Data Min Knowl Disc (2010) 20:328–360
DOI 10.1007/s10618-009-0151-4

A real-time temporal Bayesian architecture for event surveillance and its application to patient-specific multiple disease outbreak detection

Xia Jiang · Gregory F. Cooper

Received: 17 March 2009 / Accepted: 9 September 2009 / Published online: 30 October 2009

Abstract Reliable and accurate detection of disease outbreaks remains an important research topic in disease outbreak surveillance. A temporal surveillance system bases its analysis on data not only from the most recent time period, but also on data from previous time periods. A non-temporal system only looks at data from the most recent time period. There are two difficulties with a non-temporal system when it is used to monitor real data which often contain noise. First, it is prone to produce false positive signals during non-outbreak time periods. Second, during an outbreak, it tends to release false negative signals early in the outbreak, which can adversely affect the decision making process of the user of the system. We conjecture that by converting a non-temporal system to a temporal one, we may attenuate these difficulties inherent in a non-temporal system. In this paper, we propose a Bayesian network architecture for a class of temporal event surveillance models called BayesNet-T. Using this Bayesian network architecture, we can convert certain non-temporal surveillance systems to temporal ones. We apply this architecture to a previously developed non-temporal multiple-disease outbreak detection system called PC and create a temporal system called PCT. PCT takes Emergency Department (ED) patient chief complaint data as its input. The PCT system was constructed using both data (non-outbreak diseases) and expert assessments (outbreak diseases). We compare PCT to PC using a real influenza outbreak. Furthermore, we compare PCT to both PC and the classic statistical methods CUSUM and EWMA using a total of 240 influenza and Cryptosporidium disease

Responsible editor: R. Bharat Rao and Romer Rosales.

X. Jiang (✉) · G. F. Cooper
Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
e-mail: xij6@pitt.edu

G. F. Cooper
e-mail: gfc@pitt.edu

outbreaks created by injecting stochastically simulated outbreak cases into real ED admission data. Our results indicate that PCT has a smaller mean time to detection than PC at low false alarm rates, and that PCT is more stable than PC in that once an outbreak is detected, PCT is better at maintaining the detection signal on future days.

Keywords Temporal disease outbreak detection · Bayesian network · Patient-specific model · Mining ED chief complaint data · Uncertainty modeling · Biosurveillance

1 Introduction

Event surveillance consists of analyzing a region in order to detect patterns that are indicative of some event of interest. As examples, we may look for patterns that are indicative of a forthcoming disaster or a disaster that is in its early stages. Examples of such disasters include hurricanes, terrorist attacks, and outbreaks of diseases. A classic example of event surveillance involves monitoring some geographical region in order to detect a disease outbreak. In what follows, the focus will be on disease outbreak surveillance. *Disease outbreak surveillance* monitors a community for the onset of a disease outbreak. A popular term for *disease outbreak surveillance* is *biosurveillance*. Reliable and accurate detection of disease outbreaks remains an important research topic in disease outbreak surveillance.

On a given day, the number of disease cases could of course exceed the expected number by chance, and then return to normal. Ordinarily, this would not be considered a disease outbreak. A disease outbreak is characterized by a statistically increasing trend (with daily fluctuations) in cases until some peak is reached, then a decline, and then possibly an increase to a second peak, and so on. So, the pattern of a disease outbreak is emerging over time. A temporal disease outbreak surveillance system looks for emerging patterns by analyzing how the situation has changed recently in time. The analysis is based not only on the data from the most recent time period, but also on the data from previous time periods. A non-temporal outbreak detection system only looks at data from the most recent time period such as the previous 24 h. The analysis would not look at data from previous time periods. A non-temporal method can in principle be used to investigate an emerging disease outbreak, and, in fact, many of the existing disease outbreak detection systems are non-temporal (see Sect. 2.1.1). However, there are two difficulties with such a method.

First, a non-temporal system should have difficulty providing a small mean time to detection at a low false alarm rate. Ordinarily, an event detection system returns a numeric signal, and we issue an alert/alarm when the value of that signal exceeds some threshold. A non-temporal detection system looks only at the data from the current day (or whatever the unit of time might be). On a given day it is not uncommon for the signal to be relatively high due to a random occurrence or to some non-outbreak event. For example, a drug store may have increased drug sales on a given day due to a store promotion. If we were using sales of some particular drug as our signal, that signal could become high due to the promotion rather than due to an outbreak. When the background data contains many such incidental anomalies, a non-temporal system

will issue many false alarms unless we set the threshold to a high value. However, if we do this we will increase our time to detection during an actual outbreak.

On the other hand, a temporal system looks at data from the current day and previous days, which means a 1-day spike would be less likely to lead to the system returning a high value of its signal. Therefore, we conjecture that a temporal system will have a smaller mean time to detection at a low false alarm rate than a non-temporal system. In general, we want our systems to run with few false alarms because if the system issues too many false alarms, it loses its credibility.

Second, a non-temporal system can confuse the user. Although ordinarily there is a general trend for the number of disease cases to increase during an outbreak, there can be great fluctuation in the daily numbers. Figure 1a shows an epidemic curve carefully reconstructed by health care officials for a *Cryptosporidium* outbreak that occurred in North Battleford, Saskatchewan in spring, 2001. The outbreak starts on about March 20 and ends on about April 5. Although there is a general trend for daily counts to increase until the peak is reached, we see significant fluctuation in the daily counts, and early in the outbreak the counts sometimes return to pre-outbreak values. Due to the fluctuation in the daily counts, a system that looks only at the current day's data may exhibit considerable fluctuation in its posterior probability for an outbreak early in the outbreak, thereby confusing the user as to whether or not there truly is an outbreak.

As an example, PANDA-CDCA (PC) (Cooper et al. 2007) is a disease outbreak detection system that uses a Bayesian network to model the relationships among the events of interest and those observed. PC is a patient-specific system, because rather than analyzing data aggregated over the entire population (i.e., daily counts of some observable events), it monitors each individual patient case in the population. We shall discuss the patient-specific system further in Sect. 2. PC is also a multiple-disease outbreak detection system, which monitors simultaneously 12 outbreak diseases and their variations. However, PC does not use a temporal model of disease outbreaks. Cooper et al. (2007) obtained results that were surprising at the time when evaluating the ability of PC to detect a laboratory validated outbreak of influenza in Allegheny County. Under a false alarm rate of zero, PC detected influenza approximately 1 day before the first positive viral cultures of influenza were taken. However, near the beginning of

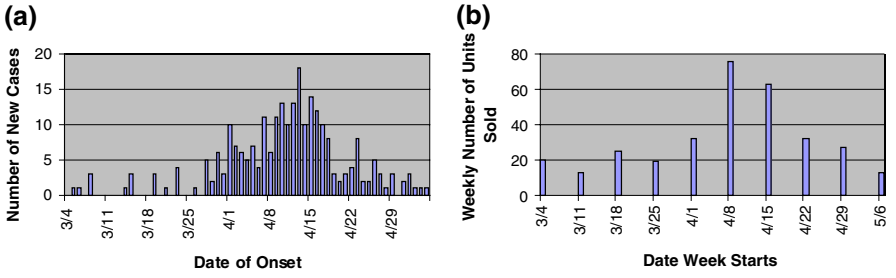

Fig. 1 **a** An epidemic *curve* for a *Cryptosporidium* disease outbreak in North Battleford, Saskatchewan. **b** Weekly OTC sales of antidiarrheal drugs at one pharmacy in North Battleford. The data for these *curves* were obtained from Stirling et al. (2001)

the influenza outbreak, the posterior probability of influenza fluctuated between very high and very low values.

We conjecture that by converting a non-temporal system to a temporal one, which models that during an outbreak the number of outbreak cases is expected to steadily increase but possibly with daily variation, we would attenuate the problems in a non-temporal system.

In this research, we developed a high-level Bayesian network architecture representing a class of temporal event surveillance models called BayesNet-T. Using this high-level Bayesian network architecture, it is possible to construct a temporal model from an existing Bayesian network model for non-temporal event surveillance. Using this high-level Bayesian network architecture, we developed a system which detects the same outbreaks as PC but considers temporal aspects of disease outbreaks. We call this temporal system PANDA-CDCA-TEMPORAL (PCT). We hypothesize that (1) PCT will have a smaller mean time to detection than PC at low false alarm rates; and (2) PCT will be more stable than PC in that once an outbreak is detected, PCT will be better at maintaining the detection signal on future days. Like PC, PCT also monitors multiple outbreak diseases and during the outbreak bases its analysis on information from each individual case in the population. So we can describe PCT as a temporal, patient-specific, multiple-disease outbreak detection system. To our knowledge, PCT is the first such system to incorporate all of these elements. PCT is real-time in nature. It is designed to run repeatedly each day and detect the event of interest that is currently occurring.

In Sect. 2, we review representative methods for event surveillance. We then describe BayesNet-T and PCT in Sect. 3. Section 4 presents results of experiments evaluating PCT's performance.

2 Methods for event surveillance

This section presents a representative, although not exhaustive, review of methods for event surveillance.

2.1 Methods that analyze counts

Often the count of occurrences of some phenomenon increases during a disease outbreak. For example, as noted previously, Fig. 1a shows an epidemic curve constructed from a sample of the population affected by a *Cryptosporidium* outbreak in North Battleford, Saskatchewan in spring 2001. The outbreak was caused by a contamination of public drinking water. *Cryptosporidium* infection causes diarrhea. Figure 1b shows the weekly counts of units of over-the-counter (OTC) antidiarrheal medicine sold at one pharmacy in North Battleford during the time period affected by the outbreak. The correlation between these two curves suggests that by monitoring OTC sales of such medicine we can possibly detect a *Cryptosporidium* outbreak at an early stage. Similarly, the number of patients visiting the emergency department (ED) with respiratory symptoms ordinarily increases during an influenza outbreak.

To monitor and analyze the counts, we first choose a unit of time, which is ordinarily 1 day, but could be 1 hour (h), or any other unit. For the sake of discussion, in what follows it is assumed that the time unit is 1 day. A count of some characteristic of the outbreak is then obtained separately for each day.

2.1.1 Non-temporal methods

Non-temporal methods consider counts from some recent period of time only, such as the previous 24 h. One method for analyzing these daily counts is to first derive the mean μ and standard deviation σ of the daily counts over a period of time when no outbreak is presumed to be occurring, and fix these values in the outbreak detection system. An alert is then issued whenever the daily count exceeds μ by $k\sigma$, where k is usually 2 or 3. Wong and Moore (2006) discuss problems with this method and improvements to it.

