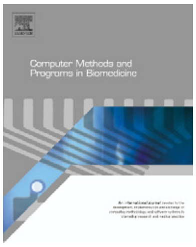
ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



# Multivariate Bayesian modeling of known and unknown causes of events—An application to biosurveillance

*Yanna Shen*[a,*,1], *Gregory F. Cooper*[b]

[a] *The Lister Hill National Center for Biomedical Communications, Bethesda, MD, USA*
[b] *Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA*

ARTICLE INFO

*Article history:*
Received 24 January 2010
Received in revised form
29 November 2010
Accepted 30 November 2010

*Keywords:*
Bayesian modeling
Unknown causes of events
Disease outbreak detection

ABSTRACT

This paper investigates Bayesian modeling of known and unknown causes of events in the context of disease-outbreak detection. We introduce a multivariate Bayesian approach that models multiple evidential features of every person in the population. This approach models and detects (1) known diseases (e.g., influenza and anthrax) by using informative prior probabilities and (2) unknown diseases (e.g., a new, highly contagious respiratory virus that has never been seen before) by using relatively non-informative prior probabilities. We report the results of simulation experiments which support that this modeling method can improve the detection of new disease outbreaks in a population. A contribution of this paper is that it introduces a multivariate Bayesian approach for jointly modeling both known and unknown causes of events. Such modeling has general applicability in domains where the space of known causes is incomplete.

## 1. Introduction

Bayesian modeling of unknown causes of events is an important and pervasive problem. However, it has received relatively little research attention. In general, an intelligent agent (or system) has only limited causal knowledge of the world. Therefore, the agent may well be experiencing the influences of causes outside its model. For example, suppose a robot that is exploring a dangerous physical environment is experiencing metal corrosion from an ambient gas that has not been characterized before (by it or anyone else). If the robot is limited to reasoning about corrosion using only the causes of corrosion that are in its knowledge base, it may well diagnose the cause as being the most probable one in its knowledge base. Such a faulty diagnosis could lead it to take incorrect countermeasures to stop the corrosion, rather than to investigate the chemical properties of the new cause of corrosion, which in turn could lead it to discover more effective counter measures. As another example, which is closely related to this paper, a clinician may be seeing a patient with a virus that is new to humans; historically, the HIV virus is one such example. It is important that clinicians be able to recognize that a patient is presenting with a heretofore unknown disease.

In general, intelligent agents (and systems) need to recognize under uncertainty when they are likely to be experiencing influences outside their realm of knowledge. This paper illustrates a Bayesian approach to doing so in the context of disease-outbreak detection, which we briefly survey in the remainder of this section.

Detection of anomalous events in data is an emerging area of research with important applications in domains such

[*] *Corresponding author at*: The Lister Hill National Center for Biomedical Communications 38A/9N912A, U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA. Tel.: +1 301 435 3181.
E-mail address: yanna.shen@nih.gov (Y. Shen).

[1] This research was conducted and completed when the author was in the Intelligent Systems Program, University of Pittsburgh, Pittsburgh, Pennsylvania.

0169-2607/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2010.11.015

as disease outbreak detection [1], fraud detection [2], and electronic intrusion detection [3]. In a typical scenario, a monitoring system examines a sequence of data to determine if any recent activity can be considered a deviation relative to historical baseline behavior. Many detection algorithms, such as statistical quality control [4], regression [5], time series models [6], and wavelets [7,8], use frequentist statistical techniques that derive statistics, such as *p* values.

With such approaches, it can be difficult to incorporate any prior knowledge and information that we may have, as for example our prior beliefs about the size, location, and temporal progression of a potential outbreak. In contrast, Bayesian methods excel at incorporating such prior knowledge and information. The Bayesian approach introduced in this paper uses informative prior probabilities to model known outbreak diseases (e.g., influenza and anthrax), and relatively non-informative priors to model unknown outbreak diseases.

Bayesian approaches have been developed that can be applied to anomaly detection, such as dynamic linear models [9] and hidden Markov models [10]. These methods can detect a wide range of anomaly types, but usually at the expense of being less effective at detecting any particular type, as for example an outbreak due to inhalational anthrax. Thus, they are at the generic-outbreak-detection end of the spectrum.

At the other end of the spectrum, we can use Bayesian methods to model specific diseases. Consider, for example, that a large-scale airborne release of inhalational anthrax has known spatio-temporal characteristics such as a specific incubation time and a plume-like spatial distribution. Thus, when monitoring for such an outbreak, a detection algorithm can be vigilant in watching for these characteristics. BARD [11] is a Bayesian outbreak-detection algorithm that models the effects of an outdoor airborne anthrax release using the Gaussian plume model of atmospheric dispersion and a disease-specific model of inhalational anthrax.

The number and variety of possible outbreak diseases that could in theory appear, but have not yet appeared, is so large that it is not practical to represent them explicitly by using disease-specific models, even if we could predict well what they might be. An example is a new, highly contagious respiratory virus that has never been seen before. This paper introduces a Bayesian approach for modeling both known and unknown diseases within a single framework. We combine an unknown-disease model with models of known diseases to obtain a hybrid modeling approach. The goal is to detect known causes of anomalies well and to detect unknown causes at all.

If an outbreak due to disease *d* occurs in the population, patients infected with disease *d* are often expected to exhibit several disease symptoms of *d*. Although the joint appearance of evidential features may be highly predictive of an outbreak, many detection algorithms monitor only a single evidential feature, which may limit the surveillance system's detection capabilities. The Bayesian approach introduced in this paper extends the univariate approach in [12] to model multiple evidential features of every person in the population. We call this approach the *multivariate Bayesian hybrid detection* algorithm or the *MBH* algorithm. Although this paper focuses on Bayesian modeling on unknown diseases, the general ideas transfer to modeling unknown events and entities in many other domains.

## 2. Background

This section describes a Bayesian framework that is used for combining models of known and unknown diseases. In addition, we provide a brief background regarding non-informative prior distributions and Beta distributions that are used as priors in our disease models.

### 2.1. Bayesian framework

Let $H$ be a hypothesis and $E$ denote some available evidence. We are often interested in knowing the posterior probability of $H$ in light of $E$, that is $P(H|E)$. Assume we can estimate the likelihood $P(E|H)$. Frequently such likelihoods are derived from a model that represents the probability that $H$ is associated with $E$. A Bayesian approach requires the specification of a prior probability of $H$, namely $P(H)$, which is our belief in $H$ before seeing evidence $E$. Eq. (1) shows the well-known use of the Bayes' theorem (rule) to derive $P(H|E)$.

$$P(H|E) = \frac{P(E|H)P(H)}{\sum_{H' \in S} P(E|H')P(H')}, \tag{1}$$

where the sum is taken over all hypotheses $H'$ in a mutually exclusive and presumed exhaustive set $S$ of hypotheses that are each modeled as having a non-zero prior probability.

The hypotheses in $S$ can be at different levels of abstraction. Consider an anomaly detection application in which we are monitoring population evidence $E$ for new outbreaks of disease. Such evidence might include the symptoms of patients who have recently visited emergency departments in a given region. Suppose $S$ includes some set of disease-specific disease outbreaks (e.g., outbreaks due to inhalational anthrax, SARS, and influenza), another hypothesis in $S$ might represent the absence of any disease outbreak in the population. Traditionally a Bayesian diagnostic system contains only hypotheses for specific disease outbreaks and for the non-outbreak condition. However, in this paper we propose to also represent all the diseases (known and unknown) that are not being modeled by a given set of disease-specific disease outbreaks. For example, such an outbreak disease could be smallpox, if smallpox is not modeled in the set of disease-specific disease outbreaks. That is, we know about smallpox, but for whatever reason are not explicitly modeling it currently. As another example, such a disease could be a new infectious disease that has never been seen before. In this case, we do not know well how to model it.

