Journal of Biomedical Informatics 42 (2009) 90–99

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Bayesian prediction of an epidemic curve

Xia Jiang *, Garrick Wallstrom, Gregory F. Cooper, Michael M. Wagner

Department of Biomedical Informatics, University of Pittsburgh, Parkvale Building, M-183, 200 Meyran Avenue, Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 1 February 2008
Available online 13 June 2008

Keywords:
Biosurveillance
Disease surveillance
Epidemic curve
Bayesian network

ABSTRACT

An epidemic curve is a graph in which the number of new cases of an outbreak disease is plotted against time. Epidemic curves are ordinarily constructed after the disease outbreak is over. However, a good estimate of the epidemic curve early in an outbreak would be invaluable to health care officials. Currently, techniques for predicting the severity of an outbreak are very limited. As far as predicting the number of future cases, ordinarily epidemiologists simply make an educated guess as to how many people might become affected. We develop a model for estimating an epidemic curve early in an outbreak, and we show results of experiments testing its accuracy.

1. Introduction

An *epidemic* is a term used in epidemiology that refers to the appearance of new cases of a particular disease in a given human population, during a given time period, at a rate that substantially exceeds the expected number based on recent experience [1]. An epidemic may affect a region, a country, or even a group of countries. If an entire continent or the entire globe is affected, we ordinarily call the occurrence a *pandemic*. A disease outbreak refers to the occurrence of cases of a disease in excess of what would normally be expected in a particular community or geographical area. We shall use the terms disease outbreak and epidemic interchangeably.

An *epidemic curve* is a graph in which the number of new cases of an outbreak disease is plotted against time. Usually, the time interval is one day. Epidemic curves are ordinarily constructed after the disease outbreak is over (if they are constructed at all). As an example, consider the epidemic curve in Fig. 1. This curve was constructed (after the outbreak was over) from clinically defined and laboratory-confirmed cases of a food borne *Cryptosporidium* outbreak that occurred on a Washington, DC university campus in fall, 1998. The curve indicates a possible food contamination through a tight clustering of cases in three days.

A good *estimate of the epidemic curve* during an outbreak would be valuable to health care officials. Based on this estimate, they can plan for sufficient resources and supplies to handle disease treatment on a timely basis. When we say we are estimating the epidemic curve during an outbreak, we mean that, on a given day of the outbreak, we are estimating the daily number of new outbreak cases for days that have occurred so far and we are predicting those daily values for days that will occur in the future (if no measures are taken to control the outbreak). We call the collection of these estimates and predictions the *estimate of the epidemic curve*. Estimation of an epidemic curve in real time is quite complex because we need a model of the outbreak (an epidemic model), a model of sickness behavior of individuals, and a model of the surveillance system (any sampling inefficiency, time delays).

At present, methods for doing real-time estimation and prediction of the magnitude of an outbreak are very limited. For the most part, investigators simply do their best to intensify surveillance in an effort to identify all cases so that the observed number of cases is as close to the real number of cases as possible [4]. PANDA [5], PANDA-CDCA [6], and BARD [7] can provide estimates of some outbreak characteristics such as outbreak type, source, and/or route of transmission of the outbreak. However, none of them estimates the epidemic curve. A model that not only detects an outbreak but estimates important characteristics of the outbreak, namely its severity and duration, is developed in [8]. Knowledge of the probable values of these variables should be more useful to public health officials than merely knowing that an outbreak is probable. However, as discussed previously, an estimate of the epidemic curve itself would be better. Another shortcoming of the model developed in [8] is that it was implemented using pharmacy data obtained during the North Battleford, Saskatchewan *Cryptosporidium* outbreak in spring, 2001, and was evaluated using simulations generated by HIFIDE [9], whose simulations are based on pharmacy data obtained during that same outbreak. Although such techniques are often used to evaluate outbreak detection algorithms, the question remains as to what the performance would be if the assumptions used to create the model and the simulations were different.

The current paper addresses these shortcomings. First, the Bayesian network model discussed here estimates the epidemic curve itself. Second, we developed an instance of the model based

* Corresponding author. Fax: +1 630 515 8309.
E-mail address: xij6@pitt.edu (X. Jiang).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.05.013

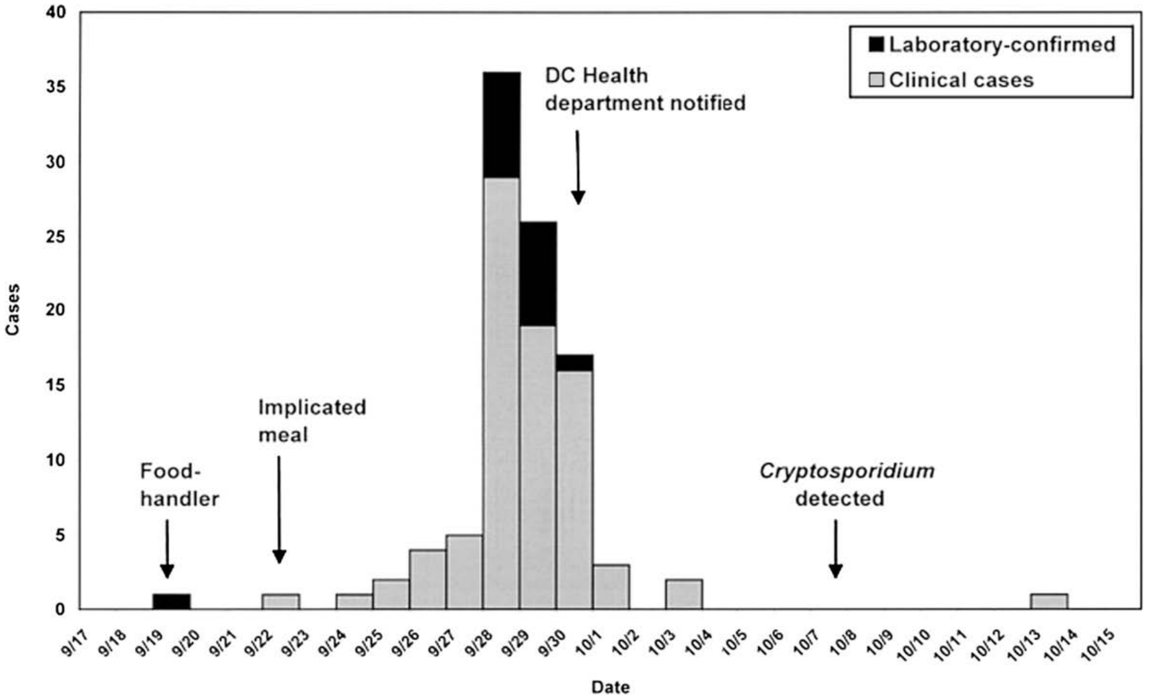


Fig. 1. An epidemic curve for the Washington D.C. *Cryptosporidium* outbreak [2].

on data obtained from a real outbreak, and we evaluated its performance using a different real outbreak. We do this for both *Cryptosporidium* and influenza outbreaks. After discussing the model, we show the results of evaluating its performance.

2. A model for epidemic curve estimation

The epidemic curve for an outbreak is often correlated with the daily counts of some observable event. For example, Fig. 2a 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. Fig. 2b 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 learn something about a *Cryptosporidium* outbreak at an early stage.

