A Bayesian spatio-temporal method for disease outbreak detection

Xia Jiang, Gregory F Cooper

▸ Additional appendix materials are published online only. To view these files please visit the journal online (http://jamia.bmj.com).

Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence to
Dr Xia Jiang, 183M Parkvale Building, 200 Meyran Avenue, Pittsburgh, PA 15260, USA; xij6@pitt.edu

Received 4 May 2009
Accepted 27 April 2010

ABSTRACT

A system that monitors a region for a disease outbreak is called a *disease outbreak surveillance system*. A *spatial surveillance system* searches for patterns of disease outbreak in spatial subregions of the monitored region. A *temporal surveillance system* looks for emerging patterns of outbreak disease by analyzing how patterns have changed during recent periods of time. If a non-spatial, non-temporal system could be converted to a spatio-temporal one, the performance of the system might be improved in terms of early detection, accuracy, and reliability.

A Bayesian network framework is proposed for a class of space-time surveillance systems called BNST. The framework is applied to a non-spatial, non-temporal disease outbreak detection system called PC in order to create the spatio-temporal system called PCTS. Differences in the detection performance of PC and PCTS are examined. The results show that the spatio-temporal Bayesian approach performs well, relative to the non-spatial, non-temporal approach.

INTRODUCTION

Early detection of disease outbreaks remains an important research topic. Even modest improvement in disease outbreak detection could have significant impact on public health in terms of lives saved and reduced economic cost. The induced economic costs were estimated to be as high 26.2 billion dollars per 100 000 persons exposed[1] and 200 million dollars per hour[2] for a disease outbreak caused by a large-scale release of inhalational anthrax.

Disease outbreak surveillance, also called disease outbreak detection and biosurveillance, consists of monitoring a community in order to recognize the onset of a disease outbreak early. A system that does this is called a *disease outbreak surveillance system*. Such a system looks for patterns of data that are indicative of an outbreak.

If an outbreak is occurring in a small subregion of a county, for example, and the system only investigates a global change for the entire county, the outbreak may go undetected until it spreads to a larger subregion. A *spatial disease outbreak surveillance system* searches for patterns in spatial subregions of the monitored region. In this way, we may detect the changes that start in small subregions earlier than if we only monitored the entire region globally. Another purpose of a spatial surveillance system is to identify the subregion where the outbreak is occurring. If a system only monitors the entire region globally, we call it *non-spatial*. A *temporal event surveillance* system looks for emerging outbreaks by analyzing how patterns have changed during recent periods of time. If a system does not do this, we call it *non-temporal*.

Even though the number of new cases during an outbreak generally exhibits an increasing trend, the daily fluctuation in this number may be dramatic. A non-temporal system, especially one that only looks at the current day's data, might issue an alarm one day when the number of new cases exceeds the normal level, and then not issue an alarm the next day when the number of new cases drops back down. As an example, the outbreak detection system PANDA-CDCA (PC)[3] does not use a temporal model of disease outbreaks. Cooper *et al*[3] obtained results that they initially found surprising when evaluating the ability of PC to detect a laboratory validated outbreak of influenza in Allegheny County, Pennsylvania. Under a false alarm rate of zero, PC detected influenza approximately one day before the first positive viral cultures of influenza were taken. However, near the beginning of the outbreak, the posterior probability of an influenza outbreak fluctuated between very high and very low values. It is generally not the case that an outbreak is present today, gone the next day, and back the following day, and so on. Rather the daily number of new cases during an outbreak can fluctuate. Our model should reflect how an outbreak behaves, including expecting typical temporal patterns of disease outbreaks. Otherwise, a non-outbreak assessment that is issued during an outbreak could confuse the user of the system, thereby delaying the correct decision that an outbreak is occurring. A temporal system that expects an outbreak to result in an emerging increase in the number of outbreak cases over time, while allowing for daily variation, may be able to maintain stable detection and reduce false negative signals without compromising initial detection performance.

We introduce a method for converting an existing non-spatial, non-temporal Bayesian outbreak detection system to a *spatio-temporal* one. We apply the method to convert PC to the spatio-temporal system PANDA-CDCA-Temporal-Spatial (PCTS). We hypothesize that: (1) when an outbreak is occurring in a small subregion, PCTS will detect the outbreak earlier than PC; (2) once an outbreak is detected, PCTS will more consistently maintain a detection signal than will PC; (3) when an outbreak is occurring in a small subregion, PCTS can estimate that subregion accurately; and (4) PCTS will perform at least as well as SaTScan, which is an existing state-of-the-art spatial detection system. We show results of evaluating these hypotheses.

BACKGROUND

An outbreak detection system looks for patterns of data that are indicative of an outbreak.

Some detection systems do this by analyzing counts of an observable event. For example, the number of patients presenting to an emergency department (ED) with respiratory-related chief complaints could be used to detect an influenza outbreak. Non-spatial, non-temporal methods of this type consider counts from some recent period of time only, such as the previous day, and do not investigate spatial subregions. A simple control chart method for analyzing these daily counts first derives the mean μ and SD σ of the counts during a non-outbreak period. The method then issues an alert if the daily count exceeds μ by $k\sigma$, where k is ordinarily 2 or 3. Wong and Moore[4] discuss this method and its variations.

Non-spatial, temporal methods that analyze counts look for changes in counts over time that are indicative of an outbreak, but they do not investigate spatial subregions. Such methods include the Serfling method,[5 6] the ARMA, ARIMA, and SARIMA models,[7 8] univariate hidden Markov models,[9 10] support vector machines,[11 12] CUSUM,[13] and Bayesian network models.[14]

Non-temporal, spatial methods that analyze counts detect spatial clusters based only on the most recent counts of an observable event, as for example counts for the current day. These methods do not look for patterns of counts over time. Kulldorff[15] developed a non-temporal, spatial method called the *spatial scan statistic*, which was implemented in the SaTScan software package.[16] A multivariate version of the spatial scan statistic that considers several counts has been developed.[17] Neill *et al*[18] developed a Bayesian spatial scan statistic. This statistic was extended to a multivariate Bayesian spatial scan statistic.[19]

Rather than analyzing data aggregated over the entire population, another approach is to model the effects of the outbreak on each individual patient in population, which is a *patient-specific approach*. Using this approach, we can often more readily model complex changes in behavior. As discussed above, PANDA-CDCA (PC) is a non-spatial, non-temporal, patient-specific method that models the CDC category A outbreak diseases[20] and several other diseases. The PC consists of a large Bayesian network that contains a set of nodes for each individual in a region. Population-wide ANomaly Detection and Assessment (PANDA)[21] is a predecessor to PC; it is a patient-specific system designed to detect a windborne release of anthrax. It has a simple temporal and spatial representation that is much more limited than the approach we describe in the next section.

Even though conceptually a patient-specific system models each patient individually, nevertheless, at the computational level we may be able to analyze it using an aggregated set of counts. More generally, however, patient-specific models may become sufficiently complex that it is no longer feasible to solve them using aggregated counts; in that situation, we can apply other inference methods. See online appendix A for a derivation of the counts analyzed by PC.

We previously showed a method for converting a non-spatial Bayesian detection system to a spatial one.[22] We also developed a method for converting a non-temporal Bayesian detection system to a temporal one.[23] If we can combine the spatial method with the temporal one appropriately, we arrive at a framework for converting a non-spatial, non-temporal Bayesian outbreak detection system to a spatio-temporal one. We describe such a framework next.