In other words, then, we will include in $S$ the union of the hypotheses for specific disease outbreaks, for the non-outbreak condition, and for unknown disease outbreaks. Unknown diseases are so numerous and oftentimes imponderable that it is not practical (or even possible) to try to represent them explicitly. A primary purpose for including a model of unknown diseases in $S$ is to identify patterns of evidence $E$ that are not similar to those associated with non-outbreak diseases or any of the specific outbreak diseases that we are modeling.

### 2.2. Priors that are used in the disease models

Non-informative priors are sometimes called “objective priors”. We use these priors to reflect a situation where there is a relative lack of knowledge about a parameter. Specifically, for modeling unknown outbreak diseases in this paper we use a non-informative prior distribution in the form of a uniform distribution on the interval [0,1] for a Binomial proportion parameter $p$, which we describe in detail below.

Castillo and Colosimo, as well as many others, suggest a non-informative prior for parameters defined over a finite range to be uniform in that range [13]. An example of this was proposed by Bayes himself [14], who used a uniform [0,1] on the Binomial proportion parameter $p$. Tuyl et al. also suggest using the uniform prior Beta(1, 1), called the Bayes–Laplace prior, on the Binomial proportion parameter $p$ to represent ignorance [15].

We use informative prior distributions to model the outbreak diseases that we know about and have modeled. We model six outbreak diseases classified by the CDC as serious bioterrorism threats, plus the following diseases: influenza, hepatitis A, cryptosporidiosis, and asthma. We call these diseases CDC-$A^+$ diseases.

We use non-uniform Beta distributions to represent prior belief in modeling the CDC-$A^+$ diseases. The Beta distribution is a continuous probability distribution that is parameterized by two positive shape parameters ($\alpha$ and $\beta$). This distribution has been used for a wide variety of applications because it can flexibly specify a range of forms of distributions from peaked ($\alpha, \beta > 1$) to uniform ($\alpha = \beta = 1$) and from U-shaped ($0 < \alpha, \beta < 1$) to skewed or either monotonically decreasing or increasing [16].

The Beta distribution can be used to represent the uncertainty or random variation of a rate or proportion. In particular, the Beta distribution is a conjugate prior of the Binomial likelihood function and, as such, it is often used to describe the uncertainty about a Binomial probability parameter, as we do in this paper.

## 3. Methodology

In this section, we describe the multivariate Bayesian hybrid detection algorithm (the MBH algorithm) in the context of disease-outbreak detection. MBH extends the univariate version of the Bayesian hybrid detection algorithm described in [12] and takes as input the binary state of emergency department patient clinical findings, such as *cough* = *present* vs. *absent*, *fever* = *present* vs. *absent*, and *diarrhea* = *present* vs. *absent*, during the most recent 24 h. Extracting specific clinical findings from electronic emergency department patient reports remains a research challenge [17], although good progress is being made [18–20]. This paper assumes that in the foreseeable future we will be able to obtain a set of clinical findings for each patient who visits the ED. Thus, the multivariate disease model uses such evidence rather than assuming we only will have a single patient chief complaint, which is typically readily available. For simplicity, MBH currently models multiple clinical findings for each person in the population by assuming conditional independence among findings given individual's disease state, as described in the sections below. As discussed in Section 5, relaxing the conditional independence assumption is possible and is an area for future work.

### 3.1. Notation

The term ED that is used below refers to emergency departments in the region being monitored. The total patient cases across all EDs are treated as a single pool.

Let $D_0$ represent all the diseases that ED patients can have in the absence of any disease outbreak in the population, and let $d_0$ represent an arbitrary member of $D_0$ (e.g., acute appendicitis would be one such non-outbreak disease). We will call these diseases *non-outbreak diseases*.

Let $D_K$ represent all the outbreak diseases that we know about and have modeled. Assume that there are $K$ types of such known outbreak diseases, as for example influenza, cryptosporidiosis, and anthrax. Let $d_k$ represent a specific outbreak disease in $D_K$, where $1 \leq k \leq K$.

Let $D_*$ represent all the outbreak diseases that are unknown or unmodeled. Let $d_*$ represent an arbitrary member of $D_*$. For example, $d_*$ might be a newly mutated type of virus that previously was innocuous to human health, but now is potentially lethal.

Let the total number of individuals being monitored in a given region be $N$.

Let $i$, $1 \leq i \leq N$, represent the index of a specific person in the population.

Let $j$, $1 \leq j \leq J$, represent the index of a specific disease symptom, where $J$ is the total number of symptoms that are modeled. The MBH algorithm takes as input patient disease symptoms, of which there can be more than one per patient, as for example, a patient presents with *cough* = *present*, *fever* = *present*, and *headache* = *absent*, and $j = 2$ represents the binary symptom *fever*.

Let $OB$ represent the state of an outbreak existing during the most recent 24-h period in the region being monitored, and let $NOB$ represent the absence of any disease outbreak during that period. Note that $OB$ and $NOB$ are mutually exclusive and exhaustive, and thus, $P(disease\_outbreak\_status = OB) + P(disease\_outbreak\_status = NOB) = 1$.

### 3.2. An entity-based disease model

The disease model we use is an entity-based Bayesian network model, which represents all the people in the population (not just the ED patients). Consider Fig. 1 that shows an example of the plate notation for such a model, where the plate is used to repeat the inner subgraph $N$ times, and $N$ represents the total number of people being monitored in a given region, as described above [21]. Fig. 1 shows an example disease model where we model a univariate symptom *cough* of every person in the population, and the cough state of every patient who came to the ED in the last 24 h is *present* or *absent*. For those individuals in the population who did not come to the ED, the *cough* variable has the value *unknown*.

When multivariate symptoms exist, such as cough, fever, and diarrhea, we assume that these symptoms (evidence) are conditionally independent given person's disease state. The assumption of conditional independence between evidential features makes it easier to convey the basic approach in this

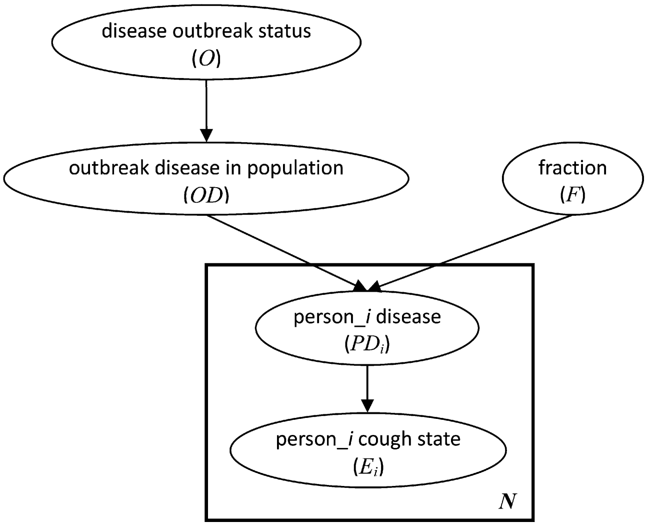


**Fig. 1 – Plate notation of a univariate Bayesian network model. The subgraph in the plate (bolded box) repeats N times, where N is the number of individuals in the population being monitored, and any links that cross a plate boundary are replicated once for each subgraph repetition. See the text next for a description of the nodes and the conditional probability tables.**