In ordinary (non-spatial) event surveillance, an entire region is monitored globally. For example, if we were monitoring whether a disease outbreak was occurring in a particular county, we would monitor the entire county globally, without considering the possibility of localized outbreaks in subregions. If an outbreak was occurring in a small subregion of a county and the entire county was monitored globally, the outbreak may go undetected until it spread to a larger subregion. In *spatial event surveillance*, we search for patterns in spatial subregions. That is, we individually monitor both small and large subregions of the region of interest. In this way, we not only may detect an emerging event sooner, but we may also learn its location. Spatial cluster detection is one statistical technique used for spatial event surveillance. Methods for *spatial cluster detection* attempt to locate spatial subregions of some larger region where the count of occurrences of some event is higher in one subregion relative to other subregions. The classic technique for analyzing these counts is the spatial scan statistic (Kulldorff 1997). A Bayesian version of the spatial scan statistic appears in Neill et al. (2005a,b).

2.1.2 Temporal methods

Temporal methods detect an outbreak based on how the situation has changed recently in time. The determination of an outbreak is based not only on the count from the most recent day, but also on counts from previous days.

There are a number of temporal (time series) methods that consider the count of occurrences of a single phenomenon. Some of these methods are discussed in Wong and Moore (2006). Known methods include the Serfling method (Serfling 1963; Tsui et al. 2001), the ARMA, ARIMA, and SARIMA models (Box et al. 1994; Hamilton 1994), univariate hidden Markov models (Rabiner 1989; Moore 2001a), Kalman filters (Burges 1998), support vector machines (Burges 1998; Moore 2001b), and CUSUM (Bos and Fetherston 1992). Other frequentist methods appear in Reis and Mandl (2003), Reis et al. (2003), and Soneson and Bock (2003). Baron (2002) developed a method based on solving a suitable change-point problem. A Bayesian method is developed in Jiang and Wallstrom (2006). Temporal versions of the spatial scan statistic appears in Kulldorff et al. (2005) and Neill et al. (2005a,b).

CUSUM is one of the most widely used temporal methods. CUSUM analyzes the counts from the previous i time units (e.g., days). Let μ_0 be the mean of the counts during some background period when no outbreak is occurring (the in-control process mean), σ be the standard deviation of the counts during the background period, and $\mu_1 = \mu_0 + \delta\sigma$ for a constant δ (μ_1 is the out-of-control process mean). Then we define a slack value K as follows:

$$K = \frac{|u_1 - \mu_0|}{2} = \frac{\delta\sigma}{2}.$$

Let $X_1, X_2, \ldots X_i$ be the counts from each of the past i time units. To determine when to signal an alert, we monitor the following time series of statistics S_i:

$$\begin{aligned} S_1 &= \max(0, X_1 - (\mu_0 + K)) \\ &\vdots \\ S_i &= \max(0, S_{i-1} + X_i - (\mu_0 + K)). \end{aligned}$$

We signal an alert whenever

$$S_i > H,$$

where $H = d\sigma$ for a constant d. In Montgomery (2001) it is recommended to let $d = 5$.

Exponentially weighted moving average (EWMA) is a statistical quality control technique. EWMA is a variation of the moving average method in which we assign weights to observations in an exponentially decreasing order according to the age of the observations, with the observations further in the past weighing less. EWMA can be considered a temporal method because, similar to other moving average algorithms, it bases its analysis not only on the value of the most time step but also on the values from the previous time steps. Wong and Moore (2006) discuss in detail how EWMA can be applied to perform outbreak detection.

Methods that look at several counts are called *multivariate*. Multivariate temporal methods appear in Moore et al. (2006) and Shmueli and Fienberg (2006). The Bayesian method developed in Jiang and Wallstrom (2006) can look at several counts, but in the implementation which they evaluated it did not. A multivariate version of the spatial scan statistic appears in Kulldorff et al. (2007). Kulldorff (2004) developed the software package SaTScan™, which allows the user to simultaneously do both multivariate and temporal spatial modeling.

2.2 Patient-specific methods

Rather than analyzing data aggregated over the entire population (i.e., daily counts of some observable events), another approach is to model the relationships between an outbreak disease and the effect of the outbreak on each individual in a population. This is a patient-specific approach. By modeling each individual in the population, we can base our analysis on more information than that contained in a summary statistic,

such as the number of patients who visited the ED with respiratory symptoms on a given day.

2.2.1 *Non-temporal patient-specific methods*

PC (Cooper et al. 2007) is a non-temporal method that models the CDC Category A diseases (see www.bt.cdc.gov/agent/agentlist-category.asp) and also several other diseases. It consists of a large Bayesian network which contains a set of nodes for each individual in a region. PC takes as input a time series of chief complaints, one for each ED patient in the region. There are 54 chief complaints, including a catchall category of other complaints. Each hour, based on the previous 24 h (1 day) of data, it outputs the posterior probability of each disease. PC not only can inform us if an outbreak is likely, but also what type of outbreak it might be. Additional details of PC are provided in Sect. 3.1.3.

PC is a non-temporal outbreak detection system, as mentioned, and it is also non-spatial. We are interested in understanding how each of spatial and temporal extensions of PC affect its performance. We previously developed a spatial extension of PC that is described and investigated in Jiang et al. (2009), which showed that such an extension can significantly improve detection performance. The purpose of the current paper is to introduce a temporal extension of PC and investigate how that extension affects its detection performance, relative to PC and to two traditional non-spatial, time-series methods, CUSUM and EWMA.

BARD (Bayesian aerosol release detector) (Hogan et al. 2007) is a Bayesian network, patient-specific system designed to compute the posterior probability of an outdoor, wind-borne release of anthrax spores. BARD's goal is to perform earlier, more sensitive detection of wind-borne outbreaks by recognizing a characteristic dispersion pattern. Its input includes meteorological data, such as wind direction and speed for the region being monitored. It not only detects an outbreak, but also characterizes release location, quantity, and time.

2.2.2 *Temporal patient-specific methods*

A predecessor to PC called PANDA (Population-wide ANomaly Detection and Assessment) (Cooper et al. 2004) is also a patient-specific, Bayesian network system that has a simple temporal model of an outbreak disease. PANDA is designed specifically to detect non-contagious outbreak diseases such as airborne anthrax or West Nile encephalitis. PANDA is able to detect disease outbreaks due to inhalational anthrax, and the only clinical evidence considered by this system is whether an individual presented to the ED with respiratory symptoms or not. The Bayesian network in PANDA contains a set of nodes for each individual in a region. These person nodes represent properties of the individual such as age, gender, home location, the anthrax infection state of the individual, and the ED admission state of the individual. The Bayesian network also contains a global node representing the location of the anthrax release and a global node representing the time of the anthrax release. Temporal information is represented by states of nodes in the network. For example, the global node *Time of Release* has states *never*, *today*, *yesterday*, and *day before yesterday*. Although PANDA models

the time period of the outbreak (in days), it does not model a progressive increase in the number of expected outbreak cases over time.

3 BayesNet-T and its application to PC

The BayesNet-T architecture builds on a non-temporal architecture called BayesNet. So first we describe BayesNet and show that PC is in the BayesNet class of event surveillance models. Then we develop BayesNet-T.

3.1 The BayesNet event surveillance models

This section presents a description of the high-level Bayesian network architecture representing the BayesNet class of event surveillance models and gives several concrete examples.

3.1.1 The high-level Bayesian network architecture

Suppose we are investigating whether there is an event of interest in some region. Let E be a random variable whose value is yes if the event of interest occurred or is occurring, and whose value is no otherwise. Besides the variable E, there can be a set of attribute variables which 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 which depend on the intermediate variables. These observable variables comprise our *Data*. Figure 2 shows a high-level Bayesian network architecture representing this class of models. Any model in this class is called a *Bayesian Network (BayesNet) model*. If each intermediate variable represents an individual in a population, and there is a set of observable variables for each such individual, it would be a patient-specific model. In this paper only BayesNet models that are patient-specific will be considered. However, the theory does not require that they be patient-specific. For example, suppose E represents the occurrence of an influenza outbreak, and the only observable variable is C, which is the count of OTC sales of thermometers. The variable C depends on E, and we can model this dependency using the DAG $E \rightarrow C$. This is a BayesNet model containing no attribute or intermediate variables and which is not patient-specific.

In a non-temporal model, new data are obtained each day (or at whatever our time unit may be) from the entire region being monitored. The Bayesian network is then used to compute

$$P(E = \text{yes}|\text{Data}).$$

3.1.2 A simple example of a BayesNet model

3.1.2.1 The model Figure 3 shows a simple example of a BayesNet model, which has no global or intermediate variables. For the sake of concreteness, let us give the

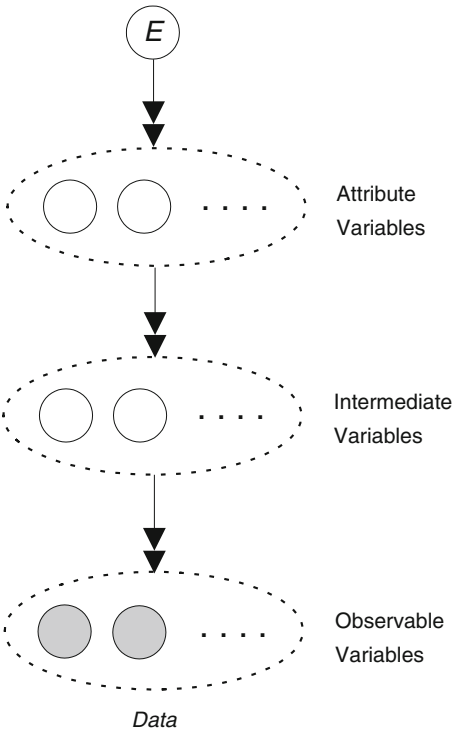

Fig. 2 The high-level BayesNet Bayesian network architecture. The value of E is yes if the event of interest occurred, and is no otherwise. The sets of variables enclosed by *ovals* represent Bayesian subnetworks. The attribute variables are properties of the event of interest, the intermediate variables depend on the properties of the event of interest, and the observable variables depend on the intermediate variables. The *shaded* observable variables are the measured variables and comprise our $Data$. The *unshaded* variables are unmeasured. The *double arrowed* edges indicate one or more edges from each variable in a given set to variables in the set below it. In general, there need not be any attribute or intermediate variables

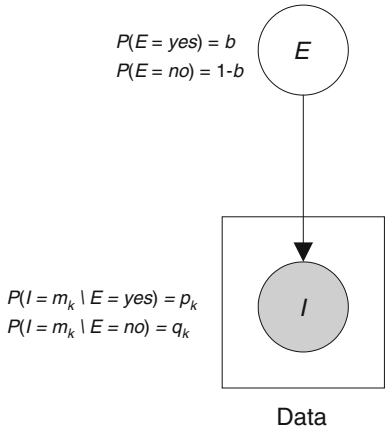

Fig. 3 A simple example of a BayesNet model

variables meaning. Suppose that the variable E has value yes if there is currently an outbreak of influenza and the value no otherwise. The plate representation in Fig. 3 indicates that there is a variable I for each individual in the entire region G being monitored for influenza. So this is a patient-specific system. There are no variables describing properties of the event per se (beyond data about entities in the population) and no intermediate variables. The possible values of I are our manifestations m_k for each individual. In this example, suppose that they are the chief complaints with

which the individual might present in the Emergency Department, where one value is noED, which means the individual did not visit the Emergency Department. In this example, other possible chief complaints might include cough, and fever/chills. Note that $I = m_k$ is an assignment of chief complaint m_k for individual I.