MODEL DESCRIPTION

The BNST framework

Suppose we are investigating whether there is an event of interest (eg, a disease outbreak) in some region. Let E be a random variable whose value is *yes* if the event of interest is occurring, and whose value is *no* otherwise. Besides the variable E, we assume a set of attribute variables that represent properties of the event of interest, a set of intermediate variables which depend on the properties of the event of interest, and a set of observable variables that depend on the intermediate variables. These observable variables comprise our *Data* or evidence. Figure 1 presents a Bayesian network framework representing the relationships among these variables. We call this the *Bayesian network spatio-temporal (BNST)* modeling framework and any model represented by it a *BNST model*. Although the framework is suitable for developing systems that detect many types of events, in this paper we focus on modeling disease outbreaks.

Three variables, *SUB*, *Y*, and *F*, are always included in the set of attribute variables. Where *SUB* represents the subregion in which an ongoing outbreak is occurring. The value of *SUB* can either be a hypothesized outbreak subregion *S* or the value *none* if there is no outbreak being hypothesized. The precise definition of a subregion depends on the application. For example, Kulldorff[15] uses a circular subregion and the center of such a subregion is moved over the region. The radius of the circle is also allowed to vary. Neill *et al*[18] represent the entire region *G* by an $m\times n$ rectangular grid, where each grid element is a cell. A subregion *S* of *G* consists of one or more cells that form a rectangle. Note that *G* itself is a subregion. Note also that subregions can overlap. *F* represents the severity of the outbreak on

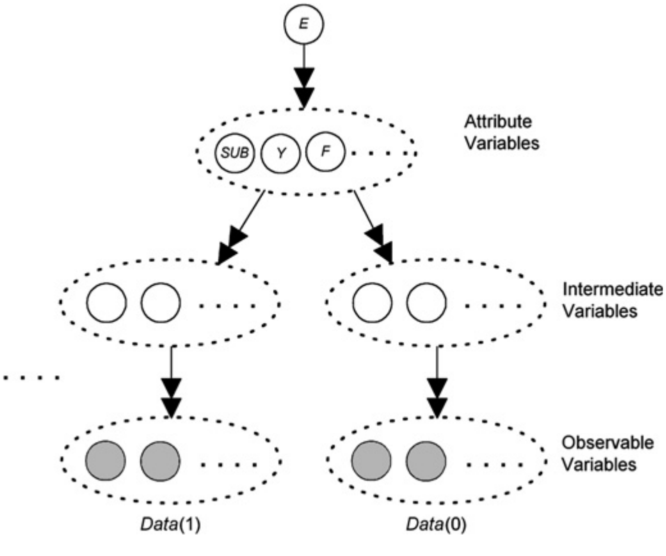

Figure 1 The high-level Bayesian network spatio-temporal (BNST) framework. The value of *E* is *yes* if the event of interest occurred, and is *no* otherwise. The sets of variables enclosed by ovals represent Bayesian subnetworks. The attribute variables are properties of the event of interest. Each intermediate variable depends on the properties of the event of interest. Each observable variable depends on the intermediate variables. The shaded observable variables are the measured variables and comprise our *Data*. The double arrowed edges indicate one or more Bayesian network edges from each variable in a given set to variables in the set below it. There is always one attribute variable *SUB*, whose value is the subregion in which a simulated outbreak is hypothesized to be occurring, one attribute variable *F* representing the severity of the simulated outbreak, and one attribute variable *Y* representing the number of days into the simulated outbreak. The data obtained on day *i* is denoted *Data*(*i*).

the day of investigation, if there is an outbreak. Y represents the number of days into the outbreak if there is an outbreak.

For the intermediate and observable variables, there is a set of these variables for today (day 0) and for each day preceding today (day i denotes i days prior to the current day). The probability distributions of the intermediate variables are conditioned on the values of SUB, Y, and F. The nature of this dependence also depends on the application. The data used for detection consist of information about the observable variables. The data for day i is denoted $Data(i)$. If each intermediate variable represents a property of an individual, we would have a patient-specific model. However, the theory does not require a patient-specific model.

New data are obtained each day (or at whatever the time unit may be) from the entire region being monitored. The Bayesian network is then used to compute the posterior probability of an outbreak along with the posterior probability that each subregion S contains an outbreak.

The conversion of a simple model to a spatio-temporal model

Figure 2A shows a simple example of a non-spatial, non-temporal Bayesian network model. Suppose that the variable E has value *yes* if there is currently an outbreak of influenza and value *no* otherwise. There is a variable I_r for each individual r in the entire region being monitored. The possible values of I_r are the manifestations m_k for the individual. In this example, suppose that they are the chief complaints with which an individual might present to the ED, where one value is *noED*, which means the individual did not visit the ED. Other possible chief complaints include *cough* and *fever/chills*. Note that $I_r{=}m_k$ is an assignment of chief complaint m_k for individual I_r. The probability distribution of I_r depends on whether or not $E{=}yes$. The $Data$ consists of the values of I_r for each individual r in the entire region.

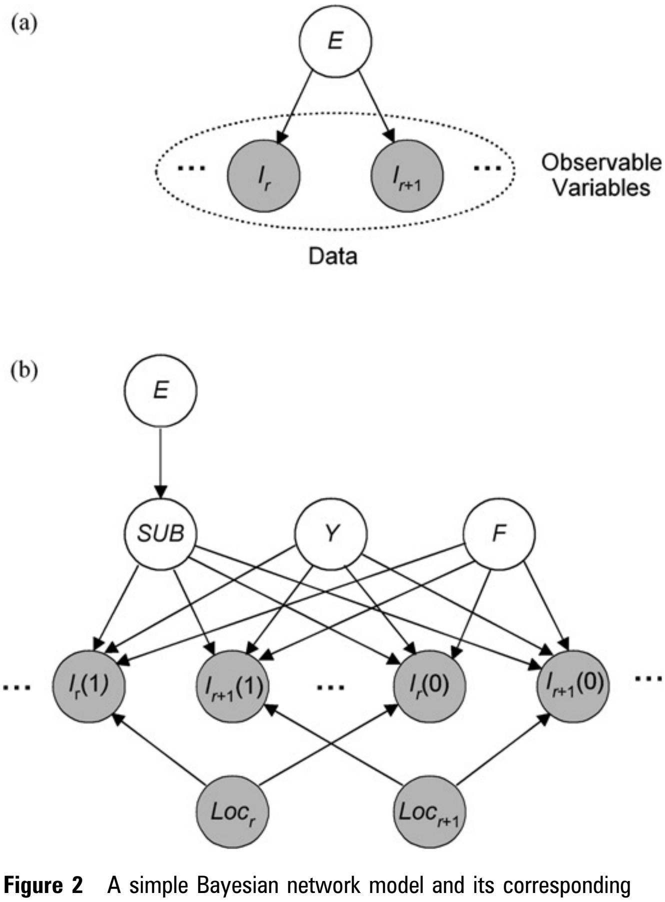

Figure 2 A simple Bayesian network model and its corresponding Bayesian network spatio-temporal (BNST) model.

Figure 2B is the spatio-temporal Bayesian network that results from applying the framework in figure 1 to the model in figure 2A. If there is an outbreak ($E{=}yes$), suppose we assume that the probability of a subregion S containing the event is $1/b$, where b is the number of subregions (ie, a uniform distribution over the subregions is assumed), while if there is no outbreak ($E{=}no$), the probability that there is no event in any subregion is 1. For each individual r, there is a location variable Loc_r, whose value is known at run-time and which represents the individual's home location, such as a zip code. For each value of Loc_r and each value S of SUB, we need to know whether Loc_r is in S. The $Data(i)$ consists of the values of I_r (i) for each individual r in the entire region being monitored. These are the data obtained i days prior to the current day. The probability distribution of I_r (i) depends on whether there is an outbreak in subregion S, whether r is located in S, the severity F of the outbreak today, and the number of days Y into the outbreak. We do not present the quantitative probabilistic details of this dependence here. They are similar to the ones shown below for a more complex model.

A conversion of PC to a spatio-temporal model

This section describes the conversion of PC to a spatio-temporal model.