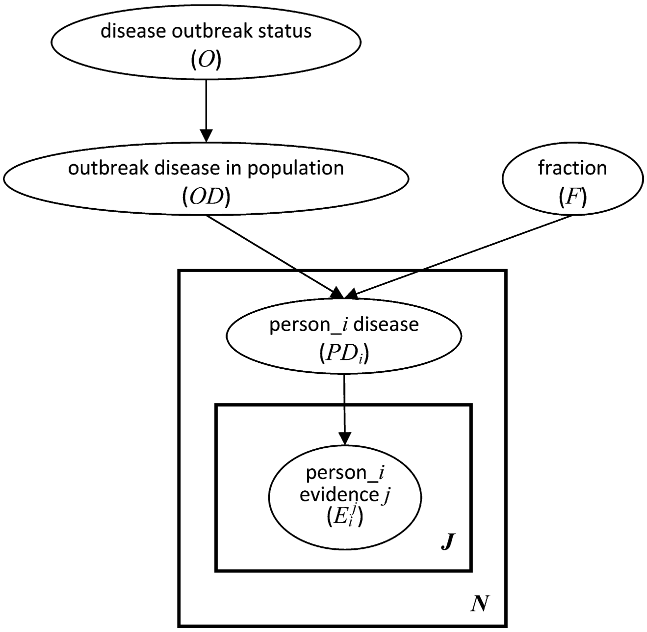


**Fig. 3 – Plate notation of the multivariate Bayesian network model showing the MBH disease model where each person's disease state and evidence state are modeled using a naïve Bayes model. *J* on the inner plate denotes that there are a total of *J* evidential features modeled for person i, and N on the outer plate denotes the total number of population being monitored in the region. See the text next for a description of the nodes and the conditional probability tables.**

paper. Based on this assumption, we model a person's symptom states and his or her disease state using a naïve Bayes model, as shown in Fig. 2. Thus, by modeling multivariate symptom states using Fig. 2, we obtain the plate notation for a multivariate disease model as shown in Fig. 3.

### 3.2.1. *The nodes*

The node *disease outbreak status* represents the outbreak status in the population during the most recent 24-h period. Let $O$ represent this node, where $O = OB$ or $NOB$.

The node *outbreak disease in population* represents the particular outbreak disease that is hypothesized to be present in the population. Let $OD$ denote this node. $OD$ can have the value *none* (no outbreak) or $d_k$ for $k > 0$ (outbreak of known disease $d_k$) or $d_*$ (outbreak of an unknown disease $d_*$). We assume in the current model that different disease outbreaks would not occur simultaneously; however, the model could be extended to allow for multiple disease outbreaks.

The node *fraction* represents the hypothetical fraction of the total population that has the outbreak disease and has visited the ED in the last 24 h with the outbreak disease in population (if any). Let $F$ denote this node. Let $f$ denote an arbitrary value of $F$. For example, $f$ might be $10^{-4}$ or $2 \times 10^{-5}$ or any of a wide range of fractions. We assume that the probability that an individual in the population will visit the ED with the outbreak disease on any given day is equal to the hypothesized fraction $f$ of the population with the outbreak disease who will visit the ED on that day.

The node *person_i disease* represents the possible diseases that person $i$ can have, given outbreak disease $OD$ in the population. Let $PD_i$ denote this node. For the people who did not come to the ED in the previous 24 h, we have that $PD_i = noED$. For the people who came to the ED in the previous 24 h, $PD_i$ is a random variable that can take on values $d_0, d_1, \ldots, d_K, d_*$.

If $OD$ = none, a specific person $i$ either has $d_0$ or his (her) status is *noED*. Note that $d_0$ represents that an individual (1) went to the ED during the last 24-h period and (2) has a non-outbreak ED disease.

When $OD = d_k$ (for $1 \le k \le K$), a specific person $i$ could either present to the ED with outbreak disease $d_k$, present with non-outbreak disease $d_0$, or not present (*noED*). That person cannot have another outbreak disease, because as mentioned in the current model we assume that there is at most one outbreak disease present in the population at any given time. Similarly, when $OD = d_*$, a specific person $i$ could present to the ED with $d_*$, present with $d_0$, or not present (*noED*).

Given the disease state of a specific person $i$ in the population, we use the *person_i evidence state* node to model the

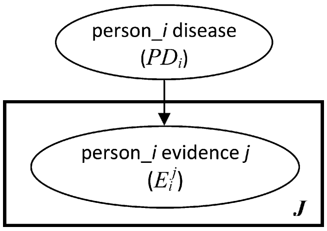


**Fig. 2 – A naïve Bayes model (plate notation) representing the total *J* evidential features for a specific person *i* in the population.**

symptom state of that person. We model a total number of $J$ disease symptoms of person $i$ as $E_i^1, \ldots, E_i^j, \ldots, E_i^J$, where $E_i^j$ represents disease symptom $j$ for person $i$. Let $e_i^j$ represent the value of $E_i^j$. For a person who came to the ED in the last 24 h, his (or her) symptom state $E_i^j$ is modeled as having symptom $j$ as *present* ($e_i^j$) or *absent* ($\sim e_i^j$). For people who did not visit the ED, our convention is to assign $E_i^j$ to be the value *unknown*.

### 3.2.2. *The conditional probability tables*

Estimating the prior probability $P(O = OB)$ can be difficult due to limited literature and lack of previous outbreak surveillance data on which to base such estimates. Let $E$ be the status of the multiple symptom states for every person in the population, and $e$ be its value. The MBH algorithm described in this paper actually derives the likelihood ratio $LR = P(E = e|O = OB)/P(E = e|O = NOB)$, instead of the posterior probability $P(O = OB|E = e)$, in order to remove the need to specify this difficult prior probability. Moreover, the evaluation measures that we use are not sensitive to this particular prior probability; that is, these performance measures are the same, regardless of the value of $P(O = OB)$.

If $O = NOB$, the model represents that there is no disease outbreak occurring in the population in the last 24 h, i.e., $P(OD = \text{none}|O = NOB) = 1$. If $O = OB$, the model represents that some outbreak due to disease $d_k$ (or $d_*$) is occurring in the population. In Section 4.2, we discuss how we estimate the value of $P(OD = d_k$ (or $d_*$)$|O = OB)$.

In this paper we do not model a dependency between $F$ and $OD$; however, in general the disease model in Fig. 3 could be readily extended to represent a dependency between these two variables.

We derive the values of $f$ of the fraction node $F$ as $n/N$, where $N$ is the total number of individuals in the population who could potentially visit the EDs in the region, and $n$ represents the number of outbreak cases who visited the ED when there is a disease outbreak in the population. We model 15 values of $n$ that increases over the mean number of patients who came to the ED during days when there presumptively was no disease outbreak in the population. For example, one value of $n$ is equal to one standard deviation above this mean. The values range up to at most five standard deviations above the mean. The fraction $F$ is assumed to be uniformly distributed over these 15 discrete values. See [22] for details regarding how we estimated the values of $f$.

If $OD = \text{none}$, a specific person $i$ either has $d_0$ or his (her) status is *noED*; the probability that the person has $d_0$ and presents to the ED, which is denoted as $\theta$, is estimated from past ED data during which it is assumed no outbreak was occurring. Then $P(PD_i = noED|OD = \text{none}, F = f) = 1 - \theta$.

When $OD = d_k$ (for $1 \leq k \leq K$), a specific person $i$ could have disease $d_0$, $d_k$, or *noED*. That person cannot have another outbreak disease, because, as mentioned, in the current model we assume that there is at most one outbreak disease present in the population at any time. Recall that $d_*$ represents an unknown disease, which by definition means it is not a modeled disease $d_k$. Therefore, we have $P(PD_i = d_*|OD = d_k, F = f) = 0$. The probability of person $i$ having $d_k$ is equal to the value of the *fraction* node, $f$, by the construction of that node. Thus, there is $1 - f$ fraction of the total population who do not present to the ED with $d_k$ (i.e., who have $d_0$ or *noED*). It is assumed that a fraction $\theta$ of these people present to the ED with $d_0$. Thus, the probability of person $i$ presenting to the ED with $d_0$ in light of $d_k$ as an outbreak disease in the population is modeled as being equal to $(1 - f)\,\theta$. Finally, $P(PD_i = noED|OD = d_k, F = f) = 1 - f - (1 - f)\,\theta = (1 - f)(1 - \theta)$.