3.1.2.2 The inference algorithm The *Data* consist of the values of I for all individuals in region G. Since there could be thousands, or even millions, of individuals in G, we would not explicitly construct the Bayesian network in Fig. 3, and instantiate I for each individual. Rather, due to the fact that the Bayesian network structure entails that individuals' chief complaints are conditionally independent given the value of E, we can compute the likelihoods of the data as follows:

$$P(\text{Data}|E = \text{yes}) = \prod_k (p_k)^{C_k}$$
$$P(\text{Data}|E = \text{no}) = \prod_k (q_k)^{C_k},$$

where C_k is the number of individuals with the kth chief complaint, and p_k and q_k are defined in Fig. 3. Then using Bayes' Theorem, we compute that

$$P(E = \text{yes}|\text{Data}) = \frac{P(\text{Data}|E = \text{yes})P(E = \text{yes})}{P(\text{Data}|E = \text{yes})P(E = \text{yes}) + P(\text{Data}|E = \text{no})P(E = \text{no})}.$$

3.1.3 PANDA-CDCA

We now describe a more complex example of a BayesNet model, namely the Bayesian network in PANDA-CDCA (PC) (Cooper et al. 2007). Note that although PC was previously developed, we generalized it to create the BayesNet architecture (Fig. 2), which is an innovative contribution of this paper.

Figure 4 shows the Bayesian network in PC. Each node in the network along with its parameters is described. PC is a hybrid network in that some of the parameter values were obtained from expert knowledge and some were learned from data. Since we often do not know for sure when an outbreak starts and whether a patient has the outbreak disease, there is little reliable data concerning patients presenting in the ED with outbreak diseases. However, we have much more data that is reliable concerning patient visits to the ED when there is no outbreak. So the probabilities concerning outbreak diseases were obtained from expert judgment, whereas those concerning non-outbreak diseases were obtained from data. We believe that outbreak detection is an excellent domain in which to investigate this interesting hybrid approach to modeling, which is based combining expert knowledge and data mining.

We now describe each node in the network.

E: This node represents whether there is an ongoing outbreak. The value yes represents that there is an ongoing outbreak of one of the outbreak diseases represented by O during all or some of the previous 24-h period.

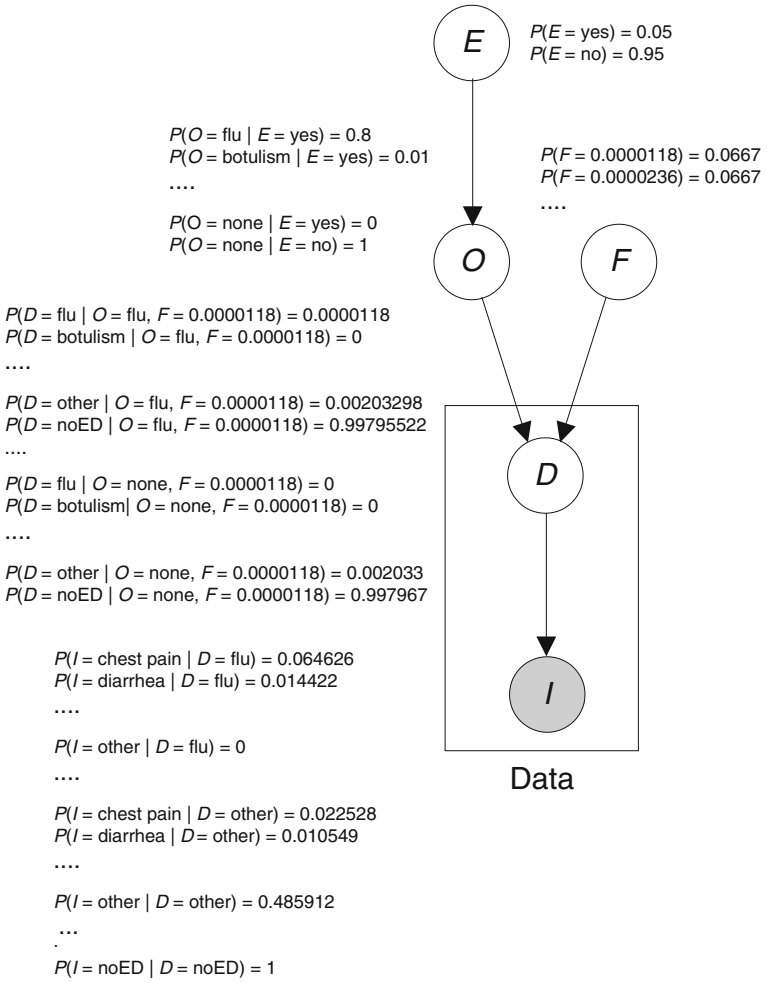

Fig. 4 The PC Bayesian network. See the text for a description of the variables

O: This node represents which outbreak disease is occurring if there is an outbreak. The prior probabilities for variable O were assessed by the project's infectious disease expert based on the literature and subjective estimates. There are 13 possible outbreak diseases, two of which are shown in Fig. 4 (influenza and botulism). The possible outbreak diseases include the following CDC Category A diseases: anthrax stage 1, anthrax stage 2, plague stage 1, plague stage 2, smallpox, tularemia, botulism, marburg hemorrhagic fever stage 1, and marburg hemorrhagic fever stage 2. PC also models the following outbreak diseases: influenza, *Cryptosporidium*, and hepatitis A.
The 13th value of O is none, which represents a population-disease state in which there is no ongoing outbreak disease. Note that $P(O = \text{none}|E = \text{no}) = 1.0$.

Thus, each of the outbreak diseases listed above has a probability of 0, when $E = \text{no}$.

PC assumes that outbreak diseases are mutually exclusive. For example, it assumes there would not be influenza and botulism outbreaks occurring simultaneously. Although in reality different outbreaks could occur concurrently, this event is unlikely, and therefore the model currently assumes it does not happen.

F: The value of this node is itself a probability. Given that an outbreak is ongoing, this node represents the probability of an individual in the population both being afflicted with the outbreak disease and going to the ED on the current day. This node indicates the extent (severity) of the outbreak, if one is occurring. For computational efficiency reasons, the states of this node were discretized into 15 numerical values, two of which are shown in Fig. 4. This node indicates the extent of the outbreak, if one is occurring. Since the value f of F is a probability, the probability distribution of F is a higher order probability distribution.

The possible values of F correspond to expected number of outbreaks cases of 5, 10, 15, 20, 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, and 275 according to the following calculations. The mean number of ED cases per day when there is not an outbreak is estimated to be 577 according to the 2004 biosurveillance data, and the standard deviation is about $\sigma = 54$. It was assumed that the expected value of the increased number of ED cases during an outbreak ranged between $0.1\sigma = 0.1 \times 54 \approx 5$ to $5\sigma = 5 \times 54 \approx 275$. The 15 values above were then taken from this range. Finally, the values of F were obtained by dividing these numbers by 423,076, which is the number of people in the population. For example, $5/423,076 = 1.18 \times 10^{-5}$.

D: This node represents the ED disease state of the an individual. The plate representation in Fig. 4 indicates that there is one such node for each individual in the population. Thus PC is a patient-specific system. The node's value could be any one of the outbreak diseases (values of O) if the individual has the outbreak disease. Additionally, it could have value other which means the individual arrives in the ED only with some non-outbreak disease (e.g., a broken arm). Finally, it could have value noED which means the individual does not visit the ED.

The probabilities for node D were obtained as follows. If there is no outbreak occurring in the population, it is assumed the individual could not have an outbreak disease. Therefore, when there is no outbreak, the individual could arrive in the ED only with a non-outbreak disease. The probability of this event is called p_{other}. So when there is no outbreak, the probability of not visiting the ED is $1 - p_{\text{other}}$. These probabilities are estimated using the ED data from the previous year. In the experiments described in Sect. 4, we used data collected in Allegheny County from 2005 for testing outbreak detection performance. Therefore, we used data from 2004 for estimating model parameters. In 2004 in Allegheny County the probability estimates for PC were as follows:

$$P(D = \text{other} | O = \text{none}, F = f) = p_{\text{other}} = 0.002033$$
$$P(D = \text{no}ED | O = \text{none}, F = f) = 1 - p_{\text{other}} = 0997967.$$

The probabilities of arriving in the ED with outbreak diseases given there is an outbreak of disease d is based on the value of F as follows:

$$P(D = d|O = d, F = f) = f$$
$$P(D = c|O = d, F = f) = 0,$$

where c is an outbreak disease not equal to d. We assume that the probability of someone having both an outbreak disease and an non-outbreak disease (other) is $p_{\text{other}} \times f$. We therefore have

$$\begin{aligned} P(D = \text{no}ED|O = d, F = f) &= 1 - p_{\text{other}} - f + p_{\text{other}} \times f \\ &= (1 - p_{\text{other}})(1 - f). \end{aligned}$$

Finally,

$$\begin{aligned} &P(D = \text{other}|O = d, F = f) \\ &= 1 - P(D = d|O = d, F = f) - P(D = \text{no}ED|O = d, F = f) \\ &= 1 - f - (1 - p_{\text{other}})(1 - f) = p_{\text{other}}(1 - f). \end{aligned}$$

I: This node represents each of the possible chief complaints that an individual could have when arriving in the ED. The plate representation in Fig. 4 indicates that there is one such node for each individual in the population. There are 54 possible chief complaints, one of which is other, which means the chief complaint was not one of the 53 specific chief complaints represented in the network. The 55th value of the node is noED, which means the individual did not visit the ED and thus did not have a literal chief complaint. As mentioned previously, we do not have good ED data during disease outbreaks. So the conditional probabilities for this node were based on the knowledge of an infectious disease expert.

3.2 The BayesNet-T temporal event surveillance models

This section describes the high-level Bayesian network architecture representing the BayesNet-T class of temporal event surveillance models and then gives a concrete example.

3.2.1 The high-level Bayesian network architecture

We start with the high-level Bayesian network architecture in Fig. 2. Then two additional random variables, Y and F, are added to the set of attribute variables. These variables are defined as follows:

F: severity of the outbreak if there is an ongoing outbreak.
Y: number of days into the outbreak, if there is an ongoing outbreak.