The PC model

Figure 3 shows the structure of PC.[3] It has two features that make it unique among outbreak detection systems. First, it models a broad set of diseases. Other Bayesian approaches to outbreak detection of which we are aware either model a single disease or do not model any disease in particular. Second, PC represents the effect of an outbreak disease on *individuals* in the population. The current version of PC only uses chief complaints as evidence. However, the PC architecture allows us to model more complex evidence, such as images and narrative text in clinical reports. Such modeling goes well beyond using population counts of observable events.

Table 1 shows the variables in PC and what they represent. Variable O represents which outbreak disease is hypothesized to be occurring, if an outbreak is hypothesized. There are nine possible outbreak diseases resulting in 13 values for O. The diseases include the six CDC category A diseases, which are anthrax (two stages/values), plague (two stages/values), smallpox, tularemia, botulism, and hemorrhagic fever (two stages/values). PC also models influenza, cryptosporidiosis, and hepatitis A. The 13th value of O is *none,* which represents that there is no outbreak. The states of variable F were discretized into 15 real values. This variable indicates the severity of the outbreak. There is a variable D_r for each individual r in the population. Its possible values include all the values of variable O and also the value *other* which means the individual arrived in the ED only with a non-outbreak disease (eg, a broken arm) and the value *noED* which means the individual did not visit the ED. There is a variable I_r for each individual in the region. Its possible values include 54 chief complaints, one of which is *other*, which means the chief complaint was not one of the 53 chief complaints represented in the network. The 55th value of I_r is *noED*, which means the individual did not visit the ED. Appendix A provides more information about PC, including its inference algorithm.

Extending PC to be a spatio-temporal system (PCTS)

We converted PC to the spatio-temporal system PC-Temporal-Spatial (PCTS) according to the BNST framework in figure 1.

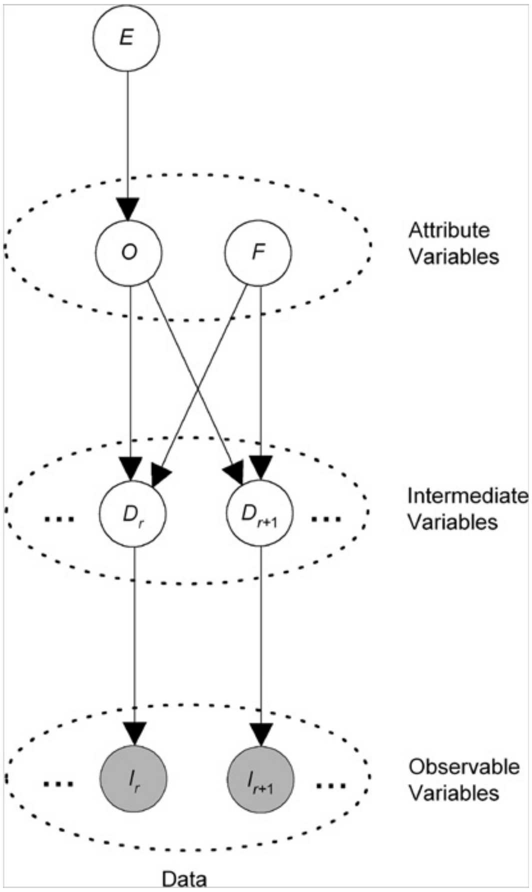

Figure 3 The PANDA-CDCA (PC) Bayesian network structure. See the text for a description of the variables.

The Bayesian network structure of PCTS appears in figure 4. The PCTS computes the posterior probability of each outbreak disease based on the most recent T days of data. Table 2 shows the variables in PCTS that are not in PC.

As in Neill *et al*,[18] the entire region is covered by an $m \times n$ rectangular grid, and each rectangular subregion of the grid constitutes a subregion. These rectangles are the possible values of SUB. PCTS assumes that the prior distribution of Y is uniform over $\{1, 2, \ldots T\}$.

The variables E, O, and F in PCTS are the same variables as in PC. $D_r(i)$ is the ED disease state of the r-th individual i days ago, where $i=0$ represents today. It has the same values as variable D_r in PC. Its conditional probability distributions are discussed below. $I_r(i)$ represents the chief complaint of the r-th individual i days ago. It has all the same properties as variable I_r in PC. Its conditional probability distributions are the ones in PC.

The Bayesian network structure in figure 4 entails that, given values of SUB, Y, O, and F, the ED disease states (values of $D_r(i)$) for an individual on different days are independent. Jiang[23] provides a justification for this independence assumption.

Table 1 The variables in PC and what they represent

Variable	What the variable represents	Variable values
E	Whether there is an ongoing simulated outbreak	Yes, no
O	Simulated outbreak disease	Anthrax, plague, …, none
F	Probability of an individual both having the simulated outbreak disease and going to the ED, given that there is a simulated outbreak	A finite set of numerical values
D_r	ED disease state of the r-th individual	Anthrax, plague, …, other, *noED*
I_r	Chief complaint of the r-th individual	Chest pain, cough, …, other, *noED*

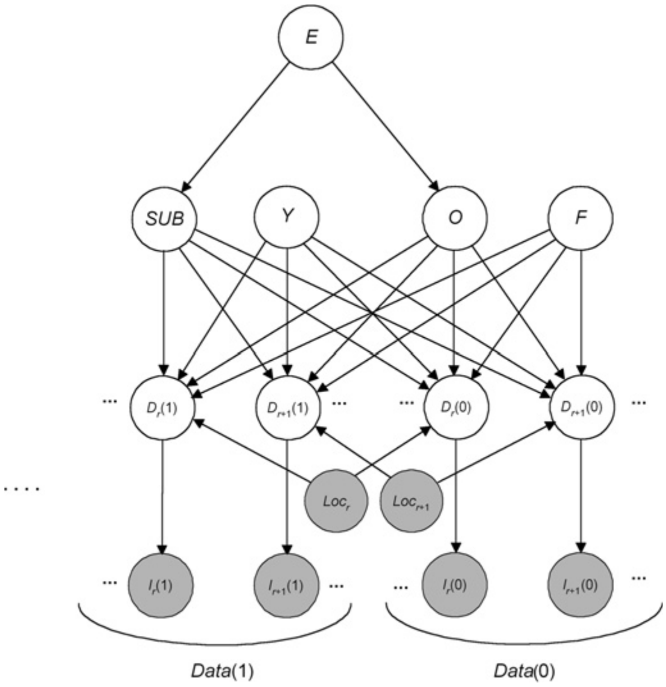

Figure 4 The PANDA-CDCA-Temporal-Spatial (PCTS) Bayesian network structure. Each temporal variable has a numerical label in parentheses. The remaining variables are not temporal.

The probability distribution of $D_r(i)$ is conditioned on O, F, and Y only when $SUB=S$ and individual r is located in S. For each such value of SUB and individual r, the conditional distributions of $D_r(i)$ are derived in the same manner for day i. For individuals not in S, the conditional probability distribution of $D_r(i)$ is the same as the one when O equals *none*, and there is no dependence on F or Y. These are the same conditional distributions as those in PC.