**Table 1 – The conditional probability table for $P(E_i^j|PD_i)$.**

| | $PD_i = d_0$ | $PD_i = d_k$ | $PD_i = d_*$ | $PD_i = \text{noED}$ |
|---|---|---|---|---|
| $E_i^j = e_i^j$ | $p_0^j$ | $p_k^j$ | $p_*^j$ | 0 |
| $E_i^j = \sim e_i^j$ | $1 - p_0^j$ | $1 - p_k^j$ | $1 - p_*^j$ | 0 |
| $E_i^j = \text{unknown}$ | 0 | 0 | 0 | 1 |

When $OD = d_*$, we can similarly derive $P(PD_i = d_*|OD = d_*, F = f) = f$, $P(PD_i = d_k|OD = d_*, F = f) = 0$, $P(PD_i = d_0|OD = d_*, F = f) = (1 - f)\,\theta$, and $P(PD_i = noED|OD = d_*, F = f) = (1 - f)(1 - \theta)$.

To facilitate describing the basic approach in this paper, we assume that the symptoms $E_i^1, \ldots, E_i^J$ are conditionally independent given the disease state of a person ($PD_i$). This assumption is not required, but it makes the exposition of the key concepts in the paper more straightforward. Extending the work to include symptoms dependencies is useful in future research, as we mention in Section 5. We use a naïve Bayes model (Fig. 2) to represent the conditional independence of symptoms given a disease state. This model has been used extensively in biomedical informatics and other fields, and it often performs classification remarkably well [23,24]. Recall that for a person who came to the ED in the last 24 h, his or her evidence state $E_i^j$ is modeled as having symptom $j$ as *present* ($e_i^j$) or *absent* ($\sim e_i^j$). The Bernoulli distribution provides a simple and natural way to model such a binary symptom [25,26], $P(E_i^j = present|PD_i)$. A standard Bernoulli distribution requires that such "success rate" probabilities be constant. However, we do not have confidence in these probabilities. To represent our uncertainty in how diseases are manifested clinically, we model $P(E_i^j = present|PD_i)$ as a random variable. Table 1 describes the conditional probability assignments for $P(E_i^j|PD_i)$, where $p_0^j$ is a random variable that represents the probability that a person came to the ED in the last 24 h, and that person has symptom $j$ as *present* given he (or she) has disease $d_0$. Random variables $p_k^j$ and $p_*^j$ can be defined analogously.

The next two sections describe how we model random variables $p_0^j$, $p_k^j$ and $p_*^j$ in the disease-specific model (DSM) and the unknown-disease model (UDM). Recall that a total of $J$ disease symptoms are modeled as conditionally independent. For simplicity, we thus describe disease modeling in terms of a specific symptom $j$ and ignore the superscript $j$ that represents the index of that symptom. Each symptom $j$ is modeled using an informative or non-informative prior probability distribution based on the person's disease state, as described below. All the multiple disease symptoms are modeled as being conditionally independent given the person's disease state.

### 3.2.3. *The disease-specific model (DSM)*

As stated, this model represents that a person has a specific disease $d_0$ or $d_k$ (for $0 \leq k \leq K$). Recall that $p_0$ ($p_k$) represents the probability of a specific symptom $j$ given a person having $d_0$ ($d_k$). We assume $p_0$ is distributed according to a Beta distribution, namely, $p_0 \sim \text{Beta}(\alpha_0, \beta_0)$. We also assume $p_k \sim \text{Beta}(\alpha_k, \beta_k)$.

Next, we describe how we modeled $p_0$ and $p_k$ using informative priors.

We estimated the parameters $\alpha_0$ and $\beta_0$ based on real ED reports from a large healthcare system in Pittsburgh from January to December 2002. Ref. [22] provides details regarding how we estimated these parameters.

Let $p_k = P(E_i^j = e_i^j | PD_i = d_k)$ as above, where $1 \le k \le K$. For the purpose of assessment, $p_k$ may be viewed as a fraction in the large sample limit of patient cases. We assessed parameters $\alpha_k$ and $\beta_k$ based on expert judgments for $1 \le k \le K$. The expert provided his expectation $\mu_k$ of $p_k$ and an interval assessment $[a_k, b_k]$ for which he stated a belief that there is a 90% chance that $p_k$ is between $a_k$ and $b_k$. Parameters $\alpha_k$ and $\beta_k$ were then estimated by solving Eqs. (2) and (3) in terms of the distribution Beta($p_k$; $\alpha_k$, $\beta_k$).

$$\mu_k = \frac{\alpha_k}{\alpha_k + \beta_k}. \tag{2}$$

$$\int_{a_k}^{b_k} \mathrm{Beta}(p_k;\ \alpha_k,\ \beta_k) dp_k = 90\%. \tag{3}$$

### 3.2.4. *The unknown-disease model (UDM)*

This model represents that a person has an unknown outbreak disease $d_*$ that we know little about. We model $p_*$, the probability of the symptom state of a specific symptom *j* in a patient with $d_*$, using a non-informative prior. As described in Section 2.2, many researchers have advocated the use of a uniform [0,1] distribution as a non-informative prior on a binary outcome. We model $p_*$ using an uniform distribution over [0,1], or equivalently, $p_* \sim \mathrm{Beta}(1, 1)$. Thus, prior to consideration of any data, this approach models every probability of the symptom as being equally likely given the presence of the unknown disease.

## 3.3. *Inference*

The objective of inference is to derive the posterior probability of an outbreak occurring given the observed evidence. In this paper, we apply a common outbreak-detection measure, the likelihood ratio (*LR*) method, that is not sensitive to the prior probability of there being an outbreak [27], and thus we do not specify disease outbreak priors here. Although these outbreak priors affect the magnitude of the posterior probabilities, they do not affect the relative order of the posterior probabilities that are obtained by running the MBH algorithm on a specific outbreak dataset (scenario). The evaluation method described in this paper determines the expected detection time (at a specific false positive rate) based on the relative order of the output *LR*s, which yields the same relative order as posterior probabilities.

We derive the likelihood ratio *LR* as $LR = P(E = e|O = OB)/P(E = e|O = NOB)$, where *e* denotes the status of the multiple symptom states for all the people in the population. By expanding the numerator of the above equation, we obtain the following equation:

$$LR = \frac{\sum_{OD \neq d_0} P(E = e|O = OD)P(OD|O = OB)}{P(E = e|OD = d_0)}. \tag{4}$$

We derive $P(E = e|OD)$ by setting *OD* to be one of $d_0$, $d_k$ or $d_*$, and then performing inference on the Bayesian network in Fig. 3. Inference is complicated by the fact that $P(E_i^j = e_i^j | PD_i)$ is not a point probability, but rather a distribution, as described in Sections 3.2.3 and 3.2.4. Recall that each person is modeled as having *J* symptoms, and each symptom state of person *i* is modeled as being *present* or *absent* using a Bernoulli distribution. For example, a person who came to the ED could have symptom states as being *cough* = *present*, *fever* = *present*, and *diarrhea* = *absent*. Given that we are modeling distributions over probabilities, it turns out that using existing exact inference methods to perform inference on the Bayesian network in Fig. 3 would require exponential time complexity [22]. Thus, we applied stochastic methods to approximate $P(E = e|OD)$. In particular, we applied Monte Carlo integration [28] to approximate $P(E = e|OD)$. Monte Carlo integration is a method of approximating an expectation by the sample mean of a function of sampled random variables.