The model presented here is applicable to estimating the epidemic curve for certain types of outbreaks from the daily[1] counts of some observable event, whose daily count is positively correlated with the daily count of outbreak cases. For example, it could be used to estimate the epidemic curve for a *Cryptosporidium* outbreak from the daily counts of units of OTC antidiarrheal medicine sold, and it could be used to estimate the epidemic curve for an outbreak of influenza from the daily counts of patients presenting in the ED with respiratory symptoms. We will describe the model in its most general form, without referring to any particular type of outbreak. The model has two distinct components. The first component models the causal relationships among the daily counts and the outbreak severity, outbreak duration, and the number of days into the outbreak. The second component models the relationship between the variables that constitute the epidemic curve and the outbreak severity, outbreak duration, and the number of days into the outbreak. The two components constitute a single Bayesian network. We use that entire network to estimate the epidemic curve from the daily counts. For the sake of clarity, we discuss each component separately. See [10] for an introduction to Bayesian networks.

2.1. Component one

Three important attributes of an outbreak are the following: (1) the severity of the outbreak, which we define as how many individuals eventually becomes ill due to the outbreak; (2) the duration of the outbreak, which is the time from when the infectious entity first appeared in the population until the last day that someone first showed symptoms due to the outbreak; and (3) the number of days since the outbreak began (assuming we are currently in the midst of the outbreak). This component models the causal relationships among these variables and the daily counts of an observable event.

First, we show the network that describes the causal structure. This structure is applicable to outbreaks of different types in different regions. Then we show how the parameter values in the network can be obtained.

2.1.1. Causal structure

The network that describes the causal structure appears in Fig. 3. The network describes the causal relationships among an outbreak of some disease and daily counts of an observable event that is causally related to the outbreak disease. Next we describe the nature of each variable/node in the network.

1) O: Represents the number of days ago since an outbreak began. It is 0 if no outbreak has started.
2) D : Represents the duration of the outbreak. It depends on O because if O is 0, D must be 0.
3) S : The severity of the outbreak if there is an ongoing outbreak. If the number of individuals in the population is known, the severity would be the percent of the population

[1] Although the unit of time is usually one day, it need not be. For example, we could use weekly counts.

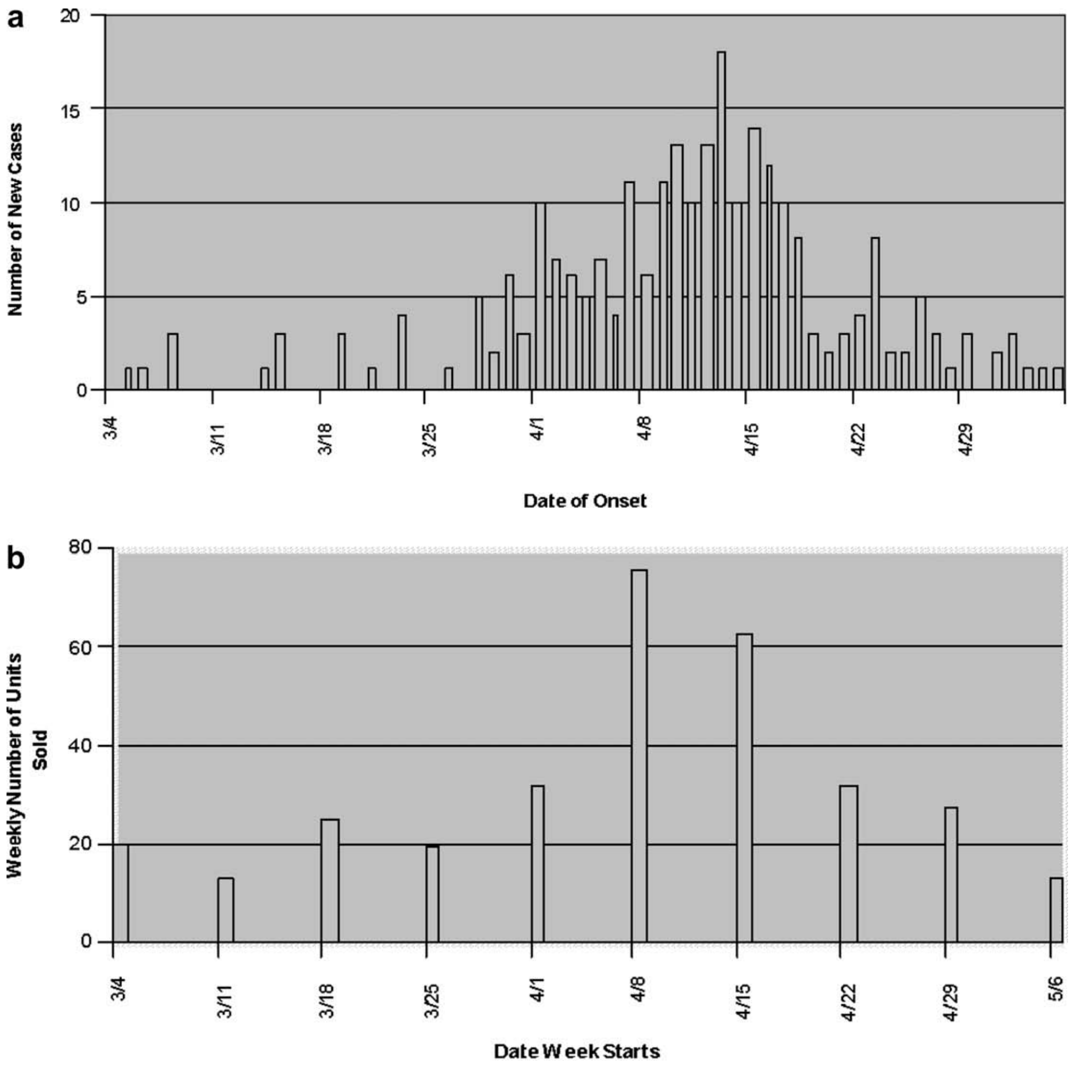


Fig. 2. An epidemic curve for a *Cryptosporidium* outbreak in North Battleford, Saskatchewan is in (a), while weekly OTC sales of antidiarrheal drugs at one pharmacy in North Battleford is in (b). The data for these curves were obtained from [3].

that eventually becomes ill due to the outbreak. However, even if that number is not known, we can use a scale of 0 to 100 to rank the severity. Note that *S* is the severity if no measures are taken to control the outbreak. *S* depends on *O* because if *O* is 0, *S* must be 0.

4) *OC*[$-i$] : The count of the observable event due to individuals who have the outbreak disease. For example, if the observable event is the purchase of one unit of some OTC drug, it would be the count of units of the drug purchased by or for individuals sick with the outbreak disease. OC[0] is the count today, and OC[$-i$] is the count *i* days before today. By today we mean the current day on which we are investigating the outbreak. We can look back as many days as deemed appropriate.
5) *BC*[$-i$] : The count of the observable event due to individuals who do not have the outbreak disease (called background counts). BC[0] is the count today, and BC[$-i$] is the count *i* days before today.
6) *TC*[$-i$] : The total count of the observable event. TC[0] is the count today, and TC[$-i$] is the count *i* days before today. TC[$-i$] is a deterministic function of its parents, OC[$-i$] and BC[$-i$]. That is, it is the sum of these two variables. TC[.] denotes the variables we observe, and which are instantiated in the network when we do inference. For example, if the observable event is the purchase of one unit of some OTC drug, and 100 units were purchased *i* days ago, we would set TC[$-i$] to 100 when we were doing inference.
7) *Day* : Represents the day of the week.
8) *C* : Represents some relevant cyclical variable other than day of the week (e.g. season). There could be more than one such variable, depending on the application.
9) *H* : Represents hidden common causes of the observable event. It mitigates the relationships among daily counts that are not due to the outbreak. This is not the only way to mitigate this relationship. For example, in a given application it may be best to model the background time series with edges between the variables representing daily counts. Cheeseman and Stutz [11] developed AutoClass to learn the range of such hidden variables from data.