The specific nature of the variable F depends on the particular application. Note that PC has a variable F that represents the severity of the outbreak (see Fig. 4). However,

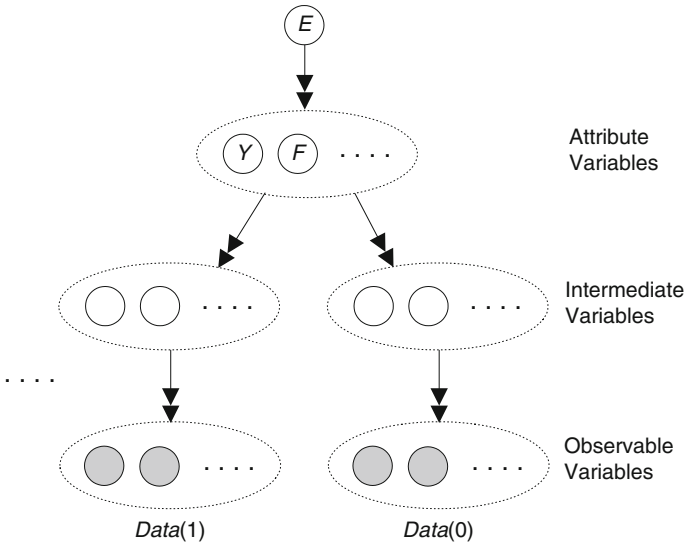

Fig. 5 The high-level BayesNet-T Bayesian network architecture. The discussion in the caption of Fig. 2 pertains to this figure. There is always one attribute variable F representing the severity of the outbreak and one attribute variable Y representing the number of days into the outbreak. The set of variables labeled $Data(0)$ denotes the data collected today, the set of variables labeled $Data(1)$ denotes the data collected yesterday, and so on

in general a BayesNet model need not have a variable F, whereas a BayesNet-T model requires one. As to the intermediate and observable variables, there are 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). Their probability distributions are conditional on the values of F, Y, and the day i. The nature of this dependence also depends on the application. The data on day i is denoted $Data(i)$. A high-level Bayesian network architecture representing this class of models appears in Fig. 5. Any model in this class is called a *Bayesian Network Temporal (BayesNet-T) model*.

3.2.2 PANDA-CDCA-TEMPORAL

In this section we develop the temporal system PCT.

3.2.2.1 The model Figure 6 shows the Bayesian network structure for PANDA-CDCA-TEMPORAL (PCT), which is the BayesNet-T system derived from PC (Fig. 4). Each day PCT bases its outbreak posterior probabilities on the most recent T days (including today) of ED data. We will now describe the nodes in the network.

- E : This node represents whether there is an ongoing outbreak. It is the same node as in PC, and its probability distribution is the same as the one in PC.
- O : This node represents which outbreak disease is occurring given that there is an outbreak. It is the same node as in PC, and its probability distribution is the same as the one in PC.

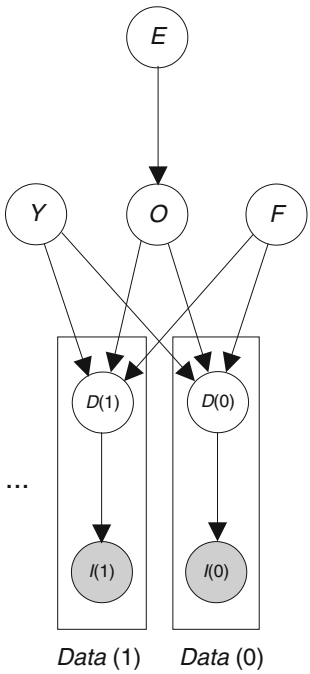

Fig. 6 The Bayesian network structure in PCT

F: If there is an ongoing outbreak then this variable represents its severity at the current time; see Sect. 3.1.3 for details.

Y: This node represents the number of days into the outbreak, as of today, if there is an ongoing outbreak. The prior probability over its values is a uniform distribution over $\{1, 2, \ldots T\}$, where T is the maximum time span over which we are modeling an outbreak.

$D(i)$: This node represents the ED disease state of the an individual i days ago, where $i = 0$ represents today. It has the same values as node D in PC. Its conditional probability distribution will be discussed shortly.

$I(i)$: This node represents the chief complaint of the an individual i days ago. It has all the same properties as the node I in PC. Its conditional probability distributions are the same as those in PC.

The probability distribution of $D(i)$ is conditional on O, F, and Y. The Bayesian network structure in Fig. 6 entails that given values of O, F, and Y, the ED disease states (values of $D(i)$) and therefore the chief complaints (values of $I(i)$) for an individual on different days are independent. For example, conditional on these three variables, if an individual went to the ED yesterday with influenza, it does not change the probability that the individual will go to the ED today with influenza. This assumption allows for a given individual going to the ED two or more times during an outbreak. This is certainly possible, especially in the case of an outbreak disease with severe symptoms. Justification for this independence assumption is provided at the end of this section.

To develop the conditional probability distribution for node $D(i)$, it is useful to first define the following random variable:

$F(i)$: Probability of an individual both being afflicted with the outbreak disease and going to the ED i days ago.

Recall that the value f of F is the probability of an individual both being afflicted with 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 the outbreak, it is reasonable to assume that the increase in cases can be approximated by a linear increase. Therefore, we assume that the value $f(i)$ of $F(i)$ is related to the values f of F and y of Y as shown in Fig. 7. This assumption entails that the outbreak extent is at level 0 when we are 0 days into the outbreak, reaches level f today, and the increase over that period of time is linear. The assumption of linearity is a first-order approximation to the way in which different types of outbreaks might exhibit an increase. In the experiments in Sect. 4.2.1 we test the robustness of this assumption by developing outbreaks with non-linear increases in the number of outbreak-disease cases.

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

which implies that

$$f(i) = \frac{y-i}{y} f. \tag{1}$$

Given the Eq. 1 and the discussion in Sect. 3.1.3 concerning PC, the conditional probability distributions for $D(i)$ are as follows:

$$\begin{aligned} P(D(i) = \text{other} | O = \text{none}, F = f, Y = y) &= p_{\text{other}} \\ P(D(i) = \text{no}ED | O = \text{none}, F = f, Y = y) &= 1 - p_{\text{other}} \\ P(D(i) = d | O = d, F = f, Y = y) &= \tfrac{y-i}{y} f \quad i < y \\ &= 0 \qquad\quad i \geq y \end{aligned} \tag{2}$$

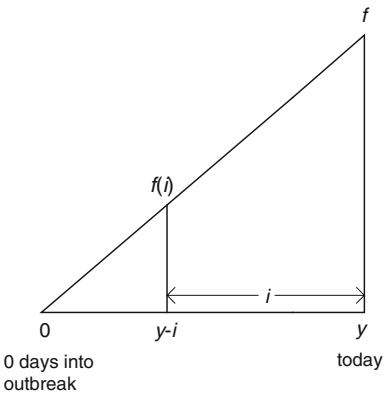

Fig. 7 Number of days into the outbreak is plotted *horizontally*, and the prevalence of the outbreak is plotted *vertically*

$$
\begin{aligned}
&P(D(i) = c|O = d, F = f, Y = y) = 0 \quad \text{for } c \neq d \\
&P(D(i) = \text{other}|O = d, F = f, Y = y) \\
&\qquad = p_{\text{other}} \left(1 - \tfrac{y-i}{y} f\right) \quad i < y \\
&\qquad = p_{\text{other}} \qquad\qquad\qquad\quad i \geq y. \qquad (3) \\
&P(D(i) = \text{no}ED|O = d, F = f, Y = y) \\
&\qquad = (1 - p_{\text{other}}) \left(1 - \tfrac{y-i}{y} f\right) \quad i < y \\
&\qquad = 1 - p_{\text{other}} \qquad\qquad\qquad\quad i \geq y. \qquad (4)
\end{aligned}
$$

Let us discuss the boundary condition in Eq. 2. Recall that Y is uniformly distributed between 1 and T. It is assumed that we must be at least 1 day into the outbreak for the individual to contract the disease and arrive with it in the ED. The value of $D(i)$ is a given individual's ED disease state i days ago. If $i \geq y$, it is the individual's disease state before we are into the outbreak. For example, if $y = 1$, we are 1 day into they outbreak today, and so if $i \geq 1$, then i days ago we had not progressed into the outbreak yet. Therefore, in that case the probability of the individual having the outbreak disease is 0 as Eq. 2 entails. A similar discussion pertains to the boundary conditions in Eqs. 3 and 4.

3.2.2.2 The inference algorithm On each day i, we know the value of $I(i)$ for each individual in the population. Data(i) is the set of these values i days ago, and Data is the set of all these values.

Since the data items are conditionally independent given that $O = none$, we have that

$$
P(\text{Data}|E = \text{no}) = P(\text{Data}|O = \text{none}) = \prod_{i=0}^{T-1} P(\text{Data}(i)|O = \text{none}). \qquad (5)
$$

Note that the product goes from 0 to $T - 1$, which means we look at T days of data.

The terms in the product on the right in Eq. 5 are given by

$$
P(\text{Data}(i)|O = \text{none}) = \prod_k (P(I(i) = m_k|O = \text{none})^{C_k(i)},
$$

where $C_k(i)$ is the number of individuals i days ago with chief complaint m_k. Note that one of the chief complaints is noED, which means the individual did not visit the ED. The reason we can compute $P(\text{Data}(i)|O = \text{none})$ by multiplying the individual probabilities is that the nodes representing the manifestations of different individuals (the nodes represented by the plate I) are conditionally independent given values of O, F, Y, and by construction F's value and Y's value are irrelevant when $O = \text{none}$.

The value of $P(I(i) = m_k|O = \text{none})$ for each patient who went to the ED could be computed by performing inference using the Bayesian network in Fig. 6. However, we can obtain this value more efficiently as follows:

$$
\begin{aligned}
&P(I(i) = m_k|O = \text{none}) \\
&= \sum_c P(I(i) = m_k|D(i) = c)P(D(i) = c|O = \text{none}) \\
&= P(I(i) = m_k|D(i) = \text{other})P(D(i) = \text{other}|O = \text{none}) \\
&\quad + P(I(i) = m_k|D(i) = \text{no}ED)P(D(i) = \text{no}ED|O = \text{none}) \\
&= P(I(i) = m_k|D(i) = \text{other}) \times p_{\text{other}} \\
&\quad + P(I(i) = m_k|D(i) = \text{no}ED) \times (1 - p_{\text{other}}).
\end{aligned}
$$

Next, we have that

$$
\begin{aligned}
P(\text{Data}|OD = d) = \sum_{f,y} \prod_{i=0}^{T-1} P(\text{Data}(i)|O = d, F = f, Y = y) \\
P(F = f)P(Y = y).
\end{aligned} \tag{6}
$$

The first term in the product on the right in Eq. 6 is given by

$$
\begin{aligned}
&P(\text{Data}(i)|O = d, F = f, Y = y) = \\
&\quad \prod_k (P(I(i) = m_k|O = d, F = f, Y = y)^{C_k(i)}),
\end{aligned} \tag{7}
$$

where $C_k(i)$ is the number of individuals i days ago with chief complaint m_k.

