Bayesian Modeling of Unknown Diseases for Biosurveillance

Yanna Shen, MS and Gregory F. Cooper, MD, PhD
University of Pittsburgh, Pittsburgh, PA

Abstract

This paper investigates Bayesian modeling of unknown causes of events in the context of disease-outbreak detection. We introduce a Bayesian approach that models and detects both (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 key contribution of this paper is that it introduces a Bayesian approach for jointly modeling both known and unknown causes of events. Such modeling has broad applicability in medical informatics, where the space of known causes of outcomes of interest is seldom complete.

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, a clinician may be seeing a patient with a virus that is new to humans; the HIV virus was at one time such an example. It is important that clinicians be able to recognize that a patient is presenting with an 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.

In a typical scenario of anomaly detection, a monitoring system examines a sequence of data to determine if any recent activity can be considered a deviation relative to historical baseline behavior. Frequentist algorithms do so by deriving *p* values. However, compared with Bayesian approaches, it is difficult to incorporate into the analysis any prior information that we may have, as for example our prior beliefs about the size, location, and temporal progression of a potential outbreak. The Bayesian approach introduced in this paper uses informative prior probabilities to model known outbreak diseases (e.g., influenza), and relatively non-informative priors to model unknown outbreak diseases.

Bayesian approaches have been developed that can be applied to biosurveillance, such as hidden Markov models[1]. 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. On the other hand, we can use Bayesian methods to model specific diseases. 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[2].

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. We call this algorithm the Bayesian hybrid detection algorithm or the BH algorithm. The goal is to detect known causes of anomalies well and to detect unknown causes at all.

Methods

In this section, we describe the BH algorithm in the context of disease-outbreak detection. BH takes as input emergency department patient symptoms, such as cough, fever, and diarrhea, of the most recent 24 hours. We describe an example of this algorithm in terms of aggregate counts of the binary symptoms *cough* vs. *no cough*.

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

Let d_0 represent an arbitrary member of the set of diseases that ED patients can have in the absence of any disease outbreak in the population (e.g., acute

appendicitis would be one such non-outbreak disease).

For $k > 0$, let d_k represent an outbreak disease 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.

Let d_* represent an outbreak disease that is unknown or unmodeled.

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). Fig 1 shows plate notation for this model, where the plate is used to repeat the inner subgraph *N* times, and *N* represents the total number of people in the population. Due to limited space, this paper focuses more on the model structure. See [3] for details regarding how we specify the conditional probability tables of the Bayesian network model.

Let *OB* denote the presence of an outbreak in the population, and let *NOB* represent no outbreak. The node *disease outbreak status* represents the outbreak status in the population during the most recent 24-hour period. Let *O* represent this node, where $O = OB$ or *NOB*. If an outbreak occurred in the population at any time during the most recent 24-hour period, then $O = OB$; otherwise, $O = NOB$.

The node *outbreak disease in population* represents the outbreak disease that is 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.

If $O = NOB$, the model represents that there is no disease outbreak occurring in the population in the last 24 hours, i.e., $P(OD = none \mid O = NOB) = 1$. If $O = OB$, the model represents that some outbreak due to disease d_k (or d_*) is occurring in the population. We leave discussion of specifying $P(OD = d_k \text{ (or } d_*) \mid O = OB)$ to the Experiments section.

The node *fraction* represents the hypothetical fraction of the total population that has the outbreak disease d_k and has visited the ED in the last 24 hours 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 2×10^{-5} or any of a wide range of fractions.

The values of *F* depend on the temporal progression of the outbreak disease *OD* and what type of disease *OD* is, because an outbreak disease in an earlier stage tends to affect a smaller fraction of the population than a disease at a later stage, and some outbreak diseases tend to affect a larger fraction of population than other outbreak diseases. Since we do not represent temporal progression in the model in this paper, there is uncertainty in the disease stage of a potential outbreak disease. Thus, in this paper we do not model a dependency between *F* and *OD*. However, in general the disease model in Fig 1 could be readily extended to represent a dependency between *F* and *OD*.

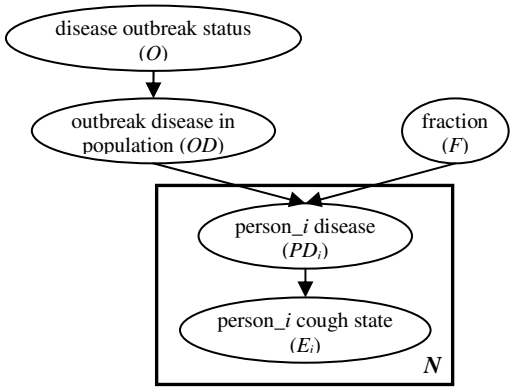

Fig 1 Plate notation for the entity-based disease model.

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 hours, we assume that $PD_i = noED$. For the people who came to the ED in the previous 24 hours, 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-hour period and (2) has a non-outbreak ED disease.