For simplicity, we described a network in which there is only one observable event. In general, there could be more than one. If so, for each such event there would be count variables as described above.

2.1.2. Parameter values

The conditional probability distributions of the variables in the network depend on the application. For the sake of concreteness,

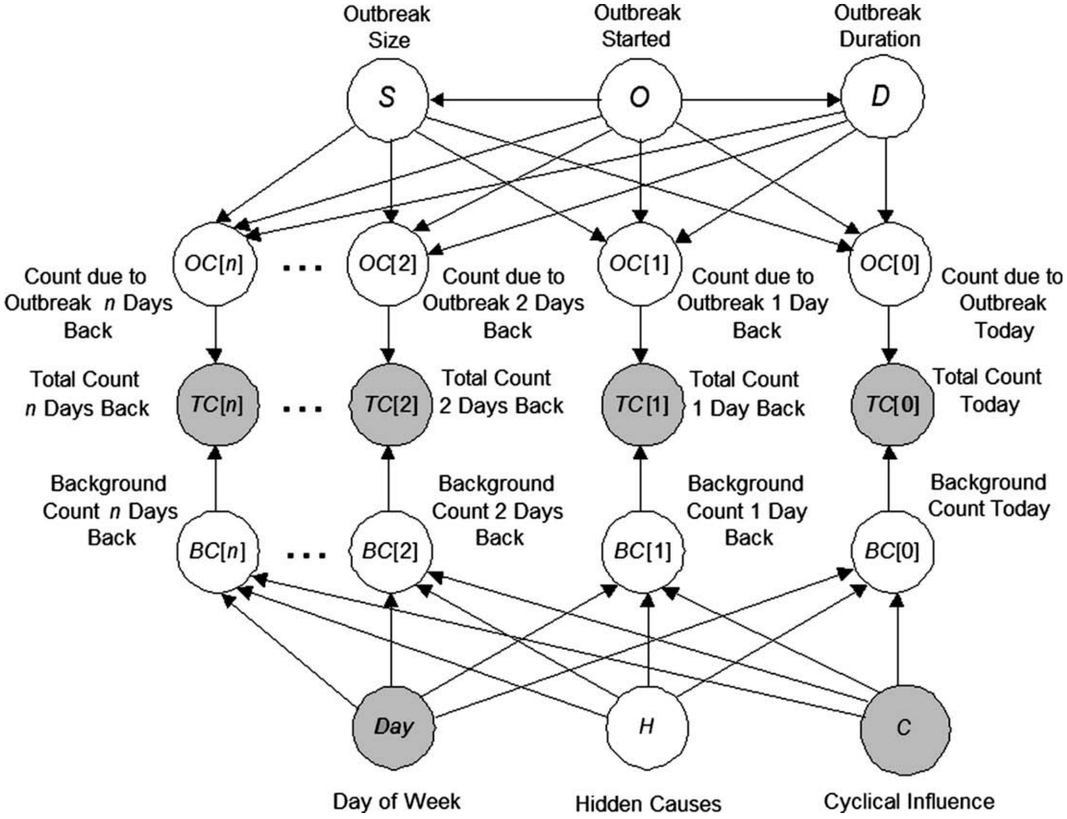


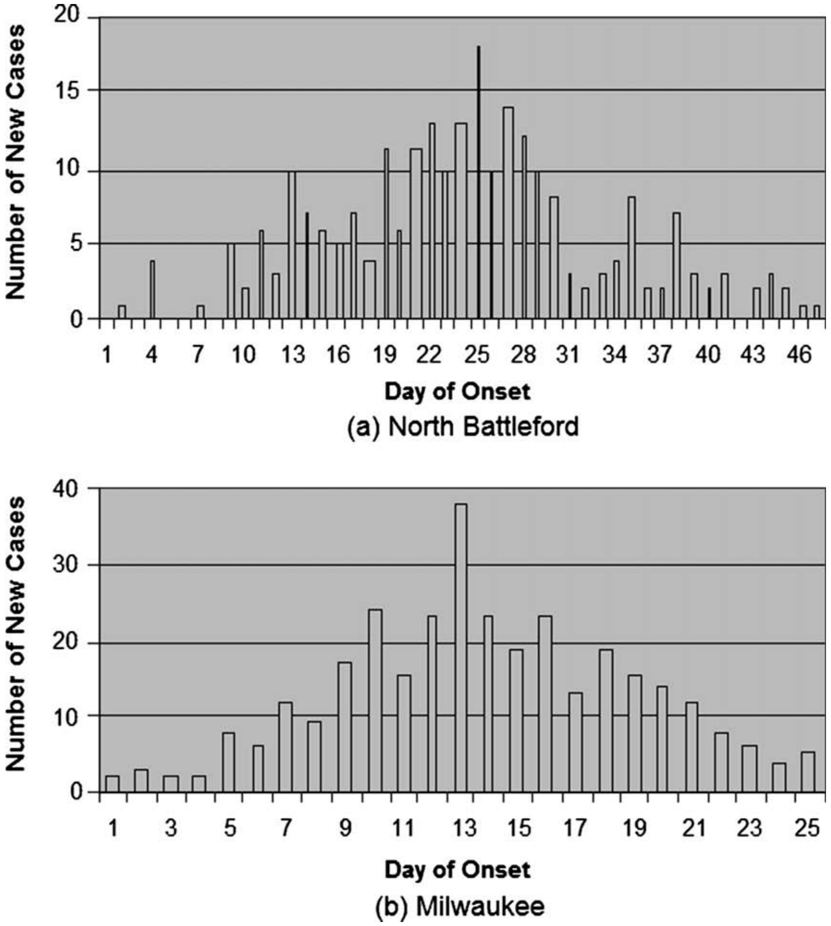
Fig. 3. The component that models the causal relationships among the daily counts and the outbreak severity, outbreak duration, and the number of days into the outbreak.

we give the ones for the *Cryptosporidium* epidemic curve estimation system, whose evaluation is discussed in Section 3.2.1.