The term in the product on the right in Eq. 7 above is computed as follows:

$$
\begin{aligned}
&P(I(i) = m_k|O = d, F = f, Y = y) \\
&\quad = \sum_c P(I(i) = m_k|D(i) = c)P(D(i) = c|O = d, F = f, Y = y) \\
&\quad = P(I(i) = m_k|D(i) = d)P(D(i) = d|O = d, F = f, Y = y) \\
&\qquad + P(I(i) = m_k|D(i) = \text{other})P(D(i) = \text{other}|O = d, F = f, Y = y) \\
&\qquad + P(I(i) = m_k|D(i) = \text{no}ED)P(D(i) = \text{no}ED|O = d, F = f, Y = y).
\end{aligned}
$$

The conditional probabilities of values of $D(i)$ in the previous expression are computed using Eqs. 2, 3, and 4.

Using Bayes' Theorem, we have that

$$
P(O = d|\text{Data}) = \frac{P(\text{Data}|O = d)P(O = d)}{\sum_c P(\text{Data}/O = c)P(O = c)}.
$$

The prior probability of an outbreak disease is computed as follows:

$$
\begin{aligned}
P(O = d) &= P(O = d|E = \text{yes})P(E = \text{yes}) + P(O = d|E = \text{no})P(E = \text{no}) \\
&= P(O = d|E = \text{yes})P(E = \text{yes}).
\end{aligned}
$$

Finally,

$$
P(E = \text{yes}|\text{Data}) = \sum_{d \neq \text{none}} P(O = d|\text{Data}).
$$

3.2.2.3 Justification for the independence assumption Recall that the model assumes that, given values of O, F, and Y, the chief complaints of an individual on different days are independent. We will now discuss the justification for this assumption, which serves as an approximation.

First, we assume that the number of days of data T (time window) is fairly small. In our experiments we used $T = 5$. For the sake of simplicity, in the discussion that follows, we use $T = 3$. Also, for the sake of brevity, in what follows we denote the event $O = d, F = f, Y = y$ by e.

Next, note that $P(I = \text{no}ED|e) \approx 1$ for any set of values of $O = d, F = f$, and $Y = y$. This can be seen by looking at Fig. 4. We see from that figure that, for example,

$$P(D = \text{no}ED|O = \text{flu}, F = 0.0000118) = .9979552 \tag{8}$$

$$P(D = \text{no}ED|O = \text{none}, F = 0.0000118) = .997967 \tag{9}$$

$$P(I = \text{no}ED|D = \text{no}ED) = 1. \tag{10}$$

Eqs. 9 and 10 are exemplary of what is true in general. Namely, regardless for realistic values of O and F(and of Yin the temporal model), an individual will most probably not go to the ED. We see from Equation 11 that if an individual does not go to the ED, then the value of I_r is noED.

Now we will discuss the justification for the independence assumption separately for three types of individuals.

1. The individual does not visit the ED during the time window.

Our assumption relevant to these individuals is that if an individual does not go to the ED on one or more days in a row, it is still most probable that the individual will not go to the ED the following day. Let $I(k)$ be the individual's chief complaint k days ago. Due to the chain rule, we then have that

$$\begin{aligned} P(I(0) = \text{no}ED, I(1) = \text{no}ED, I(2) = \text{no}ED|e) \\ = P(I(0) = \text{no}ED|I(1) = \text{no}ED, I(2) = \text{no}ED, e) \\ \times P(I(1) = \text{no}ED|I(2) = \text{no}ED, e) \times P(I(2) = \text{no}ED|e) \\ \approx P(I(0) = \text{no}ED|e) \times P(I(1) = \text{no}ED|e) \times P(I(2) = \text{no}ED|e) \\ \approx 1. \end{aligned}$$

The approximations hold because all terms in the first product are assumed to be near 1, all terms in the second product are near 1, and we have assumed the time window is small. Now

$$P(I(0) = \text{no}ED, I(1) = \text{no}ED, I(2) = \text{no}ED|e)$$

is the actual conditional probability of the data concerning the individual, and

$$P(I(0) = \text{no}ED|e) \times P(I(1) = \text{no}ED|e) \times P(I(2) = \text{no}ED|e)$$

is the value used by model.

The assumption made here is suspect only if not going to the ED several days in a row somehow made it probable an individual would go to the ED the following day, which does not seem reasonable. So the assumption concerning these individuals is reasonable, and this assumption concerns most of the individuals in the population since most individuals do not visit the ED in a short time window (unless there was a very severe outbreak in which case computer-assisted outbreak detection would probably not be needed).

- The individual visits the ED once during the time window.

Our assumption relevant to these individuals is that if an individual went to the ED 2 days ago, the probability of not going to the ED 1 day ago remains very high. Furthermore, if the individual went to the ED 2 days ago, and did not go to the ED 1 day ago, the probability of not going to the ED today remains high. In general, the assumption is that if an individual went to the ED i days ago and did not go the ED $i-1, i-2, \ldots,$ and $j+1$ days ago, then the probability of not going to the ED j days ago remains high, where $j < i$ and j and i are both in the window. This assumption seems reasonable. If an individual goes to the ED 1 day, one might argue that it would increase the probability of going to the ED a second day because the individual is sick. Or one might argue that it would decrease the probability of going to the ED another day because the individual has already been to ED. Our assumption would only be incorrect if an ED visit on 1 day made the probability of an ED visit another day substantially different. So this assumption seems reasonable, but perhaps not as compelling as the assumption for Type 1 individuals, which was discussed above.

Without loss of generality, assume that the individual's sole ED visit is 2 days ago and that the chief complaint is m_k. Given the assumption above, due to the chain rule we then have that

$$\begin{aligned} P(I(0) = \text{no}ED, I(1) = \text{no}ED, I(2) = m_k|e) \\ = P(I(0) = \text{no}ED|I(1) = \text{no}ED, I(2) = m_k, e) \\ \times P(I(1)|I(2) = m_k, e) \times P(I(2) = m_k|e) \\ \approx P(I(2) = m_k|e) \\ \approx P(I(0) = \text{no}ED|e) \times P(I(1) = \text{no}ED|e) \times P(I(2) = m_k|e). \end{aligned}$$

- The individual visits the ED more than once during the time window.

Assuming conditional independence means that the naive Bayes assumption is being made. Although this assumption is often not literally true, it often has been shown to perform well in practice on classification tasks (Sun and Shenoy 2007). This seems to be the least compelling of our assumptions. However, there should be very few individuals of this type.

It is believed that no outbreak occurred in Allegheny County during the calendar year 2004. Using ED data from that county during that calendar year, Jiang (2008) evaluated the accuracy of the assumptions just presented. The results indicated that the assumption for Type 1 individuals was accurate to the third decimal place, the assumption for Type 2 individuals was accurate to the fifth decimal place, but the assumption

for Type 3 individuals was not very accurate. However, the data also indicated that few individuals are of Type 3, most are of Type 1, and an intermediate number are of Type 2.

3.2.3 The type of temporal modeling in PCT

This paper focuses on modeling atemporal information about individual patients, as patients arrive in the ED over time. This focus is reasonable, since the type of information that is available electronically in real time about patients is usually a "snapshot" of the patient at the time of the ED visit (e.g., the patient's age, gender, and chief complaint), rather than information about the patient over time (e.g., how the patient's temperature varies over time). Thus, the evidence about a given patient is modeled as atemporal, even while we have temporal information about when those individuals visited the ED. Moreover, the goal of the research reported in this paper is to detect the presence of an outbreak at any given time, rather than to model the dynamics of an outbreak over time. Thus, the outbreak node (O) is only temporal in the sense that it represents the current time, which is continually changing.

4 Experiments

In Experiment 1, we ran PCT and PC with ED data collected during a real influenza outbreak and compared the results. In Experiment 2, we created a set of simulated influenza outbreaks and a set of simulated *Cryptosporidium* outbreaks by injecting outbreak cases in real ED data. This type of simulated outbreak is called a semi-synthetic outbreak. We then compared the results obtained by monitoring these semi-synthetic outbreaks using PCT, PC, CUSUM, and EWMA. This section describes the two experiments and how we evaluated the results from these experiments.

4.1 Experiment 1

4.1.1 Method

In this experiment, we evaluated the ability of PCT to detect a laboratory validated outbreak of influenza in Allegheny County based on chief-complaint data from EDs in that county. We compared PCT's detection performance to that of PC. We used a value of $T = 5$ as the number of days of data considered by PCT. Jiang (2007) estimated that the outbreak started on 11/18/2003 and lasted for 66 days. Allegheny County Health Department reported the first confirmed influenza case on 11/18/2003 (www.county.allegheny.pa.us/news/2003/231118.asp). The start of the outbreak was subtle and based largely on the first officially reported influenza case.

4.1.2 Results

Figure 8a and b show the probabilities of an outbreak determined, respectively by PC and PCT starting on 11/7/2003. On 11/29/2003 PC reported that the probability

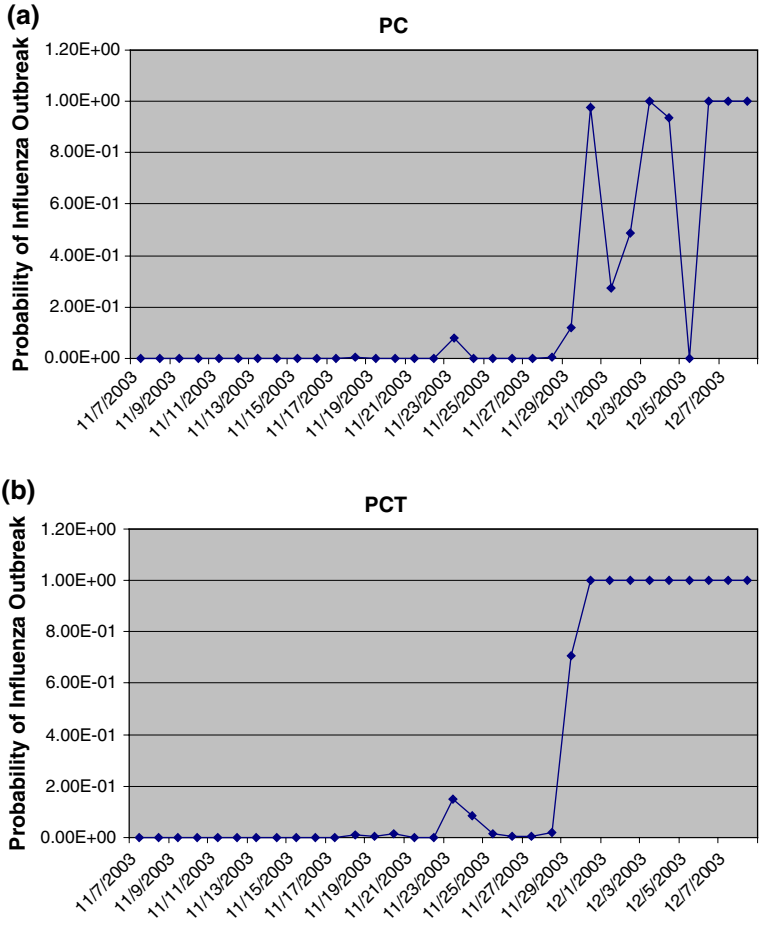

Fig. 8 A comparison of the performance of PC and PCT when detecting a real influenza outbreak. (**a**) PC's posterior probability of influenza outbreak (O = influenza) as a function of date between November 7 and December 8, 2003. (**b**) PCT's posterior probability of influenza outbreak (O = influenza) as a function of date between November 7 and December 8, 2003

of an outbreak was 0.12, and on 11/30/2003 PC reported that this probability was 0.98. However, on 12/1/2003 the probability reported by PC dropped down to 0.27. After that the probabilities fluctuated, reaching a value of almost zero on 12/5/2003. Finally, on 12/6/2003 the probabilities stabilized close to a value of one. On the other hand, on 11/29/2003 PCT reported that the probability of an outbreak was 0.71, and on 11/30/2003 PCT reported that this probability was 0.999. Not only did PCT report higher probabilities than PC early in the outbreak (on 11/29/2003 and 11/30/2003), but PCT's probabilities stabilized close to 1.0 6 days earlier than PC's probabilities (11/30/2003 verses 12/6/2003).