In particular, for each symptom *j* and each disease state that person *i* could have, we sample *M* times from the Beta distribution of $P(E_i^j = e_i^j | PD_i)$ to get a total number of *M* sampled values. For each sample *k*, we used the sampled value as the value of $P(E_i^j = e_i^j | PD_i)$. Given a point value of $P(E_i^j = e_i^j | PD_i)$, we can use exact inference to efficiently compute $P_k(E = e|OD)$ from the Bayesian network in Fig. 3, where the subscript *k* denotes inference for the *k*th sample. Finally, we approximated the expectation of $P(E = e|OD)$ over an infinite number of samples by computing the expectation of *M* values of $P_k(E = e|OD)$.

We also investigated the use of importance sampling [28] to approximate $P(E = e|OD)$ and found that Monte Carlo integration (with and without importance sampling) converged well. Since the inference method is not the focus of this paper, we do not describe it in further detail here. Additional information is provided in [22].

# 4. Evaluation

We chose three diseases from the CDC-A$^+$ diseases for use in the experiments that we performed. The three diseases are *cryptosporidiosis*, *early stage anthrax*, and *inhalation tularemia*. We use each of the three diseases to simulate an outbreak due to disease $d_k$ for $1 \le k \le 3$, as described below. In each experimental simulation, for each disease we modeled three disease symptoms: *cough*, *headache*, and *abdominal pain*. MBH takes as input the three symptom states for each individual in the population, as for example *cough* = *present*, *headache* = *absent*, and *abdominal pain* = *absent*. We selected the three diseases and the three symptoms because these diseases and their symptoms contain a wide variety of distributional patterns (over $P(E_i^j | PD_i)$) among all the CDC-A$^+$ diseases.

## 4.1. *Datasets*

We obtained real ED cases for the year 2005 from a large hospital in Allegheny County, PA. The mean number of patients who visited the ED of this hospital per day was about 130. The time series of real ED cases of the hospital was used to estimate the number of people who are expected to come to the ED on a given day without any disease outbreak. Next, we describe how we simulated one outbreak dataset (scenario) due to disease $d_k$, where $d_k$ is a specific outbreak disease out of the three outbreak diseases that we selected for evaluation (*cryptosporidiosis*, *early stage anthrax*, and *inhalation tularemia*).

The background time series of *non-outbreak* cases was simulated based on the time series of real ED cases. On any given day (on or after midnight that day and before midnight the next day), we sampled from Beta($\alpha_0^j$, $\beta_0^j$) to determine the probability $p_0^j$ of a person having a specific symptom $j$ given that person had disease $d_0$. We then sampled from Binomial($n_0$, $p_0^j$) to determine the number of people having that symptom when there was no disease outbreak in the population on that day, where $n_0$ is the number of people who in reality came to the ED on that day, and persons that have symptom $j$ were selected randomly from $n_0$ people.

Recall that we assume that an individual's symptom states are conditionally independent given his or her disease state. Thus, assuming the state of independence, we did the above procedure for each of the three symptoms we selected (cough, headache, and abdominal pain). Therefore, for example, a possible ED patient case that might be generated by this process is (*cough* = present, *headache* = absent, *abdominal pain* = absent). A total of $n_0$ such cases would be generated for the current day. These generated cases with simulated symptom states are called *background cases* for that day. Note that we only created a single time series of background cases for all the experiments described below. Every dataset (outbreak scenario) was created by overlaying the simulated outbreak cases (as described below) onto this time series of background cases.

We simulated outbreak cases with disease $d_k$ by using a linear outbreak model called "Fictional Linear Onset of an Outbreak" (or "FLOO") that is described in [29]. A simulated FLOO($\Delta$,$T$) outbreak has duration $T$. It generates $t\Delta$ cases on day $t$ of the outbreak ($0 < t \le T/2$), and then generates $T\Delta/2$ cases per day for the remainder of the outbreak. The outbreak onset date (when $t = 0$) was generated randomly as described later in this section. We note that the FLOO model is but one of many possible alternative models that could be used to simulate the epidemic curve of an outbreak. Nonetheless, we view that FLOO provides a reasonable initial evaluation, and it has been used in previous studies [29,30].

Let $n_k$ be the number of simulated outbreak cases generated by the FLOO model during the previous 24-h period. We sampled from the distribution Beta($\alpha_k^j$, $\beta_k^j$) to determine the probability $p_k^j$ of the symptom $j$ appearing in each of the $n_k$ cases. We then sampled from Binomial($n_k$, $p_k^j$) to determine the number of the outbreak cases having disease $d_k$ and symptom $j$, where outbreak cases that have symptom $j$ were selected randomly from $n_k$ outbreak cases. We did this for each of the three symptoms we selected by assuming they are conditionally independent. Thus, for example, a possible outbreak patient case having disease $d_k$ that might be generated by this process is (*cough* = present, *headache* = present, *abdominal pain* = absent). A total of $n_k$ such cases were generated for the previous 24-h period.

We generated the onset dates of the simulated outbreak due to disease $d_k$ by randomly selecting 8 unique dates from each of the 12 consecutive months in 2005. We created one dataset by overlaying the simulated outbreak cases produced by the FLOO model onto the background ED cases starting at the onset date and continuing for the outbreak duration. We thus created 8 × 12 = 96 datasets (scenarios) of outbreaks due to disease $d_k$.

In order to evaluate the MBH algorithm using different magnitudes of disease outbreaks, we generated outbreak cases using three sets of FLOO parameters, which correspond to a low, medium, and high severity of disease outbreak. For each FLOO parameter setting and each disease that we selected, we generated 96 datasets, as described above. We thus generated 3 (FLOO settings) × 3 (diseases) × 96 (outbreak scenarios) = 864 datasets, with each dataset containing the symptom states of three disease symptoms of every person in the population. For the many people who did not visit the ED, their symptoms have the value *unknown*.

There are several reasons why we used simulated outbreak data in these initial experiments, rather than real outbreak data. As described in Section 3, we are not yet able to obtain a set of clinical findings for each patient who visits the ED because most of those findings are not electronically available in the EDs to which we have research access. Another problem with real data is that the date and time of real outbreaks are seldom known precisely. Thus, there are downsides to evaluating MBH using real data. While there are limitations to using simulated data, rather than real data, using simulated data does allow us to readily evaluate a detection algorithm using a variety of patterns of simulated disease outbreaks, such as different severities of disease outbreaks and different disease outbreak onset dates. For this reason, simulated data have frequently been used in research that evaluates biosurveillance algorithms [8,11,29,31,32]. Simulated data provide a useful approach to performing an initial set of experiments, such as those reported here. In future work, it will be worthwhile to evaluate these algorithms using real data as well.

### 4.2. *Experimental methods*

Let $d_u$ and $d_v$ be two distinct outbreak diseases. Table 2 shows our experiments for one such pair of $d_u$ and $d_v$. In this table, both experiments have simulated outbreaks due to disease $d_u$. However, disease $d_u$ is modeled in Exp. 1 but not modeled (e.g., $d_u$ is an unknown disease) in Exp. 2. DSM and UDM represent two versions of the detection system that are constructed by using either the DSM or the UDM model, respectively, as described in Sections 3.2.3 and 3.2.4.

In Exp. 1, UDM models an unknown disease $d_*$, as well as the known outbreak disease $d_u$. We conjectured that including $d_*$ here would not detract significantly from detecting the outbreak due to $d_u$. In contrast, DSM does not model $d_*$. We expected this model to detect $d_u$ somewhat faster than UDM, because the simulated outbreak was in fact due to $d_u$, but we conjectured it would not be appreciably faster.