We now describe how the conditional probability distributions of $D_r(i)$ are determined for an individual r located in subregion S. Recall that the value f of F is the probability of an individual both having the outbreak disease and going to the ED today, given that an outbreak is ongoing. Early in the outbreak, which is when we hope to detect it, we assume that the increase in cases can be approximated by a linear function. As discussed in the next section, we tested the robustness of this assumption by simulating outbreaks that did not show a linear increase. We assume that the outbreak is not apparent on the first day it begins, that y days later it reaches level f on the current day, and that the increase over that period of time is linear. Let $F(i)$ denote the probability of an individual being afflicted with the outbreak disease and going to the ED i days ago where $i \leq y$. Based on the previous discussion, $f(i)/(y-i) = f/y$, which implies that

$$f(i) = \frac{y-i}{y} f.$$

Table 2 The variables in PCTS that are not in PC

Variable	What the variable represents	Variable values
SUB	Simulated outbreak subregion	S_1, S_2, …, none
Y	Number of days into the simulated outbreak, if there is a simulated outbreak	1, 2,…, T

The conditional probability distributions for $D_r(i)$ are as follows. Assume that $SUB{=}S$ and that individual r lives in region S. First

$$P(D_r(i) = \text{other}|O = \text{none}, F = f, Y = y) = p_{other},$$

$$P(D_r(i) = noED|O = none, F = f, Y = y) = 1 - p_{\text{other}},$$

where p_{other} is the probability that an individual visits the ED with only a non-outbreak disease, when there is no outbreak. This probability is estimated using ED data from the previous year.

In appendix A we show for the PC non-temporal system with outbreak disease d that

$$P(D_r = \text{other}|O = d, F = f) = p_{\text{other}}(1 - f).$$

Therefore, for PCTS we have that

$$\begin{aligned} P(D_r(i) = d|O = d, F = f, Y = y) &= \frac{y-i}{y}f \quad & i < y \\ &= 0 & i \geq y \end{aligned}$$

$$P(D_r(i) = c|O = d, F = f, Y = y) = 0 \quad \text{for } c \neq d$$

$$\begin{aligned} P(D_r(i) = \text{other}|O = d, F = f, Y = y) &= p_{\text{other}}\left(1 - \frac{y-i}{y}f\right) \quad & i < y \\ &= p_{\text{other}} & i \geq y \end{aligned}$$

$$\begin{aligned} P(D_r(i) = noED|O = d, F = f, Y = y) &= (1 - p_{\text{other}})\left(1 - \frac{y-i}{y}f\right) \quad & i < y \\ &= 1 - p_{\text{other}} & i \geq y. \end{aligned}$$

The inference algorithm for PCTS appears in appendix A.

EVALUATION

In experiment 1 we compared PCTS to PC and to the multivariate, space—time extension of SatScan,[24] which we designate as SaTScan-MT.[i] In experiment 2 we further compared PCTS and SaTScan-MT by evaluating how well they detected outbreaks emerging in time and space.

The dataset for experiment 1

Real ED admission data that were collected from Allegheny County, Pennsylvania in the year 2004 were used as background data. This dataset contains all 110 zip codes in Allegheny County. It contains an average of about 580 ED visits per day. We injected simulated outbreak cases into this real background data to create *semi-synthetic data*. The observed data consisted of chief complaints presented by patients in the ED.

We considered both simulated influenza and cryptosporidiosis outbreaks. The simulated number of new cases increased according to linear, quadratic, and cubic functions during an outbreak before the outbreak reached its peak. For each outbreak disease, we generated 40 outbreaks for each type of increase. Therefore a total of 120 influenza outbreaks and 120 cryptosporidiosis outbreaks were simulated. We chose 40 outbreaks in part based on the statistical rule of thumb[25] that when the sample size is greater than 40, and without outliers, the sampling distribution of many common statistics is approximately normally distributed. We also chose 40 in part for computational tractability reasons.

We simulated outbreaks in subregions in Allegheny County. This county covers 730 square miles, which we modeled using a 16×16 grid. Each grid element is one cell. A zip code was mapped to a cell if the zip code's centroid was in the cell. All 110 zip codes in Allegheny County were mapped to cells in this manner. All the cases for that zip code were placed in its centroid cell. The subregion where an outbreak was simulated to occur was stochastically selected from 12 possibilities. These 12 simulated outbreak subregions consisted of four 2×1 cells, four 2×2 cells, and four 3×2 cells rectangles.

To control the severity of the outbreak, we determined the number of new daily cases based on the SD of the number of real background daily ED visits in the subregion in which the outbreak was simulated. As mentioned previously, a simple control chart method issues an alert if the daily count of some observable event exceeds μ by $k\sigma$, where k is ordinarily 2 or 3. Using this traditional practice as a guideline, we considered an outbreak moderate if the average daily number of ED visits exceeded the mean number of such visits by about 2σ. So for each cell in the outbreak subregion, we determined the mean and SD σ_{cell} of the number of real ED visits during the entire year 2004. The severity level of the outbreak was based on a multiple of σ_{cell}. If, for example, the severity level was based on $2\sigma_{cell}$, the average daily number of new cases was $2\sigma_{cell}$. The duration of all outbreaks was set equal to 30 days. The multiples of σ_{cell} used for each epidemic curve function were as follows. For linear increasing outbreaks, severity level 1 used a daily average of $1.5\sigma_{cell}$ and severity level 2 used $2\sigma_{cell}$; for quadratic increasing outbreaks the values were respectively $2\sigma_{cell}$ and $2.5\sigma_{cell}$; and for cubic increasing outbreaks the values were respectively $2.5\sigma_{cell}$ and $3\sigma_{cell}$. We used larger multiples in the case of the non-linear functions because otherwise it would have taken too long for the number of new cases to reach a detectable level.

We made the simplifying assumption that the daily number of new outbreak cases reaches its peak in the middle of the outbreak and then declines in the same manner in which it increased. So it was assumed that half of the new cases occurred during the first half of the outbreak. In the case of outbreaks that methodically exhibited a linear increase in outbreak cases, we assumed that Δ of them occur on day one of the outbreak, 2Δ occur on day 2, and so on. The value of Δ can therefore be determined by solving $\Delta + 2\Delta + \ldots + 15\Delta = \text{tot}_{\text{cell}}/2$. Similar formulas were used for outbreaks that showed quadratic and cubic increases.

To force daily fluctuations, we deviated from simply making the number of new cases on day t equal to $t\Delta$ (linear case), $t^2\Delta$ (quadratic case), or $t^3\Delta$ (cubic case). We set two daily fluctuation levels, 25% and 50%. Half of the outbreaks were randomly selected to have the 25% level, and the remaining outbreaks

[i] Classic SaTScan looks for clusters of an event in circular subregions of the monitored region. Its space—time extension looks for clusters in three dimensions, where the 3rd dimension is time. So it will investigate 1-day cylinders, 2-day cylinders, etc. This system would not be considered a temporal system according to our definition in the introduction because it does not investigate how patterns change over time. However, it is not a classical non-temporal system either because it does not simply look at one unit of time. It can be viewed as a system that simultaneously investigates clusters using various units of time (1-day unit, 2-day unit, etc).

were given the 50% level. If the level was 25%, on even numbered days we would make the number of new cases 25% of the previous day's number of new cases. Similarly if the level was 50%, on even numbered days we would make the number of new cases 50% of the previous day's number of new cases. We imposed daily fluctuations so we could evaluate the stability of the systems' outbreak assessments in the face of such variation.

Both PC and PCTS contain a distribution P of the chief complaints given the outbreak disease. Simulating chief complaints using P would be biased in favor of PC and PCTS. Therefore, to generate (via stochastic simulation) the chief complaint of each case, we used a variety of different chief-complaint probability distributions that differed significantly from P. For each type of outbreak (influenza or cryptosporidiosis) and for each type of increase (linear, quadratic, or cubic), 10 different distributions were used to generate the data. Each of these 10 distributions was applied to generate data for four of the 40 outbreaks. Appendix A and Jiang[23] discuss these distributions in detail.

The dataset for experiment 2

We generated a total of 120 influenza outbreaks and 120 cryptosporidiosis outbreaks in the same manner as those used in experiment 1 except that the simulated outbreaks were also made to emerge in space. We modeled space emergence by injecting cases into one cell on days 1 and 2 of the outbreak, into two cells on days 3 and 4, into three cells on days 5 and 6, and so forth until all cells in the outbreak subregion were receiving new cases. We chose the first cell to receive outbreak cases by randomly selecting one of the cells in the outbreak subregion; we chose the second cell by randomly selecting a cell from all the cells that touched the first cell, and so on.