Note that on 11/23/30 both PCT and PC showed a slight spike in the probability of an outbreak, with PC reporting a probability of 0.079 and PCT reporting 0.151. The next day PC's probability dropped back down to zero, whereas PCT's probability was

at 0.08366. So very early in the outbreak PCT also provided a stronger warning of an outbreak than PC.

4.2 Experiment 2

This experiment involved simulated outbreaks that use semi-synthetic data based on ED data obtained from Allegheny County, Pennsylvania. First we discuss the data sets and the method used to evaluate the systems; then we show the results.

4.2.1 The data sets

We used real ED admission data that we collected from Allegheny County, Pennsylvania in the year 2004 as the background data. This data set contains all 110 zip codes in Allegheny County. The average daily number of ED visits included in this date set is about 580. We added simulated outbreak cases to this background data to create semi-synthetic outbreaks. These outbreaks were semi-synthetic because the background data is real and the overlaid outbreak data is synthetic. In all the experiments, influenza and *Cryptosporidium* outbreak cases were simulated because outbreaks of these types have been well studied (Stirling et al. 2001; Cooper et al. 2007; Jiang 2007). The observed data for both types of outbreaks consisted of chief complaints presented by patients in the ED. We simulated a total of 120 outbreaks for each type of outbreak.

Allegheny County, which covers 730 square miles, was modeled using a 16×16 grid. Each grid element is one cell. A zip code was considered entirely within a cell if the zip code's centroid was in the cell. To use a variety of background regions for the outbreaks and to simulate the way outbreaks ordinarily initiate, outbreaks cases were simulated in rectangular subregions of that county. Equal number of outbreaks that occur in rectangles that are 2 cells by 1 cell, 2 cells by 2 cells, and 3 cells by 2 cells were developed. The 2 by 1 rectangles and 3 by 2 rectangles could go either north-south or east-west.

To control the severity of the outbreak, we determined the number of daily injected cases based upon the standard deviation σ_{cell} of the number of real background daily ED visits in each cell in the injected subregion. We simulated the same number of outbreaks for each of four levels of severity. The average (over the entire simulation) number of injected cases in a cell for severity levels 1, 2, 3, and 4 were, respectively $1.5\sigma_{\text{cell}}$, $2\sigma_{\text{cell}}$, $2.5\sigma_{\text{cell}}$, and $3\sigma_{\text{cell}}$.

To test the robustness of the assumption in the model that the increase in the number of outbreak cases is linear, we simulated outbreaks in which the number of injected cases were made to increase according to linear, quadratic, and cubic functions before the outbreak reached its peak. There were an equal number of outbreaks with each type of increase. To simulate an outbreak that, for example, methodically exhibited a linear increase in outbreak cases, we would assume that Δ of them occur on day one of the outbreak, 2Δ occur on day two, and so on. The value of Δ can therefore be determined by solving

$$\Delta + 2\Delta + \cdots + \frac{30}{2}\Delta = \frac{\text{tot}_{\text{cell}}}{2},$$

where tot_{cell} is the total number of cases injected during the outbreak. The solution is $\Delta = \text{tot}_{\text{cell}}/240$.

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). Rather, in half the simulations on even numbered days we made the number of new cases 25% of the previous day's count, and in the remaining simulations on even numbered days we made it 50% of the previous day's count. We imposed daily fluctuations so we could evaluate the detection maintenance capability of the systems. If PC first detected an outbreak on day t when the number of injections was, for example, 100, it is likely that it would not detect it on day $t + 1$ if the number of injections was only 50. However, since PCT would be looking at the data from both day t and day $t + 1$ it seems likely that it would maintain the detection signal on day $t + 1$.

To determine the chief complaint of each injected case, the chief complaint was generated stochastically according to a probability distribution Q of the chief complaints given the outbreak disease (influenza or *Cryptosporidium*). Recall PCT contains a probability distribution of the chief complaints given the outbreak disease, and it is the same as the probability distribution P in PC. To test the robustness of the systems, Q was allowed to vary significantly from P. To obtain a conditional probability distribution Q_i, we let the probabilities in P be the means of a Dirichlet distribution. We stochastically generated ten different probability distributions Q_1 through Q_{10} according to the Dirichlet distribution. Table 1 shows, in the case of influenza, the probabilities in P, which were the means in our Dirichlet distribution, the standard deviations in that Dirichlet distribution, and the sample probability distributions Q_1 and Q_2.

Figure 9 shows the daily ED visit counts (both real background and injected outbreak cases) for one particular influenza outbreak that showed a linear increase. The actual zip codes in which outbreak cases were injected are 15084, 15014, and 15056.

4.2.2 *Evaluation methodology*

AMOC curves (Fawcett and Provost 1999) were used to evaluate the ability of the systems to detect the outbreaks. In such curves, the annual number of false alarms is plotted on the x-axis and the mean days to detection is plotted on the y-axis.

Table 1 The probabilities in P, which were the means in our Dirichlet distribution, the standard deviations in that Dirichlet distribution, and the sample probability distributions Q_1 and Q_2

Chief complaint	P(means)	Standard deviations	Q_1	Q_2
Cough	0.343	0.194	0.175	0.127
Diarrhea	0.025	0.064	0	0
Dypsnea	0.099	0.122	0.018	0.100
Fatigue/weakness	0.026	0.065	0.023	0.053
Fever/chills	0.421	0.202	0.678	0.617
Malaise	0.011	0.042	0	0
Myalgia	0.010	0.040	0	0
Nausea/vomiting	0.041	0.081	0.080	0.103
Sore throat	0.024	0.062	0.026	0

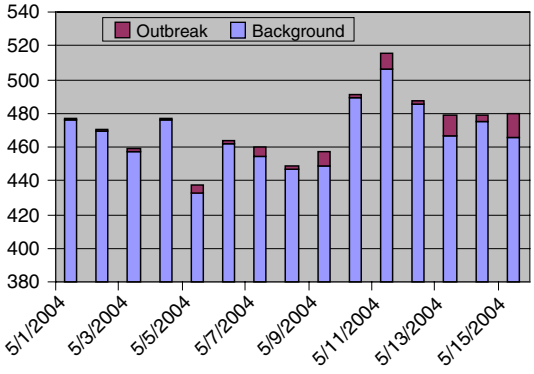

Fig. 9 The ED visit counts for one particular semi-synthetic influenza outbreak

The performances of PC and PCT were further compared using traditional significance testing, which is equivalent to computing the probability that one system's average time to detection is greater than that of another systems under the assumption of prior ignorance. We will now discuss the methodology we used to determine statistical significance.

Suppose we want to compare two systems, $System_1$ and $System_2$, which detect the same set of outbreaks. For a given false alarm rate r, we can analyze the significance of the results using a paired observation t-test. For false alarm rate r, we let $\mu_1^{(r)}$ be the mean time to detection for $System_1$, $\mu_2^{(r)}$ be the mean time to detection for $System_2$, and $\mu^{(r)} = \mu_1^{(r)} - \mu_2^{(r)}$. We are interesting in rejecting the null hypothesis that $System_1$ has a smaller mean time to detection than $System_2$. Therefore, we want to see if we can reject that $\mu_1^{(r)} \leq \mu_2^{(r)}$ in favor of $\mu_1^{(r)} > \mu_2^{(r)}$. Our hypotheses are therefore $H_0^{(r)}$:$\mu^{(r)} \leq 0$ and $H_A^{(r)}$:$\mu^{(r)} > 0$. Using the paired observation t-test, we compute the p-value $p^{(r)}$ of the result.

If we do a Bayesian analysis and assume prior ignorance as to the value of $\mu^{(r)}$, Jiang (2008) shows that

$$P(H_A^{(r)}|\text{Data}) = P(\mu_1^{(r)} > \mu_2^{(r)}|\text{Data}) = 1 - p^{(r)}.$$

where $p^{(r)}$ is the p-value obtained using the t-test.

We are not only interested in how early a system can detect an outbreak, but also in how early the system maintains the detection of an outbreak. AMOC-M curves were used to evaluate the latter. An *AMOC-M curve (AMOC-Maintain curve)* [33] is like an AMOC curve, except that the y-axis plots the average time at which an outbreak signal is detected and maintained thereafter. For example, if the threshold is 0.04, and the sequence of signals is [0.01, 0.02, 0.05, 0.03, 0.04, 0.02, **0.05**, 0.06, 0.05, 0.07], then the time at which the signal is maintained above the threshold is 7 because on the 7th day the probability is 0.05, which exceeds 0.04, and it stays at or above 0.04 for the remaining days in the analysis.

4.2.3 Results

Results of the experiments are shown next.