1) O: Based on data concerning real outbreaks and subjective judgment, we assumed that there should be a *Cryptosporidium* outbreak about once every 30 years, and that the duration of such outbreaks is uniformly distributed between 21 and 56 days. Since such outbreaks are so rare, for simplicity we assumed there could be at most one in the past 56 days. Therefore, $P(O = 0) = 1\text{–}56/(30 \times 365)$ and $P(O = i) = 1/(30 \times 365)$ for $1 \leqslant i \leqslant 56$.
2) D : If $O = 0$, $P(D = 0) = 1$; otherwise D is uniformly distributed over all integers between 21 and 56.
3) S : If $O = 0$, $P(S = 0) = 1$; otherwise S is uniformly distributed over all integers between 1 and 50.
4) $\mathrm{OC}[-i]$: We assume that the epidemic curves for the outbreaks under consideration are unimodal. That is, the number of outbreaks cases starts at 0, rises to a peak, and then declines back to 0. However, along the way, there can be daily fluctuations. Since we assume the epidemic curves are unimodal, we also assume that the corresponding curves of the counts of the observable event are unimodal. For example, we assume epidemic curves for all *Cryptosporidium* outbreaks are unimodal, and therefore the corresponding curves of the counts of units of OTC antidiarrheal medicine sold are unimodal. Similarly, we assume epidemic curves for all influenza outbreaks are unimodal, and therefore the corresponding curves representing counts of patients presenting in the ED with respiratory symptoms are unimodal. Admittedly, this is a strong assumption. For example, some influenza outbreaks are not unimodal. In Section 4 we discuss possible extensions to the current model which could take this into account. Furthermore, we assume that outbreaks of a given type reach their peak about the same fraction of days into the outbreak. Fig. 4 shows epidemic curves for the North Battleford, Saskatchewan *Cryptosporidium* outbreak in spring 2001 and the Milwaukee, Wisconsin *Cryptosporidium* outbreak in spring 1993. Note that although the outbreaks have different durations, and different peaks, they both reach their peak about half way into the outbreak. Correspondingly, we assume that the curves of the counts of the observable event for outbreaks of the same type reach their peak about the same fraction of days into the outbreak.

Making the assumption just discussed, we can develop a probability distribution for OC[0] given values of O, D, and S as follows. We first choose some actual outbreak $\mathrm{Outbreak}_{\mathrm{OJ}}$ in some jurisdiction OJ, which we call the outbreak jurisdiction. This outbreak should be of the same type as the monitored outbreak. The reason is that, as we shall see, we will base our expectations concerning the monitored outbreak on the structure of this outbreak, and we have only assumed that all outbreaks of the same type reach their peaks about the same fraction of days into the outbreak. We have not assumed that outbreaks of all types reach their peaks about the same fraction of days into the outbreak.

It is necessary that we have daily counts of the observable event both during the outbreak and when the outbreak is not occurring. It is not necessary that we have counts of all occurrences of the observable event in the jurisdiction in which the outbreak took place. For example, if the observable event is the sale of one unit of some OTC drug, we could have counts from only several pharmacies in jurisdiction OJ. Similarly, we assume we know the daily counts of the observable event in some sub-region of the jurisdiction RJ for which we are developing the system, which we call the monitored jurisdiction.



(a) North Battleford

(b) Milwaukee

Fig. 4. An epidemic curve for the North Battleford, Saskatchewan *Cryptosporidium* outbreak in spring, 2001 is in (a), while one for the Milwaukee, Wisconsin *Cryptosporidium* outbreak in spring, 1993 is in (b). The curves were respectively constructed from data obtained from [3] and from [12].

Let

a) T_{OJ} be the sub-region of OJ from which we obtain our counts of the observable event,
b) T_{RJ} be the sub-region of RJ from which we obtain our counts of the observable event,
c) B_{OJ} be the average daily count of the observable event in T_{OJ} when no outbreak is occurring (called the background count),
d) B_{RJ} be the average daily count of the observable event in T_{RJ} when no outbreak is occurring,
e) D_{OJ} be the duration of $\mathrm{Outbreak}_{\mathrm{OJ},}$
f) S_{OJ} be the severity of $\mathrm{Outbreak}_{\mathrm{OJ}}$.

We assume that T_{OJ} is an unbiased representation of OJ with regard to $\mathrm{Outbreak}_{\mathrm{OJ}}$, and that T_{RJ} is an unbiased representation of RJ with regard to any outbreak that may occur in RJ.

Recall that we know the daily counts of the observable event in T_{OJ} during $\mathrm{Outbreak}_{\mathrm{OJ}}$. These counts constitute a bar graph like the one in Fig. 2b. First we subtract B_{OJ} from each day's count to obtain a bar graph that estimates the daily counts due to $\mathrm{Outbreak}_{\mathrm{OJ}}$. We smooth this bar graph to obtain a continuous function $g(t)$ on the interval $[1, D_{\mathrm{OJ}}]$. Then, to create the conditional distribution of OC[0] given values of O, D, and S, we first define the mean of OC[0], given these values, as follows:

$$\mu(O, D, S) = \left(\frac{B_{\mathrm{RJ}}}{B_{\mathrm{OJ}}}\frac{S}{S_{\mathrm{OJ}}}\frac{D_{\mathrm{OJ}}}{D}\right) g((D_{\mathrm{OJ}} - 1)(O - 1)/(\mathrm{D} - 1) + 1). \qquad (1)$$

For given values of S and D, the result is a function of O whose shape is like $g(t)$ but which gives approximate counts of the observable event which would occur if an outbreak with duration D and severity S took place in RJ. Wallstrom [13] offers a formal justification for this. Intuitively, $B_{\mathrm{RJ}}/B_{\mathrm{OJ}}$ scales the function to describe T_{RJ} instead of T_{OJ}, D_{OJ}/D scales the function to describe an outbreak with duration D instead of D_{OJ}, S/S_{OJ} scales the function to describe an outbreak with severity S instead of S_{OJ}, and the expression in the argument of g changes the domain of the function from $[1, D_{\mathrm{OJ}}]$ to $[1, D]$.

Given values of S and D, the function μ is a scaled replica of g. We inserted random fluctuation by letting μ be the mean of a negative binomial distribution. We set the dispersion of that distribution equal to 3.52. This value is consistent with the variances we discovered for the background counts. This negative binomial distribution is then the probability distribution of OC[0] given values of O, D, and S. Similarly, $\mathrm{OC}[-i]$ for $i \neq 0$ is assumed to have the negative binomial distribution with dispersion 3.52 and mean given by Eq. (1) except O is replaced by $O-i$.

Next we apply the technique just discussed to our *Cryptosporidium* epidemic curve estimation system. First, we obtained a bar graph (like the one in Fig. 2b), whose values are the daily counts of units of OTC antidiarrheal medication sold in Pharmacy A in North Battleford during the *Cryptosporidium* outbreak in spring, 2001 (obtained from [10]). After subtracting the average daily background count from the values in that bar graph, we smoothed the resultant bar graph using cubic splines to obtain a function $g(\mathrm{t})$. Next we let D_{N} and S_{N} be the duration and severity of the North

Battleford outbreak, B_N be the average daily background counts of units of antidiarrheal medicine sold in Pharmacy A in North Battleford, and B_C be the average daily background counts of units of antidiarrheal medicine sold in the pharmacies monitored in Cook County. Their values are $D_N = 47$, $S_N = 35.8\%$, $B_N = 1.973$, and $B_C = 986$. Then we set

$$\mu(O,D,S) = \left(\frac{B_C \times D_N \times S}{B_N \times \mathrm{D} \times S_N}\right) \times g((D_{OJ}-1)(O-1)/(D-1)-1)$$
$$= \left(\frac{986 \times 47 \times S}{1.973 \times D \times 35.8}\right) \times g(46(O-1)/(D-1)+1)$$

For given values of S and D, the result is a function of O whose shape is like $g(\mathrm{t})$ but which gives approximate counts of units of antidiarrheal medicine which would be sold in the pharmacies monitored in Cook County if an outbreak with duration D and severity S took place in Cook County.

5) In this application, there is a single variable C that indicates certain holidays. Specifically, the variable was given the value "high" during the four days following July 4th and the days from December 26 through January 3, and the value "low" on all other days.