In Exp. 2, UDM did not model $d_u$, however, the simulated outbreak was due to $d_u$. Nonetheless, UDM did model $d_*$. We conjectured that modeling $d_*$ would allow UDM detect a simulated outbreak due to $d_u$ faster than would DSM, which models neither $d_u$ nor $d_*$.

If the above conjectures proved true, the experiments would provide support that modeling an unknown disease

**Table 2 – A 2 × 2 table that summarizes the experiments. In all the experiments, the simulated outbreak disease is denoted as $d_u$. In Exp. 1 $d_u$ is modeled, whereas in Exp. 2 it is not.**

| | DSM system | UDM system |
|---|---|---|
| Exp. 1<br>($d_u$ is modeled) | Model $d_0$, $d_u$<br>Simulate outbreak cases from $d_u$ | Model $d_0$, $d_u$, $d_*$<br>Simulate outbreak cases from $d_u$ |
| Exp. 2<br>($d_u$ is not modeled) | Model $d_0$, $d_v$<br>Simulate outbreak cases from $d_u$ | Model $d_0$, $d_v$, $d_*$<br>Simulate outbreak cases from $d_u$ |

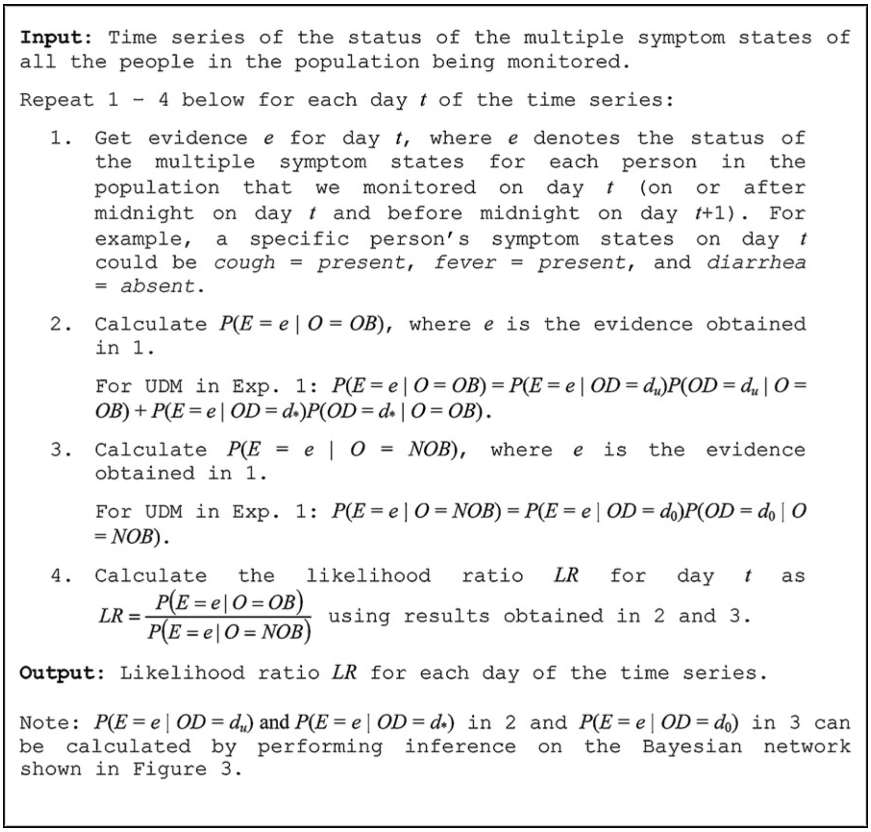
(in the form of $d_*$) provides a net benefit in detecting disease outbreaks.

In each of the four experiments represented by the cells in Table 2, we computed the likelihood ratio *LR* using Eq. (4). For the UDM model in Exp. 1, the sum in Eq. (4) is taken over *OD* equal to $d_u$ and $d_*$, and for UDM in Exp. 2, the sum is taken over $d_v$ and $d_*$. For DSM in Exp. 1, the sum consists only of the term $d_u$, and for DSM in Exp. 2, the sum of *OD* consists only of $d_v$. Fig. 4 shows pseudo-code of the MBH algorithm, in which we use the UDM detection system constructed in the context in Exp. 1 (as shown in the top right cell in Table 2) as an example to describe the process of this experiment. In this paper, due to space limitations, we only report experimental results when using a uniform prior over $P(OD|O = OB)$ for all values of *OD*. We performed a sensitivity analysis over this distribution, as is described in detail in [22].

Given the output of the likelihood ratio of an outbreak scenario for a specific experiment, we determined the detection time and false positive rate for various detection ratios. The detection time was the time from the simulated release until a detection ratio threshold *r* was exceeded. The false positive rate was derived as *FP*/*M*, where *FP* is the number of false positives that occurred using threshold *r* while monitoring a time series of simulated ED cases in which there was no (simulated) outbreak, and *M* is length in months for the time series, namely, $M = 12$.

We represent models DSM and UDM in Exp. 1 as DSM1 and UDM1, respectively, and likewise represent models DSM and UDM in Exp. 2 as DSM2 and UDM2. Let $E_{\text{DSM1}}$ be the average detection time of DSM1 over all the experiments described above at a false positive rate of one per month, since one false positive per month is commonly cited as an upper bound on a tolerable rate. Let $E_{\text{DSM2}}$ be the average detection time of DSM2 over all the experiments described above at a false positive rate of one per month. Define $E_{\text{UDM1}}$ and $E_{\text{UDM2}}$ analogously.

In order to determine the false positive rates under various detection thresholds, we ran the MBH algorithm using the DSM1, DSM2, UDM1, and UDM2 models on the background time series of ED cases in 2005, which we assumed to contain no outbreaks of the three diseases we are modeling. For each model, we selected the threshold *r* that yielded one false positive per month. Threshold *r* was applied to the output like-



```
Input: Time series of the status of the multiple symptom states of
all the people in the population being monitored.

Repeat 1 - 4 below for each day $t$ of the time series:

   1. Get evidence $e$ for day $t$, where $e$ denotes the status of
      the multiple symptom states for each person in the
      population that we monitored on day $t$ (on or after
      midnight on day $t$ and before midnight on day $t$+1). For
      example, a specific person's symptom states on day $t$
      could be cough = present, fever = present, and diarrhea
      = absent.

   2. Calculate $P(E = e \mid O = OB)$, where $e$ is the evidence obtained
      in 1.

      For UDM in Exp. 1: $P(E = e \mid O = OB) = P(E = e \mid OD = d_u)P(OD = d_u \mid O =$
      $OB) + P(E = e \mid OD = d_*)P(OD = d_* \mid O = OB)$.

   3. Calculate $P(E = e \mid O = NOB)$, where $e$ is the evidence
      obtained in 1.

      For UDM in Exp. 1: $P(E = e \mid O = NOB) = P(E = e \mid OD = d_0)P(OD = d_0 \mid O$
      $= NOB)$.

   4. Calculate the likelihood ratio $LR$ for day $t$ as
      $LR = \frac{P(E = e \mid O = OB)}{P(E = e \mid O = NOB)}$ using results obtained in 2 and 3.

Output: Likelihood ratio $LR$ for each day of the time series.

Note: $P(E = e \mid OD = d_u)$ and $P(E = e \mid OD = d_*)$ in 2 and $P(E = e \mid OD = d_0)$ in 3 can
be calculated by performing inference on the Bayesian network
shown in Figure 3.
```

**Fig. 4 – Pseudo-code of the MBH algorithm as applied to Exp. 1 using the UDM detection system.**

**Table 3 – Mean detection time (in days) of all four disease models over all the experiments, along with the *p*-values for the comparisons.**

| | DSM | UDM | *p*-Value |
|---|---|---|---|
| Exp. 1<br>($d_u$ is modeled) | 3.06 | 3.25 | $H_0$: $E_{UDM1} = E_{DSM1}$ vs. $H_a$: $E_{UDM1} > E_{DSM1}$<br>0.03 |
| Exp. 2<br>($d_u$ is not modeled) | 4.91 | 3.55 | $H_0$: $E_{UDM2} = E_{DSM2}$ vs. $H_a$: $E_{UDM2} < E_{DSM2}$<br>0.01 |

lihood ratios of an outbreak scenario of a specific experiment to determine its detection time under one false positive per month. Using this procedure, we obtained the detection time of all four disease models over all the experiments.