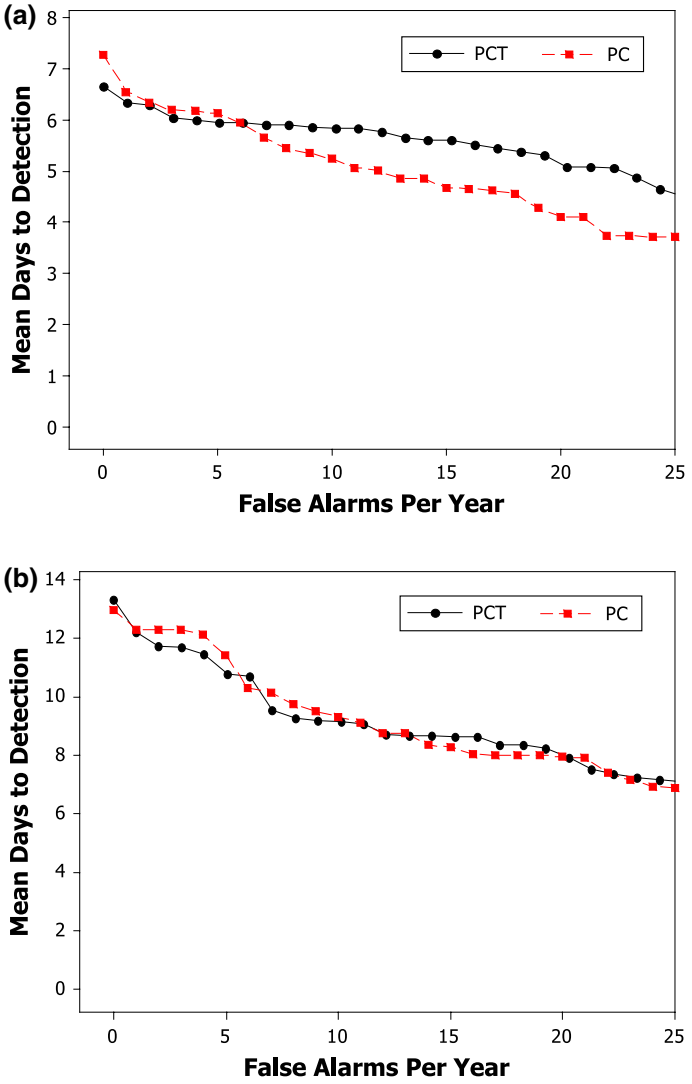

Fig. 10 AMOC *curves* comparing the detection performance of PCT and PC. (**a**) Cryptosporidium outbreaks. (**b**) Flu outbeaks

4.2.3.1 Results of comparing PC to PCT Figure 10 shows AMOC curves comparing the detection performance of PCT and PC. Table 2 shows the posterior probability that PCT has a smaller mean day to detection than PC at various false alarm rates (r). In that table and in Table 3 the following notation is used:

- PCc: PC detecting *Cryptosporidium* outbreaks.
- $PCTc$: PCT detecting *Cryptosporidium* outbreaks.
- PCf: PC detecting influenza outbreaks.
- $PCTf$: PCT detecting influenza outbreaks.

Table 2 At various false alarm rates (r), the posterior probability that PCT has a smaller mean day to detection than PC

r	$P(\mu_{\text{PCc}} > \mu_{\text{PCTc}})$	$P(\mu_{\text{PCf}} > \mu_{\text{PCTf}})$
0	0.9038	0.1291
5	0.6791	0.9132
10	0.0877	0.9087
15	0.0198	0.5630

Table 3 At various false alarm rates (r), the posterior probability that PCT has a smaller mean day to maintaining detection than PC

r	$P(\nu_{\text{PCc}} > \nu_{\text{PCTc}})$	$P(\nu_{\text{PCf}} > \nu_{\text{PCTf}})$
0	>0.9999	0.9990
5	>0.9999	>0.9999
10	>0.9999	>0.9999
15	>0.9999	>0.9999

For the sake of concreteness, let us define a small annual false alarm rate to be five false alarms or less. We see from Fig. 10 and Table 2 that in the case of *Cryptosporidium* outbreaks PCT performed better for small false alarm rates, but worse for large false alarm rates. In the case of influenza outbreaks PCT and PC performed about the same.

Note that our results indicate that PC performs better at large false alarms rates in the case of *Cryptosporidium* outbreaks. This result may be due to the following. If we set our threshold low (and thereby have a large number of false alarms), we should often be able to detect an outbreak when there is an initial small spike on an early day of the outbreak if we look only at that day's data. However, a system that looks at that day's data along with data from previous days might miss the 1-day spike.

Figure 11 shows AMOC-M curves comparing the detection maintenance performance of PCT and PC. For both *Cryptosporidium* and influenza outbreaks, the performance of PCT is superior to that of PC for all false alarm rates.

Table 3 shows the posterior probability that PCT has a smaller mean day to maintaining detection than PC at various false alarm rates (r). These results strongly support our hypothesis that PCT is more stable than PC in that once an outbreak is detected, PCT is better at maintaining the detection signal on future days.

4.2.3.2 Results of comparing PCT to CUSUM and EWMA We mentioned in Sect. 2.2 that by modeling each individual in the population, we can base our analysis on more information than that contained in a summary statistic such as the number of patients who visited the ED with respiratory symptoms on a given day. By so doing we may be able to obtain better detection performance. To test this conjecture we compared the performance of PCT to the classic temporal methods CUSUM and EWMA (see Sect. 2.1.2). We configured CUSUM and EWMA to attempt to detect the outbreak disease being simulated (either influenza or *Cryptosporidium*) by using the counts of the three chief complaints that were the best indicators of the outbreak disease (according to the probability distributions in PC).

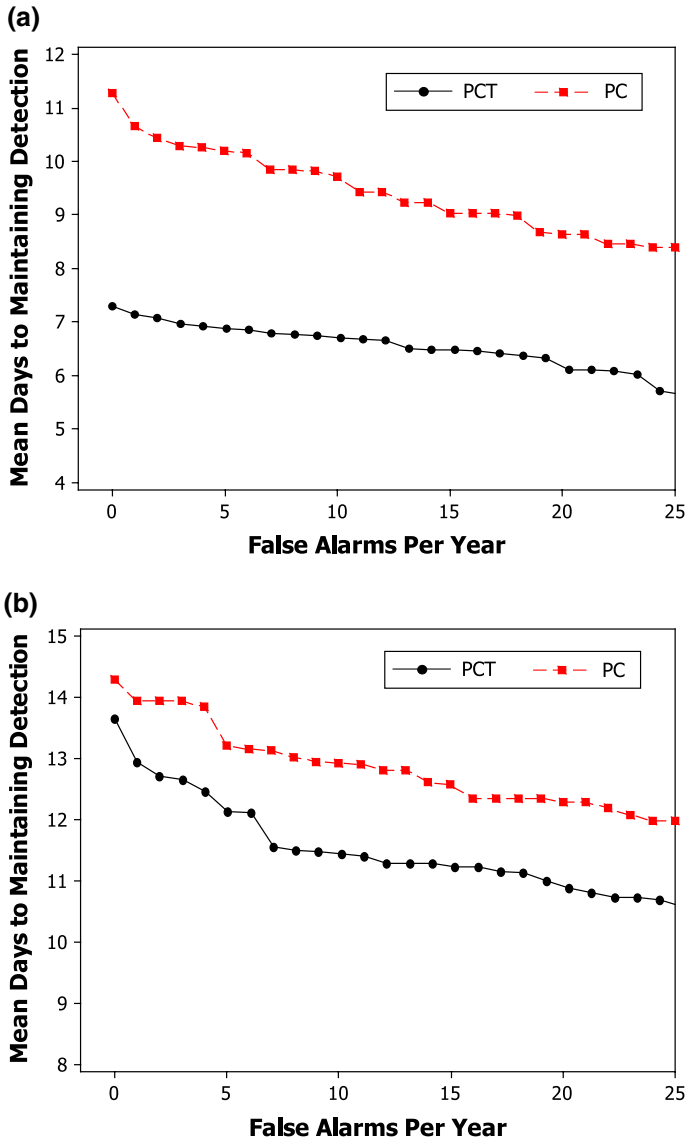

Fig. 11 AMOC-M *curves* comparing the detection maintenance performance of PCT and PC. (**a**) Cryptosporidium outbreaks. (**b**) Flu outbreaks

Figure 12 shows AMOC curves comparing the detection performance of PCT, CUSUM, and EWMA, and Fig. 13 shows AMOC-M curves comparing the detection maintenance performance of PCT, CUSUM, and EWMA. Table 4 shows the posterior probability that PCT has a smaller mean day to detection than CUSUM at various false alarm rates (r), and Table 5 shows the posterior probability that PCT has a smaller mean day to maintaining detection than CUSUM at various false alarm rates (r), Table 6 shows the posterior probability that PCT has a smaller mean day to detection than

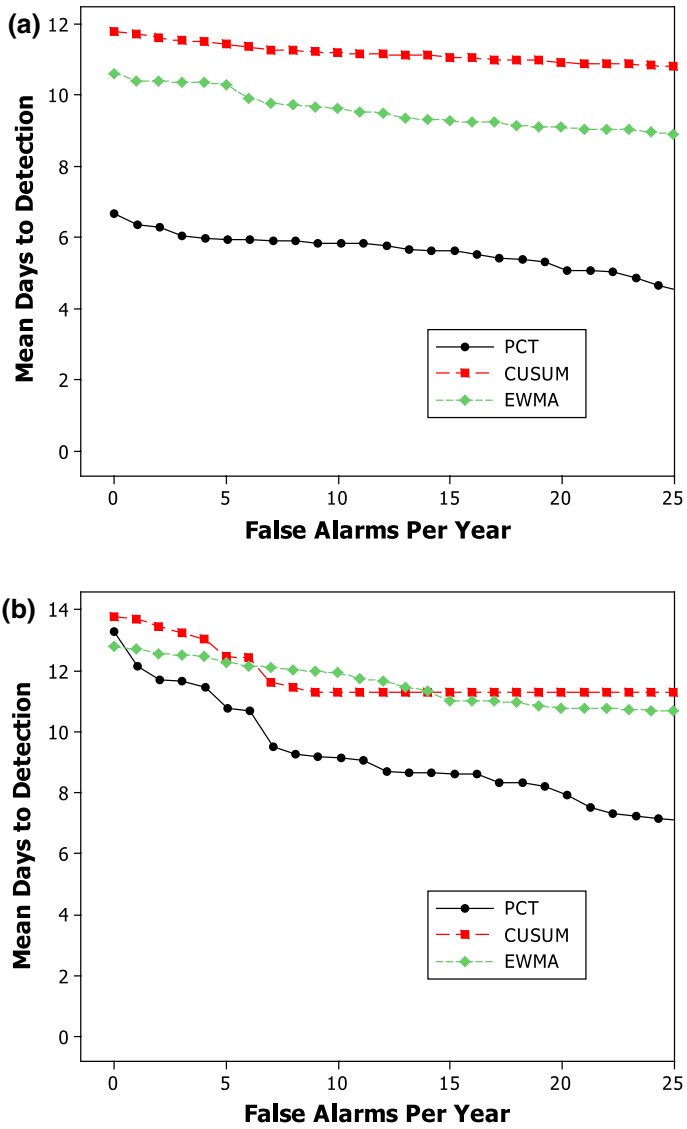

Fig. 12 AMOC *curves* comparing the detection performance of PCT, CUSUM, and EWMA. (**a**) Cryptosporidium outbreaks. (**b**) Flu outbreaks

EWMA at various false alarm rates (r), and Table 7 shows the posterior probability that PCT has a smaller mean day to maintaining detection than EWMA at various false alarm rates (r).