6) H : Based on manual inspection, we determined three values (low, medium, and high) for background sales data of antidiarrheal medicine in Cook County.

7) BC[$-i$] : For each combination of values of H, Day, and C, we computed the mean and variance of the count over all background days corresponding to the combination of values. We then made the conditional distribution of BC[$-i$], given this particular combination of values of H, Day, and C, a negative binomial distribution with this mean and variance.

Notice that the component just described is able to detect an outbreak of the monitored disease from the daily counts. That is, as $P(O = 0)$ becomes smaller, we become increasingly suspicious of an outbreak. For example, we could detect an outbreak of Cryptosporidium from the daily counts of units of OTC antidiarrheal medicine sold. So the model presented here not only can be used to estimate the epidemic curve, but also to detect the outbreak.

2.2. Component two

Next we describe the component that models the relationship between the variables that constitute the epidemic curve and the outbreak severity, outbreak duration, and the number of days into the outbreak. As in the previous subsection, we first describe the structure of the network, and then we show how to determine the parameter values.

2.2.1. Causal structure

The structure of the network appears in Fig. 5. The variable $E[0]$ represents the number of individuals' first showing symptoms of the illness today. Again, by today we mean the current day on which we are investigating the outbreak. The variable $E[-i]$ represents the number of individuals first showing symptoms i days before today, while the variable $E[i]$ represents the number of individuals first showing symptoms i days after today. We need sufficient variables in both directions so that we can estimate the entire epidemic curve on any day of an outbreak. Note that the variables for which $i > 0$ represent future values on the epidemic curve (if the outbreak goes uncontrolled). These are the values we want to predict.

2.2.2. Parameter values

Let OJ be some jurisdiction in which we have the epidemic curve for an actual outbreak $\mathrm{Outbreak}_{\mathrm{OJ}}$ This epidemic curve is a bar graph showing the daily disease counts during $\mathrm{Outbreak}_{\mathrm{OJ}}$. Examples appear in Figs. 1 and 4. Jurisdiction OJ does not need to be the same jurisdiction used to obtain the conditional distributions of OC[i] (See Section 2.1.2). Let RJ be the jurisdiction for which we are developing the system, and let

1) D_{OJ} be the duration of $\mathrm{Outbreak}_{\mathrm{OJ}}$,
2) S_{OJ} be the severity of $\mathrm{Outbreak}_{\mathrm{OJ}}$,
3) N_{OJ} be the number of people in OJ,
4) N_{RJ} be the number of people in RJ.

Using cubic splines, we first smooth the bar graph representing the daily disease counts during $\mathrm{Outbreak}_{\mathrm{OJ}}$ to obtain a continuous function $e(\mathrm{t})$ on the interval $[1, D_{OJ}]$. Then, to create the conditional distribution of $E[0]$, given values of O, D, and S, in a system which is monitoring RJ, we first define the mean of $E[0]$, conditional on these values, as follows:

$$\upsilon(O, D, S) = \left(\frac{N_{RJ}}{N_{OJ}}\frac{S}{S_{OJ}}\frac{D_{OJ}}{D}\right)e((D_{OJ}-1)(O-1)/(D-1)+1). \qquad (2)$$

For given values of S and D, the result is a function of O whose shape is like $e(t)$ but which approximates disease counts which would occur if an outbreak with duration D and severity S took place in RJ. Intuitively, $\mathrm{N}_{\mathrm{RJ}}/\mathrm{N}_{\mathrm{OJ}}$ scales the function to RJ instead of OJ, D_{OJ}/D scales the function to describe an outbreak with duration D instead of D_{OJ}, S/S_{OJ} scales the function to describe an outbreak with severity S instead of S_{OJ}, and the expression in the argument of e changes the domain of the function from $[1, D_{OJ}]$ to $[1, D]$.

Given values of S and D, the function υ is a scaled replica of e. We inserted random fluctuation by letting υ be the mean of a neg-

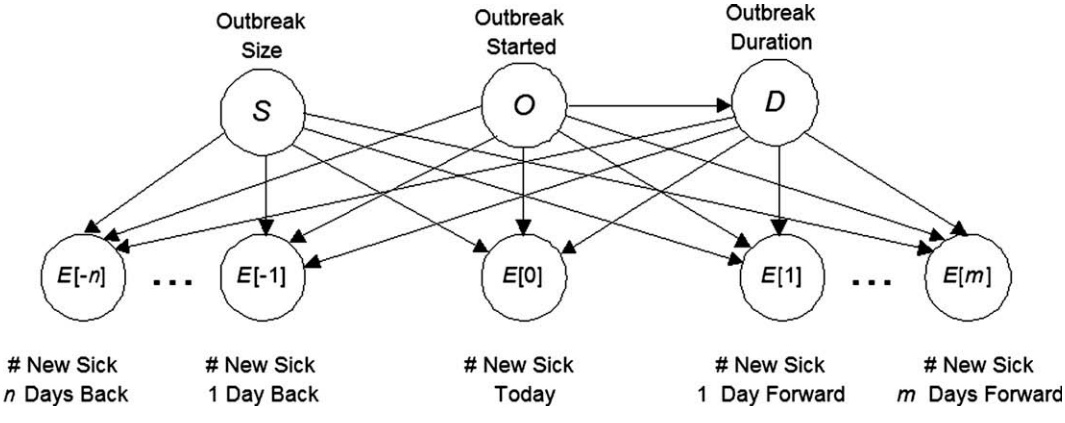


Fig. 5. The component that models the relationship between the variables which constitute the epidemic curve and the outbreak severity, outbreak duration, and the number of days into the outbreak.

ative binomial distribution with dispersion 3.52. This negative binomial distribution is then the probability distribution of $E[0]$ given values of O, D, and S. Similarly, $E[i]$ for $i \neq 0$ is assumed to have the negative binomial distribution with dispersion 3.52 and mean given by Equality 2 except O is replaced by $O + i$.

2.3. Combining Components One and Two

Our entire Bayesian network model consists of the two components just described. That is, it combines the networks in Figs. 3 and 5. By doing inference in this network, we can determine the conditional probability distributions of $E[i]$ for all i from the observed values of the variables *DAY*, *C*, and TC[i] for all i on a given day of the outbreak. If, for example, we are currently in the 9th day of the outbreak, the value of variable $E[0]$ is the epidemic curve value on the 9th day of the outbreak, the value of variable $E[-1]$ is the epidemic curve value on the 8th day of the outbreak, the value of variable $E[1]$ is the epidemic curve value on the 10th day of the outbreak, and so on. If we enter values of *DAY*, *C*, and TC[i] for all i on the 9th day of the outbreak, we will obtain conditional probability distributions of $E[0]$, $E[-1]$, $E[1]$, etc. We use the expected values relative to these distributions as our estimates of the epidemic curve values.