### 4.3. *Statistical analysis*

To evaluate the MBH algorithm, we first adopted the linear mixed effects model [33] to model the detection times that were obtained over all the experiments described above. We used a linear mixed effects model in order to take into account (1) the hierarchical nature of the detection time data and (2) the correlations between factors FLOO($\Delta$,$T$) and $d_u$, where FLOO($\Delta$,$T$) is the model that we used for generating the simulated outbreak cases, and $d_u$ is the disease that is causing the ongoing disease outbreak. Ref. [22] contains details regarding this linear mixed effects model.

We then performed Tukey's test[2] on the detection time data to evaluate the following null hypothesis $H_0$: $E_{UDM1} = E_{DSM1}$ vs. the alternative hypothesis $H_a$: $E_{UDM1} > E_{DSM1}$ for Exp. 1, and $H_0$: $E_{UDM2} = E_{DSM2}$ vs. the alternative hypothesis $H_a$: $E_{UDM2} < E_{DSM2}$ for Exp. 2. Table 3 shows the mean detection time (in days) of all four disease models over all the experiments, in which the last column shows the *p*-values of comparisons of DSM and UDM in Exp. 1 and Exp. 2 under a false alert rate of one per month. All these tests used a significance level of 0.05.

As shown in Table 3, at one false alert per month, modeling $d_*$ in Exp. 1 resulted in UDM having a detection time that was 0.19 days (= 4.6 h) slower than DSM with a statistical significance of 0.03. In Exp. 2, UDM detected the outbreak disease 1.36 days (= 32.6 h) faster than DSM with a statistical significance of 0.01. These results support the conjectures presented in Section 4.2.

### 4.4. *Decision analysis*

As described above, modeling an unknown disease $d_*$ yields a substantial decrease in detection time (~33 h) when the disease outbreak is caused by an unknown disease (Exp. 2). When the disease outbreak is due to a known outbreak disease (Exp. 1), modeling $d_*$ degrades the detection performance only modestly (~5 h). This section analyzes when modeling the possibility of an unknown outbreak disease will have a better expected detection performance than not modeling it.

Let event G denote the following event: given that an outbreak is occurring, it is due to a disease that is not being explicitly modeled in the detection system. According to Table 2, G is true in Exp. 2 and is false in Exp. 1. Let $q$ be the probability that G is true. Recall that we wish to evaluate whether modeling the possibility of an unknown disease occurring is a net positive in detecting disease outbreaks rapidly. If $q = 1$, then modeling $d_*$ will likely be helpful. If $q = 0$, however, modeling $d_*$ will be useless and possibly harmful by increasing the chance of a false positive alert. Our objective is to determine the value range of $q$ such that modeling an unknown disease $d_*$ (using UDM) yields an overall expected decrease in detection time. Based on deriving such an estimate of $q$, we can then determine whether to construct a DSM or an UDM model in a detection system. Fig. 5 shows such a decision analysis.

Recall that $E_{DSM1}$ and $E_{DSM2}$ are the average detection time of DSM1 and DSM2 over all the experiments described above at a false positive rate of one per month, respectively. Let $E_{DSM} = (1 - q) \times E_{DSM1} + q \times E_{DSM2}$. Define $E_{UDM}$ analogously. Let $q_*$ be the probability such that the equation below holds:

$$(1 - q) \times E_{UDM1} + q \times E_{UDM2} = (1 - q) \times E_{DSM1} + q \times E_{DSM2} \qquad (5)$$

Then $q_*$ is the threshold such that any probability greater than $q_*$ renders modeling $d_*$ helpful, given the conditions and assumptions of the evaluation. Solving Eq. (5) using the values in Table 3, yields $q_* = 0.12$. The standard error of computing $q_*$ is 0.1 [22]. If $q = 0.12$ then modeling $d_*$ is expected to be neither

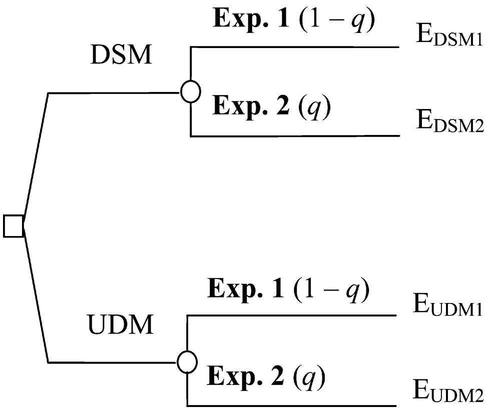


**Fig. 5 – A decision tree showing the decision analysis for selecting to use DSM vs. UDM for outbreak detection, where Exp. 1 and Exp. 2 denote as shorthand the condition represented by these experiments.**

[2] Tukey's test is frequently used as an adjustment for multiple-comparison procedure to find which means are significantly different from one another. It was performed in [22] in order to evaluate the disease detection performance of three disease models, in which two of the three disease models (DSM and UDM) are introduced in this paper. For simplicity of presentation, this paper only focus on disease model DSM and UDM, and reports experimental results of the two models.

helpful nor harmful. However, if $q > 0.12$, then including $d_*$ in the model is expected to decrease the detection time when the detection system is operating at an expected false alert rate of one per month.

It seems plausible that there are disease-outbreak monitoring situations in which if there is an outbreak then the probability exceeds 0.12 of it being due to an unknown disease. The Olympics provide one possible scenario, where a bioterrorist might attempt to use a new infectious disease agent to maximize terror. In such situations, the methods described in this paper could be beneficial.

## 5. Discussion and future work

This paper introduced a Bayesian method for disease-outbreak detection that combines models of known diseases and unknown diseases. In particular, we modeled the known non-outbreak disease $d_0$ using an informative prior estimated from past ED data, and we modeled a known outbreak disease $d_k$ (for $k > 0$) using informative priors that were assessed from an infectious disease expert. The unknown disease model uses a non-informative prior to model some unknown disease $d_*$. Simulation results show that this hybrid modeling approach can improve the detection of unknown disease outbreaks in the population.

Recall that the disease model in this paper does not model multiple disease outbreaks simultaneously. If this circumstance occurred, we conjecture that modeling $d_*$ would still improve the detection performance because we model $d_*$ using a uniform prior, which allows the disease model (UDM) to match a wide variety of outbreak-disease patterns.

As mentioned, the Bayesian approach that we described for modeling unknown diseases is based on specifying non-informative priors. There are numerous ways of specifying such non-informativeness, and in other work we have investigated several approaches beyond just using uniform distributions [22]. In particular, we studied semi-informative priors, in which some constraints are placed on the parameters of a disease model (e.g., the symptom *cough* has an increased rate of occurrence, relative to background rates), but otherwise the parameter distributions are uniform [26]. We also studied a semi-informative prior in the form of a mixture prior to model an outbreak disease that we partially know (e.g., a disease that has characteristics of an influenza-like illness) and that might manifest some disease symptoms similar to one or more known outbreak diseases. In particular, the mixture prior consists of several component priors of known outbreak diseases and a uniform component prior that represents our uncertainty about how partially-known diseases would appear [22]. Simulation results support that using a mixture of priors to model a partially-known disease is beneficial to the detection system's detection performance. We believe the investigation of non-informative and semi-informative priors holds significant promise in domains where causes of events may sometimes be unknown, including the medical domain.