Converting patient-specific data to counts

Our datasets consist of semi-synthetic, patient-specific ED data. However, SaTScan-MT looks at one or more cumulative counts of events that are indicative of the event of interest. So, it was necessary to convert the patient-specific data to counts in order to provide input for SaTScan-MT. Without loss of generality, let us discuss the data concerning influenza outbreaks. There are various ways we could develop the counts: (1) provide individual counts of each chief complaint; (2) provide individual counts of only the 20 chief complaints that are manifestations of influenza; or (3) provide individual counts of the few best manifestations of influenza (according to the probability distribution in PC). We did some preliminary experiments with all three approaches, and in the third approach we attempted to improve the input data for SaTScan-MT by trying both the seven best manifestations and the three best manifestations. In these experiments, SaTScan-MT performed best when method (3) was used with the three best manifestations. In this paper we only include that configuration. Appendix A contains additional details.

Evaluation methodology

We used AMOC curves[26] to evaluate the systems' outbreak detection capabilities. In such curves, the false alarm rate is plotted on the x-axis and the mean days to detection on the y-axis. If an outbreak is not detected, the days to detection is set to a penalty. In our experiments we set it to 30, the number of days in the simulated outbreak.

The detection performances at various false alarm rates were further compared using a Bayesian test of significance, which is the posterior probability that one system's average time to detection (at a given false alarm rate) is greater than that of another system's under the assumption of prior ignorance (see Jiang[23] for details). We will present the results using Bayesian posterior probabilities instead of p values.

We are also interested in how well the systems maintained the detection of an outbreak. The AMOC-M curves are used to evaluate this, where the *M* stands for *maintain*. An AMOC-M curve[23] is like an AMOC curve except that the y-axis plots the average of the time at which an outbreak signal is detected and maintained thereafter. For example, if the outbreak-detection threshold is 0.05, and the sequence of daily outbreak posterior probabilities from a system is (0.01, 0.02, 0.07, 0.03, 0.04, **0.08**, 0.09, 0.06, 0.08, 0.07), then the time at which the signal is maintained is 6 because on the 6th day the probability is 0.08, which exceeds 0.05, and it stays at or above 0.05 for the remaining days in the sequence that we are considering.

The PCTS and SaTScan-MT not only detect an outbreak, but also determine the subregion in which the outbreak is occurring. We compared their accuracy concerning the detection of the correct subregion using the overlap coefficient. Let S be the actual outbreak subregion, T be the hypothesize outbreak subregion, and # return the number of zip codes in a subregion. Then

$$\text{overlap coefficient}(S,T) = \frac{\#(S\cap T)}{\#(S\cup T)},$$

where $\cap$ denotes set intersection. The overlap coefficient is 0 if and only if the two subregions do not intersect (overlap), while it is 1 if and only if they overlap exactly.

Results for experiment 1

We evaluated how well PC-Temporal-Spatial (PCTS) and PC detected the outbreak disease that was injected (rather than just detecting that any outbreak is occurring). As discussed above, SaTScan-MT was configured to detect the simulated outbreak disease. The posterior probability of the outbreak disease was used as the detection signal for PC and PCTS, and the likelihood ratio of the most likely subregion was used as the signal for SaTScan-MT.

The results shown here concern the performance over all 120 outbreaks for each of influenza and cryptosporidiosis. Our purpose is to compare the systems' overall performances relative to a variety of outbreak behaviors.

Let $P(\mu_{PCc}>\mu_{PCTSc})$ be the posterior probability that PCTS has a smaller mean day to detection than PC for cryptosporidiosis outbreaks, and $P(\mu_{PCf}>\mu_{PCTSf})$ be that probability for influenza outbreaks. Let $P(\nu_{PCc}>\nu_{PCTSc})$ be the posterior probability that PCTS has a smaller mean day until *maintaining* detection than PC for cryptosporidiosis outbreaks, and $P(\nu_{PCf} >\nu_{PCTSf})$ be that probability for influenza outbreaks. For false alarm rates (r) equal to 0, 5, 10, and 15, our results showed that $P(\mu_{PCc} >\mu_{PCTSc})$, $P(\mu_{PCf}>\mu_{PCTSf})$, $P(\nu_{PCc}>\nu_{PCTSc})$, and $P(\nu_{PCf} >\nu_{PCTSf})$ are all greater than 0.9999. Each of these probabilities was obtained using Bayesian statistics under an assumption of prior ignorance. Jiang[23] shows under the assumptions being made that such a posterior probability is equal to 1 minus the corresponding p value. For example, if $P(\mu_{PCc} > \mu_{PCTSc})=0.96$ then we can reject the hypothesis that μ_{PCc} is less than or equal to μ_{PCTSc} at the 0.04 significance level. It seems more straightforward to state our results using posterior probabilities rather than p values. These results support hypothesis 1, namely that when an outbreak is occurring in a small subregion, PCTS will detect the outbreak earlier than PC. The results also support hypothesis 2, namely that PCTS is better at maintaining the detection signal.

Table 3 A comparison of PANDA-CDCA-Temporal-Spatial (PCTS) to SaTScan-MT at various false alarm rates (*r*)

r	$P(\mu_{SATc}>\mu_{PCTSc})$	$P(\mu_{SATf}>\mu_{PCTSf})$	$P(\nu_{SATc}>\nu_{PCTSc})$	$P(\nu_{SATf}>\nu_{PCTSf})$
0	>0.9999	0.5962	>0.9999	0.2471
5	>0.9999	0.0237	>0.9999	<0.0001
10	>0.9999	0.1711	>0.9999	<0.0001
15	>0.9999	0.4288	>0.9999	<0.0001

$P(\mu_{SATc}>\mu_{PCTSc})$ is the posterior probability that PCTS has a smaller mean days to detection than SaTScan-MT for cryptosporidiosis simulated *outbreaks*, and $P(\nu_{SATf}>\nu_{PCTSf})$ is the posterior probability that PCTS has a smaller mean days until maintaining detection than SaTScan-MT for influenza simulated outbreaks.

Table 3 shows the posterior probability that PCTS has a smaller mean day to detection (μ) than SaTScan-MT and the posterior probability that PCTS has a smaller mean day until maintaining detection (ν) than SaTScan-MT at various false alarm rates (*r*). The PCTS detected cryptosporidiosis outbreaks significantly earlier than SaTScan-MT at false alarm rates of 0, 5, 10, and 15. However, SaTScan-MT detected influenza outbreaks significantly earlier than PCTS at a false alarm rate of 5, while results for rates of 0, 10, and 15 were not significant (in the sense that the probability must be less than 0.05 or greater than 0.95 to be considered significant). We see also from table 3 that PCTS maintained detection of cryptosporidiosis outbreaks significantly earlier than SaTScan-MT at all false alarm rates, but SaTScan-MT maintained detection of influenza outbreaks significantly earlier than PCTS at most false alarm rates.

Figure 5A,B shows AMOC curves comparing the outbreak detection performance, while Figure 5C,D shows AMOC-M curves comparing the detection *maintenance* performance. These curves are consistent with the results discussed above.