The fact that the variables *DAY* and *C* have an effect on the conditional probabilities of the variables of interest ($E[i]$ for all i) is a subtle matter. Initially, these variables are independent of *DAY* and *C* because the edges touching each variable labeled TC[i] are both into TC[i]. This is called a head-to-head meeting of the edges. However, when TC[i] is instantiated, the variables, which are above and below it, are rendered dependent, which means OC[i] is rendered dependent on the variables *DAY* and *C*. This dependency is then passed on to the variables labeled $E[i]$. This matter is discussed in detail in [10]. The following is a classic intuitive example: Suppose earthquakes and burglars can both cause your burglar alarm to sound, and there is no causal relationship between earthquakes and burglars. We would then create a causal (Bayesian) network with only edges from the earthquake and burglar nodes to the alarm node. A priori the earthquake and the burglar nodes are independent. Suppose now that we learn the alarm has sounded. We would fear that we were burglarized. However, if we later learned that there was a mild earthquake, this event would explain away the alarm sounding, thereby making a burglary less likely. Psychologists call this "discounting". Thus the burglar and the earthquake nodes are rendered dependent by the instantiation of the alarm node.

For the experiments described next, we used the Bayesian network package Netica (http://www.norsys.com/) to develop the networks and perform the inference.

3. Experiments

3.1. Method

Health care officials carefully reconstructed epidemic curves for the North Battleford, Saskatchewan *Cryptosporidium* outbreak in spring, 2001 [3] and the Milwaukee, Wisconsin *Cryptosporidium* outbreak in spring, 1993 [12]. Furthermore, counts of units of OTC antidiarrheal medicine sold were available for certain pharmacies in both jurisdictions. The North Battleford outbreak had a severity of 35.8% and duration of 47 days, while the Milwaukee outbreak had a severity of 25% and duration of 27 days. We developed an instance of the model using the North Battleford outbreak, and we tested it using the Milwaukee outbreak. That is, we used North Battleford as the outbreak jurisdiction and Milwaukee as the monitored jurisdiction. Detection was based on the daily counts of purchases of OTC antidiarrheal medicine.

We did not have actual epidemic curves for any influenza outbreaks. However, we did have a large amount of data concerning significant influenza outbreaks in two jurisdictions. Both outbreaks started in fall, 2003 and lasted 66–68 days. One of our data sources prefers that the jurisdictions remain anonymous. So we simply labeled them A and B. Our data sources are the Centers for Disease Control and Prevention (http://www.cdc.gov) and the National Retail Data Monitor system managed by RODS laboratory (http://rods.health.pitt.edu). Using the durations of the outbreaks, the influenza-like illness (ILI) curves, counts of deaths due to influenza during the outbreaks (which were also available), and national figures concerning influenza and influenza deaths, we were able to estimate epidemic curves for these outbreaks. We developed an instance of the model using the outbreak in jurisdiction A, and we tested it using the outbreak in jurisdiction B. That is, we used jurisdiction A as the outbreak jurisdiction and jurisdiction B as the monitored jurisdiction. Detection was based on the daily counts of patients presenting in the ED with respiratory symptoms. The discussion of parameter values in Section 2 concerning the *Cryptosporidium* model applies to the influenza model except for these modifications. (1) The probability distribution of the variable O was as follows: $P(O = 0) = 1-70/(2 \times 365)$ and $P(O = i) = 1/(2 \times 365)$ for $1 \leqslant i \leqslant 70$; (2) There was no variable C representing cyclical effects. It was assumed that there is no holiday season and so forth that affect ED visits with respiratory symptoms.

We evaluated each instance by determining how well it estimated the "gold standard" epidemic curve on given days of the outbreak. In the case of the *Cryptosporidium* outbreaks, we used the epidemic curve for the Milwaukee outbreak, which was carefully reconstructed by health care officials, as the gold standard. In the case of the influenza outbreak, we used the epidemic curve for the outbreak in jurisdiction B, which we reconstructed, as the gold standard.

3.2. Results

We show results for estimates that are obtained about 1/5 of the way into the outbreaks and 1/3 of the way into the outbreaks. In the case of the *Cryptosporidium* outbreak these are the estimates obtained on day 5 and day 9 of the outbreak, while in the case of the influenza outbreak these are the estimates obtained on day 13 and day 22 of the outbreak. We consider 1/5 of the way into the outbreak *very early* estimation, and 1/3 of the way into the outbreak *early* estimation.

3.2.1. Cryptosporidium outbreak

Fig. 6 shows the estimated epidemic curves obtained on the 5th and 9th days of the 27 day *Cryptosporidium* outbreak in Milwaukee, Wisconsin in spring, 1993. The gold standard epidemic curve is also shown. Table 1 shows two ways of measuring the similarity of the two sequences. If we let x_{i} be the sequence of values in the gold standard epidemic curve and y_{i} be the sequence of values in the estimated epidemic curve, we obtain

$$\text{per cent error} = 100 \times \frac{\sum_{i=1}^{27} |\, y_{\mathrm{i}} - x_{\mathrm{i}} \,|}{\sum_{i=1}^{27} x_{\mathrm{i}}}.$$

If the sequences were the same, this value would be 0. We also show the Pearson correlation, which is given by

$$\frac{1}{27}\sum_{i=1}^{27}\left(\frac{x_{\mathrm{i}} - \bar{x}}{\sigma_{\mathrm{x}}}\right)\left(\frac{y_{\mathrm{i}} - \bar{y}}{\sigma_{\mathrm{y}}}\right),$$

where $\bar{x}$ represents the average. The Pearson correlation is 1 if there is a linear relationship between the two sequences. This value mea-

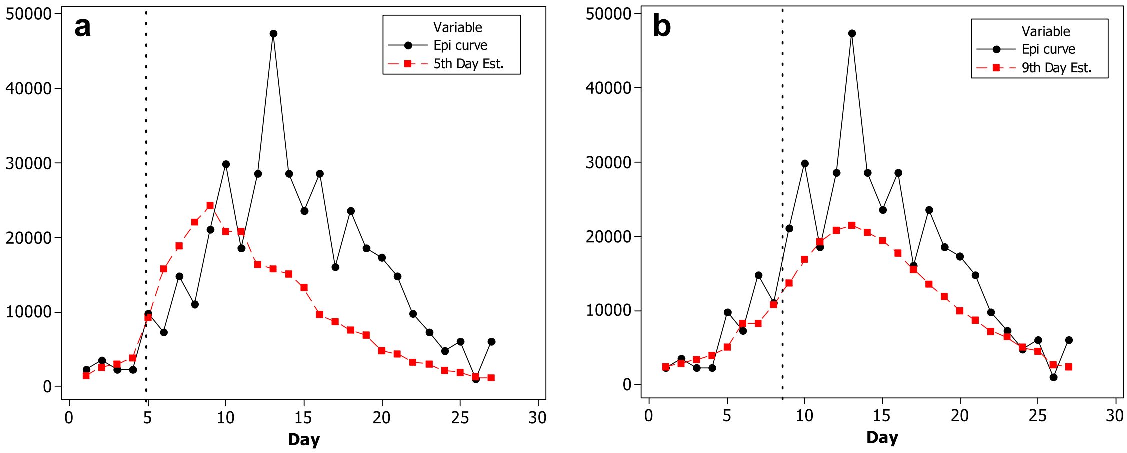


Fig. 6. Estimates of the Milwaukee *Cryptosporidium* epidemic curve on days 5 and 9 of the 27 day outbreak. Values before the dotted line are estimates, while those after it are predictions.

Table 1
Measures of similarity of estimated epidemic curves to gold standard epidemic curve for the Milwaukee *Cryptosporidium* outbreak

Day	Per Cent Error	Pearson correlation
5	51.9	.600
9	33.9	.909

Table 2
Measures of similarity of estimated epidemic curves to gold standard epidemic curve for the influenza outbreak in jurisdiction B

Day	Per Cent Error	Pearson correlation
13	91.6	.966
22	10.8	.935

sures the correlation between the sequences, but it does not reflect how close the values are to each other. So it only indicates whether the curves have similar shapes.

