximately the same run time as Kulldorff's multivariate scan statistic without randomization testing (0.79 seconds per day of data). Using multiple event models increases the time needed to compute likelihoods, proportional to the number of models, but does not affect the fixed costs of loading the data and computing baselines. For our first set of experiments, using seven event models instead of one increased the total run time only 15%, from 0.84 to 0.97 seconds per day of data.

A second advantage of not requiring randomization testing is that *calibration* of the Bayesian statistic is easier than calibration of the frequentist statistic. As we showed in Neill (2007a), the p-values reported by frequentist scan statistics tend to be oversensitive, in that the proportion of false positives reported at level α is much higher than α, e.g. 20–40% false positive rate at $\alpha = .05$ on OTC data. Using 999 Monte Carlo replications, many regions will have the smallest possible p-value ($p = 0.001$), making it difficult to distinguish between these regions. In our Bayesian approach, on the other hand, we can simply choose a threshold for the total posterior probability of relevant event types, and notify the user whenever the probability exceeds this threshold. The total number of alerts produced by the MBSS method tends to be very reasonable: for example, in our first set of disease surveillance experiments, the MBSS-EQ method found 15 days with posterior outbreak probabilities over 50%, i.e. slightly more than 1 alert per month.

Finally, the results produced by the MBSS method are easy to interpret, visualize, and use for decision-making. MBSS outputs the total posterior probability of each event type, $\Pr(E_k \mid D) = \sum_S \Pr(H_1(S, E_k) \mid D)$, as well as the posterior probability that no events have occurred, $\Pr(H_0 \mid D)$. These probabilities enable the user to decide whether to respond to the detected events, based on the costs of false positives and false negatives for each event type. Additionally, MBSS gives information about the space-time region affected by the event, distributing the total event probability $\Pr(E_k \mid D)$ over possible regions of effect S. One useful way to visualize these probabilities is to compute the total probability that each spatial location s_i has been affected by a given event type on a given day: this posterior probability $\Pr(H_1(s_i, E_k) \mid D)$ can be obtained by summing the probabilities $\Pr(H_1(S, E_k) \mid D)$ for all space-time regions S containing s_i. We can then display separate probability maps for each event type E_k for each day of data. For example, Fig. 5 shows the probability maps created by monitoring the OTC cough/cold, anti-fever, and thermometer sales in Allegheny County during a (simulated) outbreak from July 4–10, 2005, assuming a single MBSS event model with equal effects on the three data streams. Days 1, 3, 5, and 7 of the outbreak are shown here; the outbreak is not visible on Day 1, but becomes increasingly apparent (as represented by darker shading of the affected zip codes) over the course of the outbreak. Additionally, the variation in shading for Day 3 reveals uncertainty about the precise spatial extent of the outbreak during its early stages; by Day 5, MBSS is able to much more precisely identify which zip codes have been affected.

Another possibility for visualizing the outputs of MBSS is to combine the posterior probabilities of multiple event types into a single map. One way to do this is to use *color*: if the number of event types is at most three, we can encode the event posteriors for a given spatial location into the red, green, and blue components of its RGB-coded color. Then the total brightness of each area represents its total event probability, and whether the color is predominantly red, green, or blue indicates which event type is most likely. Many other visualization techniques are possible, and we are currently evaluating the utility of different approaches to visualization in the disease surveillance domain.