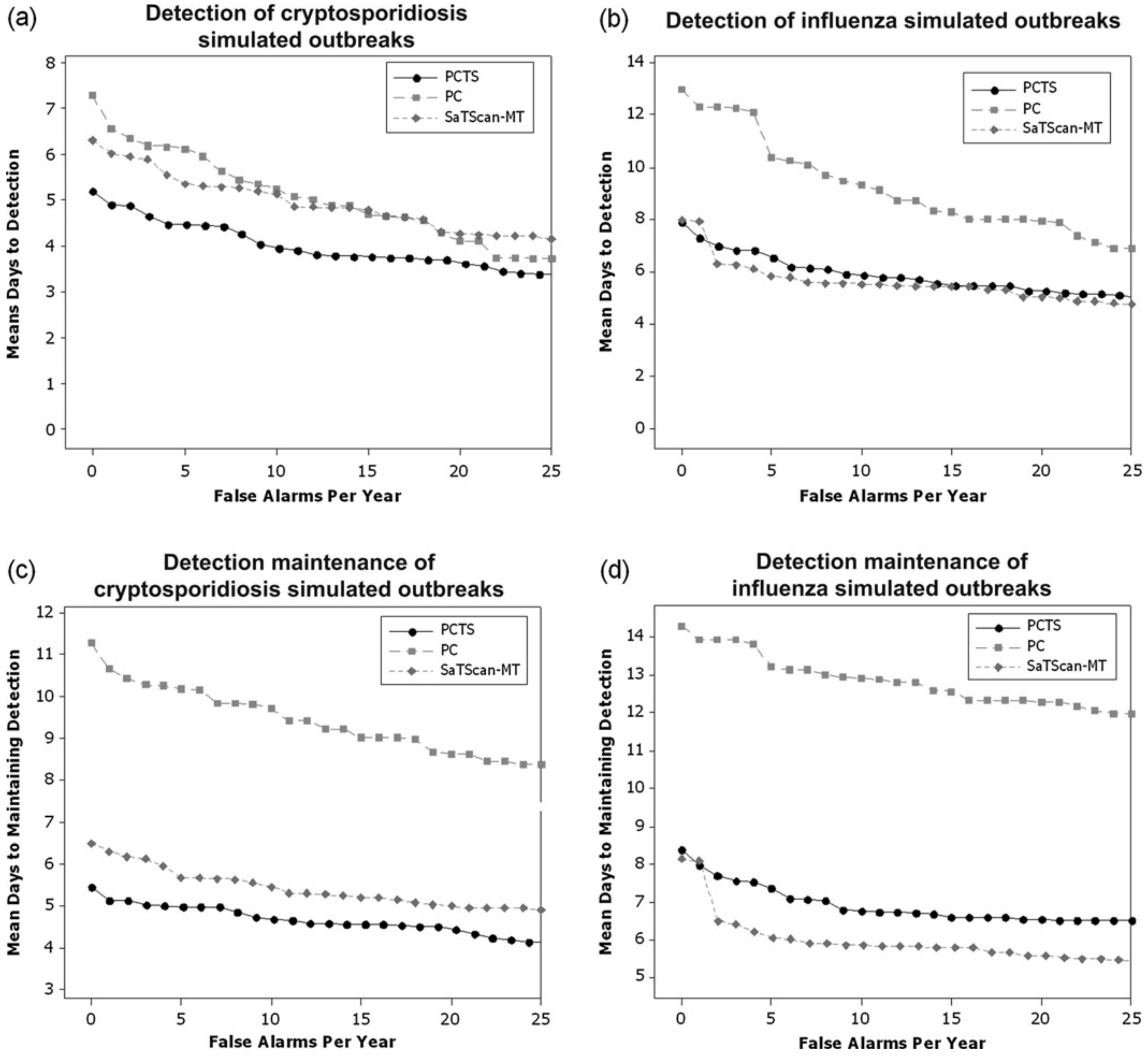

Figure 5 (A, B) AMOC curves. (C, D) AMOC-M curves.

Table 4 For various values of the day into the simulated outbreak, the posterior probability that PCTS has a larger overlap coefficient than SaTScan-MT

	Simulated outbreaks not emerging in space		Simulated outbreaks emerging in space	
Day	$P(O_{PCTSc}>O_{SATc})$	$P(O_{PCTSf}>O_{SATf})$	$P(O_{PCTSc}>O_{SATc})$	$P(O_{PCTSf}>O_{SATf})$
1	0.4008	0.7176	0.2774	0.9694
5	>0.9999	0.9575	0.8599	0.9601
10	0.9944	>0.9999	0.9379	0.8350
15	>0.9999	>0.9999	>0.9999	0.9990

The 2nd and 3rd columns concern the simulated outbreaks in experiment 1, which were emerging in time but not in space. The 4th and 5th columns concern the simulated outbreaks in experiment 2 which were emerging in both space and time.

Appendix A (tables A.1 and A.2) contains an analysis of sensitivity that shows the fraction of the outbreaks that are ever detected. The sensitivity of PCTS is greater than that of PC at all false alert rates. The sensitivity of PCTS and SaTScan-MT are both 100 per cent for cryptosporidiosis at all false alert rates. For influenza detection, SaTScan-MT tends to have a sensitivity that is 5–10% better than that of PCTS, depending on the false alert rate.

The second and third columns of table 4 shows the posterior probability that PCTS has a larger average overlap coefficient (O) than SaTScan-MT at various values of the *Day* into the outbreak. The subregion S that maximized $P(\text{Data}|\text{SUB} = S)$ was considered to be the subregion detected by PCTS, and the subregion that maximized the likelihood ratio was considered to be the one detected by SaTScan-MT. Figure 6 shows the average values of the overlap coefficient for PCTS and SaTScan-MT plotted against the *Day* into the outbreak. Notice that values fluctuate up and down on a daily basis. This is consistent with the imposed daily fluctuations in the number of new outbreak cases. The PCTS outperformed SaTScan-MT for both types of outbreaks. Furthermore, its average value was about 0.7 by the seventh day of both types of outbreaks. These results support the hypothesis that when an outbreak is occurring in a small subregion, PCTS can closely estimate that subregion.

The average running times for the three systems, taken over all 240 outbreaks, were as follows: PC: < 1 s.; PCTS: 350 s.; SaTScan-MT: 21 s. The implementation of SaTScan-MT used four processors, whereas that of PCTS used only one. So the running time of PCTS compares a little more favorably than the times indicate, but it is still somewhat slower. Not surprisingly, PC's running time was much faster, given that it is based on a simpler model. We also expected that PCTS would be slower than SaTScan-MT because PCTS is a more complex system that does more computations. As discussed above, PCTS is a true temporal system in that it models emerging patterns in time while SaTScan-MT does not. Furthermore, PCTS is a patient-specific system which is normally more computationally costly than a system that takes accumulated counts as input.

Results for experiment 2

In this experiment, we further compared PCTS and SaTScan-MT by evaluating how well they detected outbreaks that were emerging in both time and space. Table 5 shows the posterior probability that PCTS has a smaller mean day to detection (μ) than SaTScan-MT and the posterior probability that PCTS has a smaller mean day to maintaining detection (ν) than SaTScan-MT at various false alarm rates (r). The results concerning detection are similar to those for non-emerging outbreaks, except that PCTS faired better with the influenza outbreaks. That is, in the case of cryptosporidiosis outbreaks, PCTS performed significantly better than SaTScan-MT at both detection and detection maintenance for emerging outbreaks. In the case of influenza outbreaks, PCTS overall performed better at detection, although not statistically significantly so, and in most cases did not perform statistically significantly worse at detection maintenance.

Figure 7 shows AMOC curves comparing PCTS and SaTScan-MT when detecting emerging outbreaks. These curves are consistent with the results just discussed.

The fourth and fifth columns of table 4 show the posterior probability that PCTS has a larger average overlap coefficient (O) than SaTScan-MT at various values of the *Day* into the outbreak for emerging outbreaks. The results are similar to those for non-emerging outbreaks. That is, PCTS usually outperformed SaTScan-MT for both types of outbreaks.