We see that on the 5th day of the outbreak the estimated curve decreases too early and the severity of the outbreak is underestimated. It appears difficult to predict whether the outbreak will become severe when it is still very early in the outbreak. One explanation for this is that the counts that occur very early in severe and moderate outbreaks may not be very different. On the other hand, by the 9th day the estimate is fairly good.

3.2.2. Influenza outbreak

Fig. 7 shows the estimated epidemic curves obtained on the 13th and 22nd days of the 66 day influenza outbreak in jurisdiction B in fall, 2003, while Table 2 shows the Per Cent Error and Pearson correlation. Note that, even though the *percent* error as much smaller on the 22nd day than on the 13th day, the Pearson correlation is slightly smaller. Recall that the Pearson correlation only measures whether the curves have similar shapes and not whether the values in the two curves are close. Looking at Fig. 7, we see that the shapes of the two curves on the 13th day are indeed more similar in that they both reach their peaks on about the same day. On the 22nd day the estimated curve is skewed a bit to the left of the actual curve. As was the case for the *Cryptosporidium* estimates, the very early (13th day) estimate is poor, but the early (22nd day) estimate is quite good. In this case, the severity of the outbreak was significantly overestimated very early in the outbreak. Notice that even the severity during the first 13 days of the outbreak was overestimated. Recall that we do not have actual epidemic curve values (daily counts of individuals with influenza) for the first 13 days. All we have is daily counts of patients presenting with respiratory symptoms in the monitored Emergency Departments. Based on these daily counts, the system estimates the size, duration, and days since the outbreak started. The early ED visit counts indicated

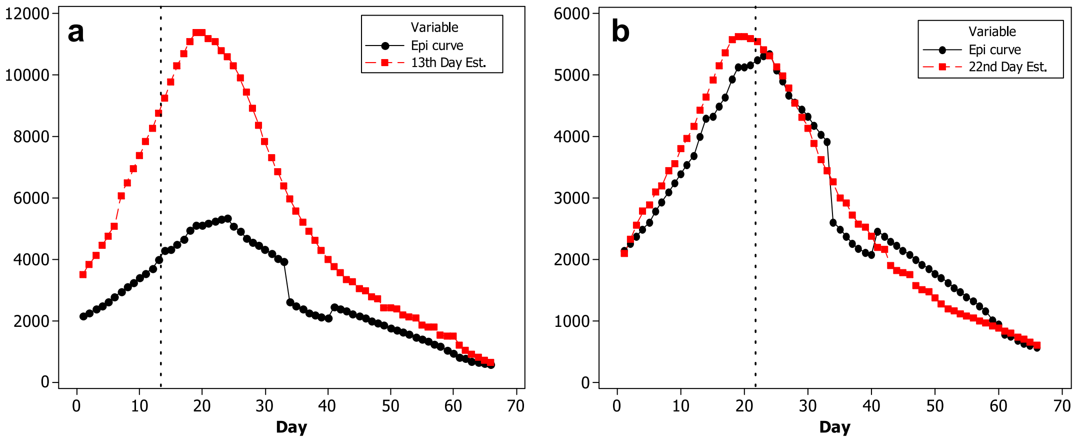


Fig. 7. Estimates of the influenza epidemic curve for jurisdiction B on days 13 and 22 of the 66 day outbreak. Values before the dotted line are estimates, while those after it are predictions.

that the severity of the outbreak was greater than that which would have been indicated by the actual (unknown) number of influenza cases if they were known. Since the severity was overestimated, the entire epidemic curve was estimated to be too large, which means the estimates on the first 13 days were be too large.

The epidemic curve estimates for both the *Cryptosporidium* and influenza outbreaks were quite good about 1/3 of the way into the outbreaks. On the other hand, while the estimate for the *Cryptosporidium* outbreak was good about 1/5 of the way into the outbreak, the estimate for the influenza outbreak was much too large 1/5 of the way into the outbreak. These results indicate that perhaps, early in the influenza outbreak, individuals presented in the ED with respiratory symptoms more often than we would expect based on the severity of the outbreak. It is difficult to know with confidence why this happened. There could, for example, have been other spurious causes of respiratory symptoms present in this community early in this particular outbreak. Furthermore, it could have been due to statistical variation. That is, in this particular influenza outbreak perhaps more people than usual, who had influenza, decided to go to the ED early in the outbreak. In the case of the Cryptosporidium outbreak we monitored purchases of an OTC medication, while in the case of the influenza outbreak we monitored ED visits. Possibly there is less statistical variation in the sale of OTC medication purchases than there is in ED visits. Additional studies are needed to resolve this issue.

3.2.3. Further evaluation

Our method uses actual epidemic curves from real previous outbreaks to construct the epidemic curves for new outbreaks. The question remains as to whether this is useful, or if we can obtain comparable results by only assuming that the curve is unimodal. In [14] we only made the assumption that the epidemic curve is unimodal, and we obtained the estimates shown in Fig. 8 for the 66 day influenza outbreak in jurisdiction B, whose estimates were obtained using our method appear in Fig. 7. Note that the totals in Fig. 8 are weekly and not daily totals. So it would be difficult to perform a detailed analysis comparing the estimates in Figs. 7 and 8. However, using visual inspection, we can see that the estimate in Fig. 8 is very bad on the 20th day of the outbreak, and even the 25-day and 30-day estimates do not appear to be as good as the 22-day estimate shown in Fig. 7. So at least for this particular outbreak the method that uses the information in a previous epidemic curve appears to yield better results.

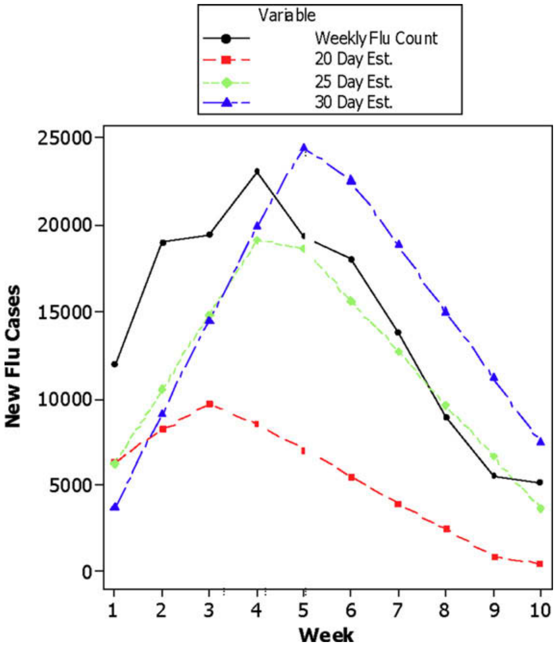


Fig. 8. Estimates of the influenza epidemic curve for jurisdiction B using the method described in [14] which only assumes the epidemic curve is unimodal.

4. Discussion

The model presented here is an initial version of a Bayesian network model that estimates an epidemic curve early in an outbreak. Preliminary investigation with two actual outbreaks indicates it is capable of providing informative estimates about 1/3 of the way into the outbreaks but not about 1/5 of the way into the outbreaks. Previously, most similar systems have been evaluated using only simulations (See [7] and [13]). We performed our evaluation using two types of real outbreaks.