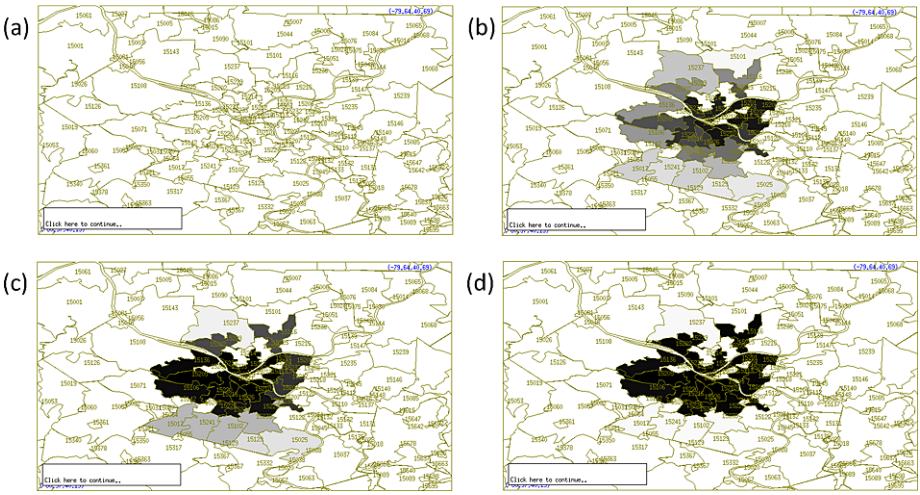

Fig. 5 Map of posterior outbreak probabilities computed by MBSS for a simulated outbreak injected into three OTC data streams from July 4–10, 2005. Days 1, 3, 5, and 7 of the outbreak are shown here. *Darker shading* indicates a higher probability that the given zip code has been affected

We are also currently working to extend the MBSS framework in several other ways, including the incorporation of incremental model learning from labeled data and active learning from user feedback, as well as extending the underlying statistical models to dynamically changing patterns and more general multivariate datasets. In the disease surveillance domain, we are currently developing models of several outbreak types (e.g. influenza, anthrax) in order to better detect and distinguish between these outbreaks. We are also developing models of other (non-outbreak) causes of a detected cluster, such as tourism and promotional sales of OTC medications. These models will allow us to discriminate between detected clusters that are due to outbreaks and those due to other irrelevant causes, reducing the number of false positives and increasing the system's power to detect true outbreaks.

Acknowledgements This work was partially supported by NSF grant IIS-0325581 and CDC grant 8 R01 HK000020-02. The authors wish to thank Andrew Moore, Jeff Schneider, Jeff Lingwall, Xia Jiang, and Maxim Makatchev for comments on earlier versions of this work.

References

Buckeridge, D. L., Burkom, H. S., Moore, A. W., Pavlin, J. A., Cutchis, P. N., & Hogan, W. R. (2004). Evaluation of syndromic surveillance systems: development of an epidemic simulation model. *Morbidity and Mortality Weekly Report*, *53*(Supplement on Syndromic Surveillance), 137–143.

Burkom, H. S. (2003). Biosurveillance applying scan statistics with multiple, disparate data sources. *Journal of Urban Health*, *80*(2 Suppl. 1), i57–i65.

Burkom, H. S., Murphy, S. P., Coberly, J., & Hurt-Mullen, K. (2005). Public health monitoring tools for multiple data streams. *Morbidity and Mortality Weekly Report*, *54*(Supplement on Syndromic Surveillance), 55–62.

Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, *43*, 671–681.

Cooper, G. F., Dash, D. H., Levander, J. D., Wong, W.-K., Hogan, W. R., & Wagner, M. M. (2004). Bayesian biosurveillance of disease outbreaks. In *Proc. conference on uncertainty in artificial intelligence*.

Cooper, G. F., Dowling, J. N., Levander, J. D., & Sutovsky, P. (2007). A Bayesian algorithm for detecting CDC Category A outbreak diseases from emergency department chief complaints. *Advances in Disease Surveillance*, *2*, 45.

Duczmal, L., & Assuncao, R. (2004). A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, *45*, 269–286.

Jiang, X., Neill, D. B., & Cooper, G. F. (2008). *A Bayesian network model for spatial event surveillance*. (Tech. rep.). University of Pittsburgh, Department of Biomedical Informatics.

Kleinman, K., Abrams, A., Kulldorff, M., & Platt, R. (2005). A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection*, *133*(3), 409–419.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, *26*(6), 1481–1496.