Recall that the MBH algorithm models the binary state of every evidential feature, as for example, *cough* = *present* vs. *absent*, and *headache* = *present* vs. *absent*, by assuming the evidential features are conditionally independent given the disease state of an individual in the population. Assuming independence between evidential features has facilitated describing the basic approach in this paper. However, the approach can be extended to model symptoms that are conditionally dependent. In particular, we could model dependent symptoms using a Dirichlet-multinomial hierarchical model. The Monte Carlo inference method we used in this paper can be readily adapted to perform inference on such a model.

Finally, we note that the experiments we have described were based on simulations of disease outbreaks. It is difficult to obtain adequate real data on a range of disease outbreaks, which is why many disease-outbreak studies rely on simulations. We used real past ED data on non-outbreak diseases and expert assessments of outbreak diseases in an effort to develop quality simulation models. Recall from Section 4.1 that the MBH algorithm was evaluated on the simulated outbreak scenarios, in which the simulated symptom state of each patient case was generated by sampling from the Beta-Binomial model. The sampling method itself brings random effects into the outbreak scenarios to be tested. In addition, as described in Section 3.2.3, the probability of a symptom state in a disease was assumed to have a Beta distribution, while the data were simulated using the Beta-Binomial model, as described above. Thus, the simulated data contains another level of random effects. Nevertheless, it will be important in the future to evaluate further the methods described here using additional simulation models and ultimately using data on real outbreaks of a variety of diseases.

## Conflict of interest statement

The authors report that there are no disclosures relevant to this manuscript.

## Acknowledgements

This research was funded by grants IIS-0325581 and IIS-0911032 from the National Science Foundation and by grant P01-HK000086 from the Centers for Disease Control and Prevention.

## REFERENCES

[1] W.-K. Wong, Data mining for early disease outbreak detection, doctoral dissertation, Carnegie Mellon University, 2004.

[2] T. Fawcett, F. Provost, Adaptive fraud detection, Data Mining and Knowledge Discovery 1 (3) (1997) 291–316.

[3] D. Denning, An intrusion-detection model, IEEE Transactions on Software Engineering 13 (2) (1987) 222–232.

[4] L.C. Hutwagner, W. Thompson, G.M. Seeman, The bioterrorism preparedness and response early aberration reporting system (EARS), Journal of Urban Health 80 (2, Suppl. 1) (2003) i89–i96.

[5] R.E. Serfling, Methods for current statistical analysis of excess pneumonia–influenza deaths, Public Health Reports 78 (1963) 494–506.

[6] B.Y. Reis, K.D. Mandl, Time series modeling for syndromic surveillance, BMC Medical Informatics and Decision Making 3 (2) (2003).

[7] A. Goldenberg, G. Shmueli, R.A. Caruana, Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales, Proceedings of National Academy of Sciences 99 (8) (2002) 5237–5240.

[8] J. Zhang, F.C. Tsui, M.M. Wagner, W.R. Hogan, Detection of outbreaks from time series data using wavelet transform, in: AMIA Annual Symposium Proceedings, 2003, pp. 748–752.

[9] M. West, J. Harrison, Bayesian Forecasting and Dynamic Models, Springer-Verlag, New York, 1989.

[10] Y. LeStrat, F. Carrat, Monitoring epidemiologic surveillance data using hidden Markov models, Statistics in Medicine 18 (1999) 3463–3478.

[11] W.R. Hogan, G.F. Cooper, G.L. Wallstrom, M.M. Wagner, J.-M. Depinay, The Bayesian aerosol release detector: an algorithm for detecting and characterizing outbreaks caused by an atmospheric release of *Bacillus anthracis*, Statistics in Medicine 26 (2007) 5225–5252.

[12] Y. Shen, G.F. Cooper, Bayesian modeling of unknown diseases for biosurveillance, American Medical Informatics Association Annual Symposium Proceedings (November (14)) (2009) 589–593.

[13] E. Castillo, B.M. Colosimo, An introduction to Bayesian inference in process monitoring, control and optimization, in: B.M. Colosimo, E. Castillo (Eds.), Bayesian Process Monitoring, Control and Optimization, Chapman and Hall, Boca Raton, 2007, pp. 3–46.

[14] S.J. Press, Subjective and Objective Bayesian Statistics, 2nd ed., John Wiley & Sons, New York, 2003.

[15] F. Tuyl, R. Gerlach, K. Mengersen, Posterior predictive arguments in favor of the Bayes–Laplace prior as the consensus prior for binomial and multinomial parameters, Bayesian Analysis 4 (1) (2009) 151–158.

[16] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman & Hall, London, 1995.

[17] C.J. McDonald, The barriers to electronic medical record systems and how to overcome them, The Journal of the American Medical Informatics Association 4 (3) (1997) 213–221.

[18] W.W. Chapman, J.N. Dowling, M.M. Wagner, Fever detection from free-text clinical records for biosurveillance, Journal of Biomedical Informatics 37 (2) (2004) 120–127.

[19] W.W. Chapman, J.N. Dowling, M.M. Wagner, Classification of emergency department chief complaints into seven syndromes: a retrospective analysis of 527,228 patients, Annals of Emergency Medicine 46 (5) (2005) 445–455.

[20] D. Chu, Clinical feature extraction from emergency department reports for biosurveillance, Master's thesis, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, 2007.

[21] W.R. Gilks, A. Thomas, D.J. Spiegelhalter, A language and program for complex Bayesian modeling, The Statistician 43 (1) (1994) 169–177.

[22] Y. Shen, Bayesian modeling of anomalies due to known and unknown causes, doctoral dissertation, Intelligent Systems Program, University of Pittsburgh, Pittsburgh, 2009. http://www.cs.pitt.edu/~shenyn/sheny_etd2009.pdf.

[23] P. Domingos, M. Pazzani, Beyond independence: conditions for the optimality of the simple Bayesian classifiers, in: Proceedings of the International Conference on Machine Learning (ICML), 1996, pp. 105–112.

[24] P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning (29) (1997) 103–130.

[25] R.L. Scheaffer, L.J. Young, Introduction to Probability and its Applications, 3rd ed., Richard Stratton, 2009, pp. 121–125.

[26] Y. Shen, G.F. Cooper, A new prior for Bayesian anomaly detection – application to biosurveillance, Methods of Information in Medicine 49 (1) (2010) 44–53.

[27] S.N. Goodman, Toward evidence-based medical statistics 2: the Bayes factor, Annals of Internal Medicine 130 (12) (1999) 1005–1013.

[28] L. Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, New York, NY, 2004.

[29] D.B. Neill, A.W. Moore, G.F. Cooper, A Bayesian spatial scan statistic, Advances in Neural Information Processing Systems 18 (2006) 1003–1010.

[30] X. Jiang, G.F. Cooper, A Bayesian spatio-temporal method for disease outbreak detection, Journal of American Medical Informatics Association 17 (2010) 462–471.

[31] G.F. Cooper, D.H. Dash, J.D. Levander, et al., Bayesian biosurveillance of disease outbreaks, in: Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence, 2004, pp. 94–103.

[32] W.-K. Wong, A.W. Moore, G.F. Cooper, M.M. Wagner, What's strange about recent events (WSARE): an algorithm for the early detection of disease outbreaks, Journal of Machine Learning Research 6 (2005) 1961–1998.

[33] M. Davidian, Linear Mixed Effects Models for Multivariate Normal Data, Class Notes for Applied Longitudinal Data Analysis, North Carolina State University, 2007, pp. 363–422.