Our preliminary results are encouraging, but they only serve to establish some foundations in the relatively new field of epidemic curve estimation. More extensive modeling and testing is needed. In particular, we need to more thoroughly study the robustness of the assumptions that the epidemic curve for an outbreak is unimodal, and that all outbreaks of a given type reach their peak about the same fraction of days into the outbreak.

A possible concern is our implicit assumption concerning the buying behavior of the population being monitored. Recall that T_{OJ} is the sub-region of the outbreak jurisdiction OJ from which we obtain our counts of the observable event, and T_{RJ} is the sub-region of the monitored jurisdiction RJ from which we obtain our counts of the observable event, B_{OJ} is the average daily count of the observable event in T_{OJ} when no outbreak is occurring, and B_{RJ} is the average daily count of the observable event in T_{RJ} when no outbreak is occurring, To model the buying behavior of individuals in RJ instead of OJ we used the scaling factor B_{RJ}/B_{OJ}. By so doing, we have modeled the behavior of individuals in RJ only when no outbreak is occurring, not when one is occurring. However, we apply this scaling factor to model their buying behavior during an outbreak. So we have implicitly assumed that individuals in RJ react during an outbreak in the same way as individuals in OJ. For example, suppose that on the average 10 units of a monitored OTC drug are sold each day in RJ when there is no outbreak, and on the average 20 units are sold in OJ when there is no outbreak. Next suppose that an outbreak of the same duration and severity occurs in RJ as the one that occurred in OJ. Then if 50 additional units were sold on day 5 of the outbreak in OJ, the expected value of the number of additional units sold in RJ on day 5 is (10/20)50 = 25.

Another concern is that the estimate for the *Cryptosporidium* outbreak was good at 1/5 of the way into the outbreak, whereas the estimate for the influenza outbreak was much too large 1/5 of the way into the outbreak. We noted at the end of Section 3 that perhaps this was due to the fact that we monitored ED visits in the case of the influenza outbreak and OTC sales in the case of the *Cryptosporidium* outbreak.

Future research could investigate the following: (1) Whether all outbreaks of a given type do reach their peak about the same fraction of days into the outbreak when the outbreaks are unimodal. (2) Whether good results can be obtained using this model when this assumption does not hold well, but the outbreaks are unimodal. That is, perhaps the performance of the system is robust relative to this assumption. Bayesian network performance has been shown to be robust relative to the parameters (probability distributions) in the network [15]. (3) Whether we can obtain better results by building models using multiple outbreaks. For example, in the case of influenza outbreaks, sometimes the epidemic can have a secondary wave. In this case, no model could predict the epidemic curve before the onset of the second wave. By using multiple outbreaks, we can explicitly build into the model the year-to-year variability in influenza outbreaks. If we ran the data against several curves shapes, some with only one wave, and some with a secondary wave, we may find that an epidemic curve with a secondary wave is more probable than one without one, once the onset of the secondary wave occurs. (4) Whether, during an outbreak, individuals in different jurisdictions react the same way regarding their purchases of the OTC drug being monitored (or relative to whatever observable event we are monitoring). Future research could investigate the robustness of the model if this assumption does not exactly hold. (5) Whether we can obtain better results by monitoring OTC sales than we can obtain by monitoring ED visits.

Investigations (2) and (3) can be performed using simulated outbreaks. Performing investigations (1), (4), and (5), will be more difficult because they require the observation of real outbreaks. The difficulty is that real outbreaks are rare, and the collection of data during outbreaks had been even rarer. However, for the past several years data on observable events related to influenza outbreaks have been collected for four cities by the RODS Laboratory at the University of Pittsburgh. A problem is that epidemic curves are not available for the influenza outbreaks that occurred during that time. Currently, we are working on a way to estimate the epidemic curves.

An improvement to the model would be to look at multivariate time series (counts of several observable events) rather than only univariate time series (counts of a single observable event). As discussed at the end of Section 2.1.1, the system can readily be modified to handle multivariate times series. We need only to include count nodes for all observable events. Future investigations can incorporate this improvement since data on counts of several observable events have been collected by the RODS Laboratory.

Acknowledgments

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

References

[1] Le Strat Y, Carrat F. Monitoring epidemiological surveillance data using hidden markov models. Stat Med 1999:18.

[2] Quiroz E, Bern C, MacArthur J, et al. An outbreak of cryptosporidiosis linked to a foodhandler. J Infect Dis 2000:181.

[3] Stirling R, Aramini J, Ellis A, Gillien L, Meyers R, Flevry M, et al. Waterborne cryptosporidiosis outbreak, North Battleford, Saskatchewan spring 2001. Health Can 2001:2.

[4] Wagner MM, Gresham LS, Dato V. Case detection outbreak detection and outbreak characterization. In: Wagner MM, editor. Handbook of biosurveillance. New York: Elsevier; 2006.

[5] Cooper GF, Dash DH, Levander JD, Wong WK, Hogan WR, Wagner MM. Bayesian biosurveillance of disease outbreaks. In: Chickering MD, Halpern JY, editors. UAI '04 proceedings of the 20th conference in uncertainty in artificial intelligence. Banff, Canada: AUAI Press; 2004.

[6] Cooper GF, Dowling, JN, Lavender, JD, Sutovsky, P. A Bayesian algorithm for detecting CDC category A outbreak diseases from emergency department chief complaints. In: Proceedings of syndromics 2006. Baltimore: Maryland; 2006.

[7] Hogan WR, Cooper GF, Wallstrom GL, Wagner MM, Dipinay JM. The Bayesian aerosol release detector: an algorithm for detecting and characterizing outbreaks caused by an atmospheric release of bacillus anthracis. Statistics in medicine 2007;26:5225–52.

[8] Jiang X, Wallstrom GL. A Bayesian network for outbreak detection and prediction. In: Proceedings of AAAI-06. Boston: Massachusetts; 2006.

[9] Wallstrom GL, Wagner MM, Hogan WR. High-fidelity injection detectability experiments: a tool for evaluation of syndromic surveillance systems. CDC morbidity and mortality weekly report 2005; August 26 (Supplement).

[10] Neapolitan RE, Learning Bayesian networks Upper Saddle River. New Jersey: Springer; 2004.

[11] Cheeseman P, Stutz J. Bayesian classification (autoclass): theory and results. In: Fayyad D, Piatesky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining Menlo Park. California: AAAI Press; 1995.

[12] Mac Kenzie WR, Hoxie NJ, Proctor ME. A massive outbreak in Milwaukee of cryptosporidium infection transmitted through the public water supply. N Engl J Med 1994;331:22.

[13] Wallstrom GL. Probabilistic interpretation of surveillance data. In: Wagner MM, editor. Handbook of biosurveillance. New York: Elsevier; 2006.

[14] Jiang X, Wagner M, Cooper GF. Modeling the temporal trend of the daily severity of an outbreak using Bayesian networks. In: Holmes D, editor. Applications of Bayesian networks. New York: Springer-Verlag; 2008.

[15] Henrion M, Pradhan M, Del Favero B, Huang K, Provan G, O'Rorke P. Why is diagnosis using belief networks insensitive to imprecision in probabilities? In: Horvitz E, Jensen F, editors. Uncertainty in artificial intelligence; proceedings of the twelfth conference. Burlington, Massachusetts: Morgan Kaufmann; 1996.