Kulldorff, M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A*, *164*, 61–72.

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, *14*, 799–810.

Kulldorff, M., Athas, W., Feuer, E., Miller, B., & Key, C. (1998). Evaluating cluster alarms: a space-time scan statistic and cluster alarms in Los Alamos. *American Journal of Public Health*, *88*, 1377–1380.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., & Mostashari, F. (2005). A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine*, *2*(3), e59.

Kulldorff, M., Huang, L., Pickle, L., & Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, *25*, 3929–3943.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W. K., Kleinman, K., & Platt, R. (2007). Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, *26*, 1824–1833.

Mollié, A. (1999). Bayesian and empirical Bayes approaches to disease mapping. In A. B. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, & R. Bertollini (Eds.) *Disease mapping and risk assessment for public health*.

Neill, D. B. (2006). *Detection of spatial and spatio-temporal clusters* (Tech. rep. CMU-CS-06-142). Ph.D. thesis, Carnegie Mellon University, Department of Computer Science.

Neill, D. B. (2007a). An empirical comparison of spatial scan statistics for outbreak detection. *Advances in Disease Surveillance*, *4*, 259.

Neill, D. B. (2007b). Incorporating learning into disease surveillance systems. *Advances in Disease Surveillance*, *4*, 107.

Neill, D. B., & Lingwall, J. (2007). A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance*, *4*, 106.

Neill, D. B., & Moore, A. W. (2004). Rapid detection of significant spatial clusters. In *Proc. 10th ACM SIGKDD conf. on knowledge discovery and data mining* (pp. 256–265).

Neill, D. B., & Moore, A. W. (2005). Anomalous spatial cluster detection. In *Proc. KDD 2005 workshop on data mining methods for anomaly detection* (pp. 41–44).

Neill, D. B., & Sabhnani, M. R. (2007). A robust expectation-based spatial scan statistic. *Advances in Disease Surveillance*, *2*, 61.

Neill, D. B., Moore, A. W., & Sabhnani, M. R. (2005a). Detecting elongated disease clusters. *Morbidity and Mortality Weekly Report*, *54*(Supplement on Syndromic Surveillance), 197.

Neill, D. B., Moore, A. W., Sabhnani, M. R., & Daniel, K. (2005b). Detection of emerging space-time clusters. In *Proc. 11th ACM SIGKDD intl. conf. on knowledge discovery and data mining*.

Neill, D. B., Moore, A. W., & Cooper, G. F. (2006). A Bayesian spatial scan statistic. In *Advances in neural information processing systems 18* (pp. 1003–1010).

Patil, G. P., & Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Envir. Ecol. Stat.*, *11*, 183–197.

Reis, B. Y., Kohane, I. S., & Mandl, K. D. (2007). An epidemiological network model for disease outbreak detection. *PLoS Medicine*, *4*, 210.

Sabhnani, M. R., Neill, D. B., Moore, A. W., Tsui, F.-C., Wagner, M. M., & Espino, J. U. (2005). Detecting anomalous clusters in pharmacy retail data. In *Proc. KDD 2005 workshop on data mining methods for anomaly detection* (pp. 58–61).

Tango, T., & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, *4*, 11.

Wagner, M. M., Tsui, F.-C., Espino, J. U., Hogan, W., Hutman, J., Hirsch, J., Neill, D. B., Moore, A. W., Parks, G., Lewis, C., & Aller, R. (2004). A national retail data monitor for public health surveillance. *Morbidity and Mortality Weekly Report*, *53*(Supplement on Syndromic Surveillance), 40–42.

Wallstrom, G. L., Wagner, M. M., & Hogan, W. R. (2005). High-fidelity injection detectability experiments: a tool for evaluation of syndromic surveillance systems. *Morbidity and Mortality Weekly Report*, *54*(Supplement on Syndromic Surveillance), 85–91.