We see that for both types of outbreaks PCT performed substantially better than CUSUM both at initially detecting outbreaks and at maintaining detection. For *Cryptosporidium* outbreaks PCT performed much better than EWMA at both outbreak detection and detection maintenance, whereas for influenza outbreaks PCT performed

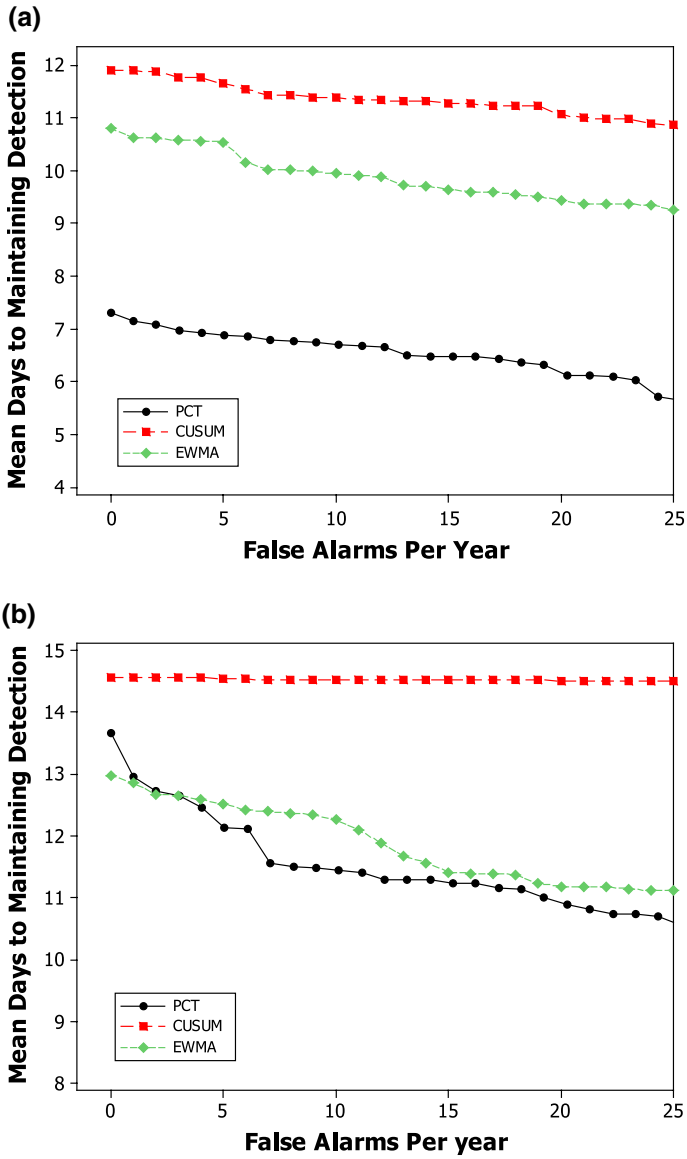

Fig. 13 AMOC-M *curves* comparing the detection maintenance performance of PCT, CUSUM, and EWMA. (**a**) Cryptosporidium outbreaks. (**b**) Flu outbreaks

much better than EWMA at outbreak detection but only moderately better at detection maintenance.

5 Discussion

This paper introduced a Bayesian network architecture called BayesNet-T for developing temporal event surveillance systems. Using this architecture, we extended the

Table 4 At various false alarm rates (r), the posterior probability that PCT has a smaller mean day to detection than CUSUM

r	$P(\mu_{\text{CUSUMc}} > \mu_{\text{PCTc}})$	$P(\mu_{\text{CUSUMf}} > \mu_{\text{PCTf}})$
0	>0.9999	0.9312
5	>0.9999	0.9751
10	>0.9999	>0.9999
15	>0.9999	>0.9999

Table 5 At various false alarm rates (r), the posterior probability that PCT has a smaller mean day to maintaining detection than CUSUM

r	$P(\nu_{\text{CUSUMc}} > \nu_{\text{PCTc}})$	$P(\nu_{\text{CUSUMf}} > \nu_{\text{PCTf}})$
0	>0.9999	>0.9999
5	>0.9999	>0.9999
10	>0.9999	>0.9999
15	>0.9999	>0.9999

Table 6 At various false alarm rates (r), the posterior probability that PCT has a smaller mean day to detection than EWMA

r	$P(\mu_{\text{EWMAc}} > \mu_{\text{PCTc}})$	$P(\mu_{\text{EWMAf}} > \mu_{\text{PCTf}})$
0	>0.9999	0.1011
5	>0.9999	>0.9999
10	>0.9999	>0.9999
15	>0.9999	>0.9999

Table 7 At various false alarm rates (r), the posterior probability that PCT has a smaller mean day to maintaining detection than EWMA

r	$P(\nu_{\text{EWMAc}} > \nu_{\text{PCTc}})$	$P(\nu_{\text{EWMAf}} > \nu_{\text{PCTf}})$
0	>0.9999	0.0455
5	>0.9999	0.8275
10	>0.9999	0.9856
15	>0.9999	0.6727

non-temporal outbreak detection system PC to the temporal outbreak detection system PCT. We hypothesized that (1) PCT will have a smaller mean time to detection than PC at low false alarm rates; and (2) PCT will be more stable than PC in that once an outbreak is detected, PCT will be better at maintaining the detection signal on future days. Results concerning both a real influenza outbreak and simulated outbreaks using semi-synthetic data support hypothesis 2 strongly and hypothesis 1 modestly.

PCT is a patient-specific system that models each individual in the population. We hypothesized that such as system might obtain better detection performance than classic time-series systems that uses a summary statistic such as daily counts. Results of comparing PCT to CUSUM and EWMA served to support this hypothesis. Another advantage of PCT over these methods is that PCT can be readily extended to include other information about individuals other than ED data.

References

Baron MI (2002) Bayes and asymptotically pointwise stopping rules for the detection of influenza outbreaks. In: Gastonis C, Kass RE, Carriquiry A et al (eds) Case studies in Bayesian statistics. Springer–Verlag, New York

Bos T, Fetherston TA (1992) Market model nonstationarity in the Korean stock market. In: Rhee SG, Chang RP (eds) Pacific-Basin capital markets research, 3rd edn. Elsevier, North-Holland, Amsterdam

Box G, Jenkins G, Reinsel R (1994) Time series analysis: forecasting and control. Prentice Hall, Englewood Cliffs

Burges C (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167

Cooper GF, Dash DH, Levander JD, Wong WK, Hogan WR, Wagner MM (2004) Bayesian biosurveillance of disease outbreaks. In: Proceedings of the 20th conference on uncertainty in artificial intelligence, AUAI Press, Arlington, Virginia, pp 94–103

Cooper GF, Dowling JN, Lavender JD, Sutovsky P (2007) A Bayesian algorithm for detecting CDC category A outbreak diseases from emergency department chief complaints. Adv Disease Surveill 2:45

Fawcett T, Provost F (1999) Activity monitoring: noticing interesting changes in behavior. In: Proceedings of the fifth SIGKDD conference on knowledge discovery and data mining, ACM Press, San Diego, California, pp 53–62

Hamilton J (1994) Time series analysis. Princeton University Press, Princeton

Hogan W, Cooper GF, Wallstrom G, Wagner M (2007) The Bayesian aerosol release detector: an algorithm for detecting and characterizing outbreaks caused by an atmospheric release of bacillus anthracis. Stat Med 26(29):5225–5252

Jiang X (2007) A Bayesian network for predicting an epicurve. Adv Disease Surveill 2:15

Jiang X (2008) A Bayesian network model for spatio-temporal event surveillance, Ph.D. Thesis, Department of Biomedical Informatics, University of Pittsburgh

Jiang X, Cooper GF, Neill DB (2009) A Bayesian network model for spatial event surveillance. Int J Approx Reason. doi:10.1016/j.ijar.2009.01.001

Jiang X, Wallstrom GL (2006) A Bayesian network for outbreak detection and prediction. In: Proceedings of AAAI-06, Boston, Massachusetts, pp 1166–1160

Kulldorff M (1997) A spatial scan statistic. Commun Stat Theory Methods 26(6):1481–1496

Kulldorff M (2004) Satscan v. 4.0: software for the spatial and space-time scan statistics, Technical Report, Information Management Services, Inc.

Kulldorff M, Heffernan R, Hartman J, Assunco R, Mostashari F (2005) Space-time permutation scan statistic for disease outbreak detection. PLoS Med 2:216–224

Kulldorff M, Mostashari F, Luiz D, Yih K, Kleinman K, Platt R (2007) Multivariate scan statistics for disease surveillance. Stat Med 26:1824–1833

Montgomery DC (2001) Introduction to statistical quality control. Wiley, New York

Moore A (2001a) A powerpoint tutorial on hidden Markov models, available at www.cs.cmu.edu/~awm/781/timetable.html

Moore A (2001b) A powerpoint tutorial on support vector machines, available at www.cs.cmu.edu/~awm/781/timetable.html

Moore A, Anderson B, Das K, Wong WK (2006) Combining multiple signals for biosurveillance. In: Wagner M (ed) Handbook of biosurveillance. Elsevier, New York

Neill DB, Moore AW, Cooper GF (2005a) A Bayesian spatial scan statistic. Adv Neural Inform Process Syst (NIPS) 18:1003–1010

Neill DB, Moore AW, Sabnani M, Daniel K (2005b) Detection of emerging space-time clusters. In: Proceedings of 11th ACM SIGKDD international conference on knowledge discovery and mining, Chicago, Illinois, pp 218–227

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286

Reis BR, Mandl KD (2003) Time series modeling for syndromic surveillance. BMC Med Inform Dec Making 3(2)

Reis BY, Pagano M, Mandl KD (2003) Using temporal context to improve biosurveillance. PNAS 100(4):1961–1965

Serfling RE (1963) Methods for current statistical analysis of pneumonia-influenza deaths. Public Health Rep 78(6):494–506

Shmueli G, Fienberg S (2006) Current and potential statistical methods for monitoring multiple data streams for biosurveillance. In: Wilson A, Wilson GD, Olwell D (eds) Statistical methods in counterterrorism. Springer, New York

Stirling R, Aramini J, Ellis A, Gillien L, Meyers R, Flevry M, Werker D (2001) Waterborne cryptosporidiosis outbreak, North Battleford, Saskatchewan, spring 2001. Can Commun Disease Rep 27(22):185–192

Soneson C, Bock D (2003) A review and discussion of prospective statistical surveillance in public health. JR Stat Soc A 166(1):5–21

Sun L, Shenoy P (2007) Using Bayesian networks for bankruptcy prediction: some methodological issues. Eur J Oper Res 180(2):738–753

Tsui FC, Wagner MM, Dato V, Chang HC (2001) Value of ICD-9-coded chief complaints for detection of epidemics. Symp J Am Med Inform Assoc 9:4–47

Wong WK, Moore A (2006) Classical time series methods for biosurveillance. In: Wagner M (ed) Handbook of biosurveillance. Elsevier, New York