DISCUSSION

We developed a method for converting a non-spatial, non-temporal Bayesian outbreak-detection system to a spatio-temporal

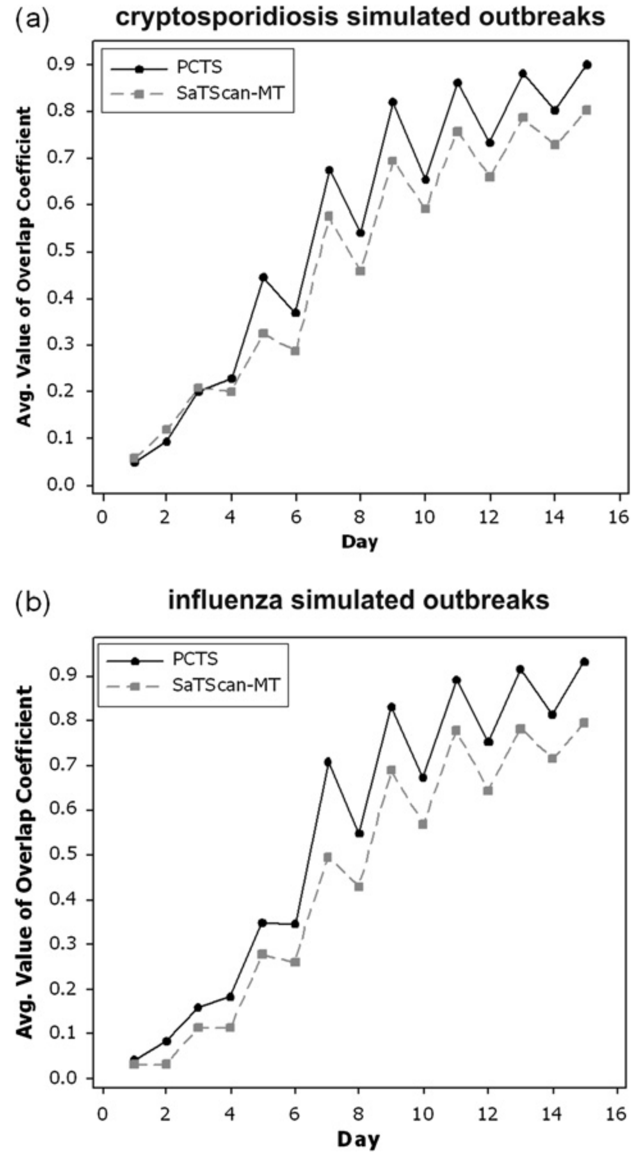

Figure 6 The average values of the overlap coefficient for PANDA-CDCA-Temporal-Spatial (PCTS) and SaTScan-MT.

Table 5 A comparison of PANDA-CDCA-Temporal-Spatial (PCTS) to SaTScan-MT at various false alarm rates (*r*) for simulated outbreaks that were emerging in both space and time

r	$P(\mu_{SATc} > \mu_{PCTSc})$	$P(\mu_{SATf} > \mu_{PCTSf})$	$P(\nu_{SATc} > \nu_{PCTSc})$	$P(\nu_{SATf} > \nu_{PCTSf})$
0	>0.9999	0.6691	>0.9999	0.6918
5	0.9951	0.0693	>0.9999	0.0125
10	>0.9999	0.6818	>0.9999	0.0971
15	0.9988	0.9404	>0.9999	0.1012

The probabilities have the same meaning as those in table 3.

one. Using this framework, we extended the PC system to create the spatio-temporal system PCTS. We hypothesized that: (1) when an outbreak is occurring in a small subregion, PCTS will detect the outbreak earlier than PC; (2) PCTS is better than PC at maintaining the detection signal; (3) PCTS can closely estimate the subregion in which an outbreak is occurring; and (4) PCTS will perform as well or better than an existing state-of-the-art spatial disease-outbreak detection system.

The experimental results support that PCTS provides improved disease outbreak detection, relative to PC from which it was derived. Besides often detecting outbreaks earlier (at a given false alert rate), PCTS was better at maintaining a stable detection signal over time.

The PCTS is a patient-specific system that models each individual in a population. We also hypothesized that such a system might obtain better detection performance than one that uses

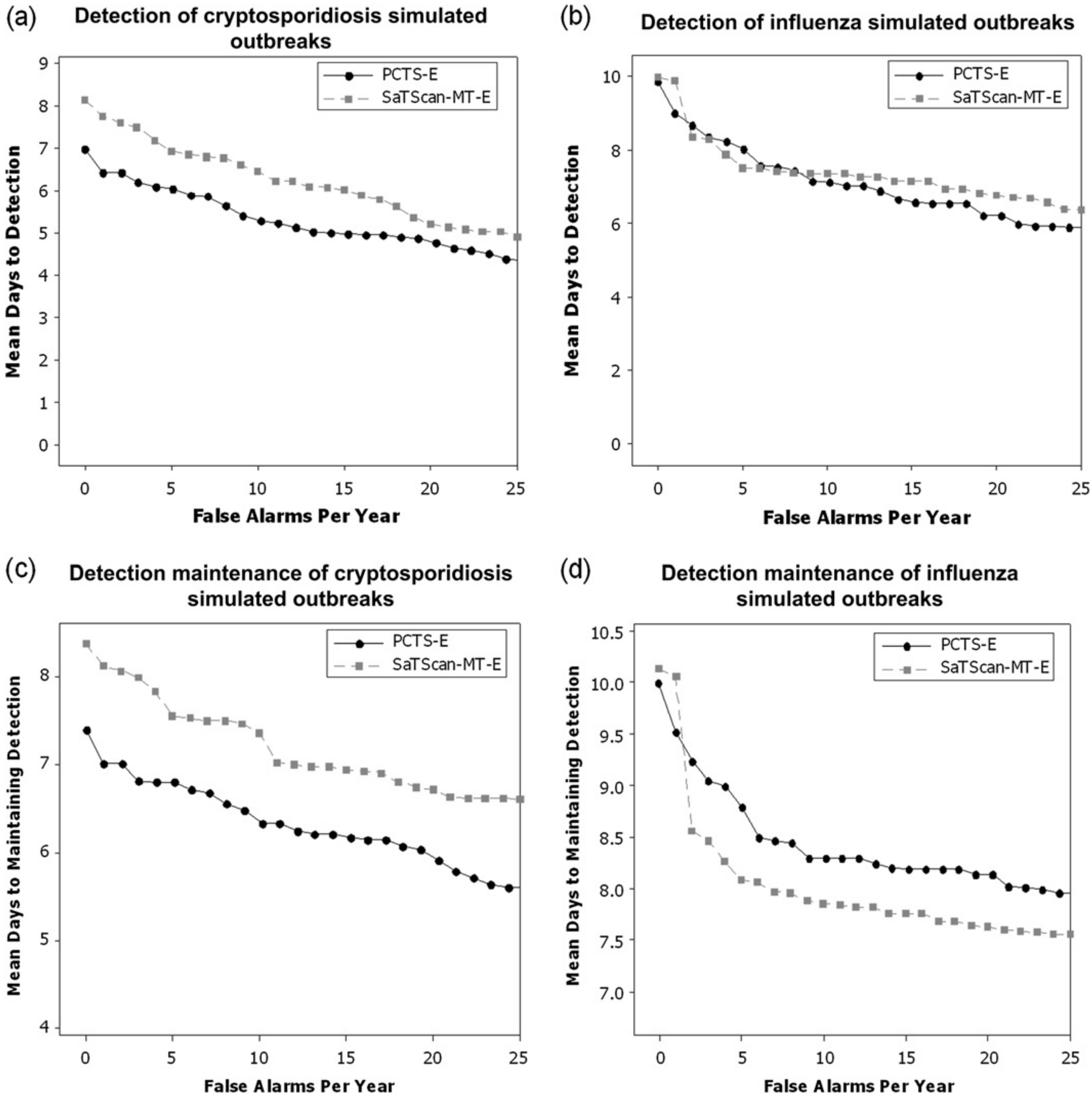

Figure 7 (A, B) AMOC curves for simulated outbreaks emerging in space and time. (C, D) AMOC-M curves for such simulated outbreaks.

a summary statistic such as daily counts. To test this hypothesis, we compared PCTS to SaTScan-MT. In the case of cryptosporidiosis outbreaks, PCTS performed substantially better than SaTScan-MT at outbreak detection, detection maintenance, and subregion detection.

In the case of influenza outbreaks, however, PCTS performed at a level comparable to SaTScan-MT at outbreak detection and worse at outbreak detection maintenance. SaTScan-MT had a modestly better sensitivity for detecting influenza. The PCTS substantially outperformed SaTScan-MT at subregion detection. One reason for these mixed results may be the following. Influenza outbreaks are fairly difficult to detect because influenza has symptoms such as cough and fever, which are common to several other outbreak diseases and often even occur when there is no outbreak. The PCTS ordinarily detects that some outbreak is occurring better than it detects that a specific outbreak is occurring, and this should hold even more for a difficult disease such as influenza. The SaTScan MT does not simultaneously investigate different outbreaks; it only looks at counts of observable events. As discussed above and in more detail in appendix A, SaTScan-MT performed poorly when we gave it counts of all 20 chief complaints that are manifestations of influenza, and performed much better when we only gave it the top three chief complaints. We only showed the results for its better performance. The PCTS simultaneously models 12 different outbreaks and reports both the overall probability of an outbreak and the probability of an outbreak of each outbreak disease. If we had compared PCTS's ability to detect any one outbreak to SaTScan-MT or if we had optimized PCTS to only investigate an influenza outbreak by looking only at the top three chief complaints, PCTS may have faired better in the comparisons.

It is reasonable to ask whether SaTScan-MT may have been at a disadvantage in the evaluation, due to it modeling circular outbreak regions. Two points attenuate concern. First, each outbreak we generated consisted of counts occurring in zip codes whose centroids were in a simulated outbreak rectangle. Thus, these outbreak subregions were not strictly rectangular in shape. Second, we used rectangles that were 2×1 cells, 2×2 cells, and 3×2 cells, which are not highly elongated shapes. These two points mean that the simulated outbreaks consisted of relatively compact subregions that were not restricted to be rectangular in shape.

Our comparisons were done using simulated outbreaks which showed methodical daily fluctuations. A more realistic comparison would be to evaluate the performance of the systems using real outbreaks. Unfortunately, there are not that many real outbreaks with available data.

Being Bayesian and allowing patient-specific modeling imparts several advantages to the BNST framework. Since it is Bayesian we can incorporate both expert knowledge and knowledge learned from data into a BNST system. A method such as SaTScan bases its analysis only on the current data and is not able to take advantage of prior knowledge. The Bayesian framework, however, has a price: to properly use the framework we need to devote considerable effort to obtaining and representing prior knowledge. This knowledge/belief could change over time and from one region to another, and thus be hard to manage and maintain. On the other hand, a data-driven system, such as SaTScan, can readily be applied in many contexts without modification. Thus, both approaches have advantages and disadvantages. Nonetheless, we believe it may be possible to significantly improve methods for acquiring, managing, and maintaining prior knowledge about disease outbreaks within a Bayesian framework, and we believe this topic to be a promising area in need of considerable additional research.

Funding This research was supported by grant IIS-0325581 from the National Science Foundation.

Competing interests None.

Ethics approval This study was conducted with the approval of the University of Pittsburgh.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Kaufmann A,** Meltzer M, Schmid G. The economic impact of a bioterrorist attack: are prevention and postattack intervention programs justifiable. *Emerg Infect Dis* 1997;**3**:83—94.
2. **Wagner MM,** Tsui FC, Espino JU, *et al*. The emerging science of very early detection of disease outbreaks. *J Public Health Manag Pract* 2001;**7**:51—9.
3. **Cooper GF,** Dowling JN, Lavender JD, *et al*. Bayesian algorithm for detecting CDC category A outbreak diseases from emergency department chief complaints. *Adv Dis Surv* 2007;**2**:45.
4. **Wong WK,** Moore A. Classical time series methods for biosurveillance. In: Wagner M, ed. *Handbook of biosurveillance*. New York: Elsevier, 2006: 217—34.
5. **Serfling RE.** Methods for current statistical analysis of pneumonia-influenza deaths. *Public Health Rep* 1963;**78**:494—506.
6. **Tsui FC,** Wagner MM, Dato V, *et al*. Value of ICD-9-coded chief complaints for detection of epidemics. *Symposium of the J Am Med Inform Assoc* 2001;**9**:41—7.
7. **Box G,** Jenkins G, Reinsel G. *Time series analysis: forecasting and control*. Englewood Cliffs, New Jersey: Prentice Hall, 1994.
8. **Hamilton J.** *Time series analysis*. Princeton, New Jersey: Princeton University Press, 1994.
9. **Rabiner LR.** A tutorial on hidden Markov models and selected applications in speech recognition. In: Waibel A, Lee KF, eds. *Readings in speech recognition*. Burlington, MA: Morgan Kaufmann, 1990:267—96.
10. **Moore A.** A powerpoint tutorial on hidden Markov models [monograph on the Internet]. 2001. http://www.autonlab.org/tutorials/hmm.html.
11. **Burges C.** A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* 1998;**2**:121—67.
12. **Moore A.** A powerpoint tutorial on support vector machines [monograph on the Internet]. 2001. http://www.autonlab.org/tutorials/svm.html.
13. **Bos T,** Fetherston TA. Market model nonstationarity in the Korean stock market. In: Rhee SG, Chang RP, eds. *Pacific-basin capital markets research 3*. North-Holland, Amsterdam: Elsevier, 1992:287—302.
14. **Jiang X,** Wallstrom GL. A Bayesian network for outbreak detection and prediction. *Proceedings of AAAI-06*; 16—20 July 2006; Boston, MA: 1166—0.
15. **Kulldorff M.** A spatial scan statistic. *Commun Stat Theory Methods* 1997;**26**:1481—96.
16. **Kulldorff M.** SaTScan v. 4.0: software for the spatial and space-time scan statistics [software on the internet]. 2004. http://www.satscan.org/.
17. **Kulldorff M,** Mostashari F, Luiz D, *et al*. Multivariate scan statistics for disease surveillance. *Stat Med* 2007;**26**:1824—33.
18. **Neill DB,** Moore AW, Cooper GF. A Bayesian spatial scan statistic. *Adv Neural Inf Process Syst* 2005;**18**:1003—10.
19. **Neill DB,** Cooper GF. A multivariate Bayesian scan statistic. *Adv Dis Surv* 2007;**2**:60.
20. Centers for Disease Control and Prevention [homepage on the internet]. www.bt.cdc.gov/agent/ agentlist-category.asp.
21. **Cooper GF,** Dash DH, Levander JD, *et al*. Bayesian biosurveillance of disease outbreaks. In: Maxwell D, Halpern J, eds. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2004:94—103.
22. **Jiang X,** Cooper GF, Neill DB. A Bayesian network model for spatial event surveillance. *International Journal of Approximate Reasoning* 2010;**51**:224—39.
23. **Jiang X.** *A Bayesian network model for spatio-temporal event surveillance [dissertation]*. Pittsburgh (PA): University of Pittsburgh, 2008.
24. **Kulldorff M,** Heffernan R, Hartman J, *et al*. Space-time permutation scan statistic for disease outbreak detection. *PLoS Med* 2005;**2**:216—24.
25. **Levine-Wissing R,** Thiel D. *AP Statistics*. Piscataway, NJ: Research and Education Association, 2006.
26. **Fawcett T,** Provost F. Activity monitoring: noticing interesting changes in behavior. *Proceedings of the 5th SIGKDD Conference on Knowledge Discovery and Data Mining*; 15-18 August 1999; San Diego, CA: 53—62.