When $OD = d_k$ (for $1 \le k \le K$), a specific person *i* could 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 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 person *i*, we use the *person_i cough state* node to model the cough state of that person. Let E_i represent this node for a specific person *i* and let e_i represent the value of E_i. In this paper, we represent E_i as a binary symptom of person *i*, where $1 \le i \le N$. It is possible to model more than one symptom, and we have done so, but we restrict this paper to an example that contains only one symptom, which makes it easier to convey the basic approach. For people who came to the ED during the past 24 hours, their evidence states are

cough or *no cough*. For people who did not visit the ED, our convention is to assign E_i to be the value *unknown*.

Recall that we model the state of a binary symptom E_i of every person in the population. For the people who came to the ED, we define $P(E_i = cough \mid PD_i = d_0) = p_0$, $P(E_i = cough \mid PD_i = d_k) = p_k$, and $P(E_i = cough \mid PD_i = d_*) = p_*$. The symptom state of a person is modeled using a Bernoulli distribution. Thus, we have $P(E_i = no\ cough \mid PD_i = d_0) = 1-p_0$, $P(E_i = no\ cough \mid PD_i = d_k) = 1-p_k$, and $P(E_i = no\ cough \mid PD_i = d_*) = 1-p_*$. In the next section we describe how we model the distributions over the probability parameters p_0, p_k and p_*. Modeling distributions over parameters allows us to represent not only the parameters themselves, which model disease expression, but also our uncertainty about what the values of those parameters should be. Both forms of uncertainty are important.

***The known disease-specific model (DSM)*:** As stated, this model represents that a person has a specific disease d_0 or d_k (for $1 \le k \le K$). Recall that p_0 (p_k) represents the probability of a cough symptom given a person having d_0 (d_k). We modeled p_0 and p_k using informative priors. We assume p_0 is distributed according to a Beta distribution, namely, $p_0 \sim \text{Beta}(\alpha_0, \beta_0)$. We also assume $p_k \sim \text{Beta}(\alpha_k, \beta_k)$.

We estimated α_0 and β_0 based on real ED reports from a large healthcare system in Pittsburgh from January to December 2002. We estimated α_k and β_k based on expert assessment. See [3] for details regarding how we estimated these parameters.

***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 cough in a patient with d_*, using a non-informative prior. Castillo et al., as well as many others, suggest a non-informative prior for parameters defined over a finite range to be uniform in that range[4]. An example of this was proposed by Bayes himself[5], who used a uniform [0, 1] on the binomial proportion parameter p. Tuyl et al. also suggest using the uniform prior, called the Bayes-Laplace prior, on the binomial proportion parameter p to represent ignorance[6]. We model p_* using an uniform distribution over [0, 1], or equivalently, $p_* \sim \text{Beta}(1,1)$. Thus, for an unknown disease, we model that any probability of cough is equally likely.

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[7], and thus we do not specify disease priors here. Although these priors affect the magnitude of the posterior probabilities, they do not affect the relative order of the posterior probabilities. 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 *LRs*, which yields the same relative order as posterior probabilities. Thus computing *LR* does not affect the detection performance measure of BH that we use in this paper.

We derive the likelihood ratio *LR* as $LR = \frac{P(E = e \mid O = OB)}{P(E = e \mid O = NOB)}$, where e denotes the status of the symptom *cough* for every person in the population. We further calculate *LR* using the following equation:

$$LR = \frac{\sum_{OD \neq d_0} P(E = e \mid OD) P(OD \mid O = OB)}{P(E = e \mid OD = d_0)}. \quad (1)$$

We derive $P(E = e \mid OD)$ by setting OD to be one of d_0, d_k or d_*, and then performing inference on the Bayesian network in Fig 1. Inference is complicated by the fact that $P(E_i = e_i \mid PD_i)$ is not a point probability, but rather a distribution, as described above. Thus, in order to perform inference we needed to integrate over these distributions. Although space does not permit a detailed description of inference, we note that we applied a variation of the exact inference method given in [8], which is polynomial time in the number of people who came to the ED.

Experiments

We chose three diseases from the CDC-A$^+$ diseases for use in the experiments 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 one of the three disease symptoms among *cough*, *headache*, and *abdominal pain*. BH will take as input one symptom state of the population at a time, as for example *cough* vs. *no cough*. We selected the three diseases and the three symptoms because these diseases and their symptoms contain a wide variety of distributional patterns (over $P(E_i \mid PD_i)$) among all the CDC-A$^+$ diseases.

Datasets: We obtained real ED cases for 2005 from a large hospital in Allegheny County, Pennsylvania. 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 by simulating a specific disease symptom (out of the three) of people in the population.

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 $Beta(\alpha_0, \beta_0)$ to determine the probability p_0 of a person having a specific symptom given that person had disease d_0. We then sampled from $Binomial(n_0, p_0)$ 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. These generated cases with simulated symptom states that we call *background cases*.

Let S be a symptom of disease d_k (e.g., cough). We simulated outbreak cases with disease d_k by using a linear outbreak model called FLOO that is described in [9]. Let n_k be the number of simulated outbreak cases generated by FLOO per day. We sampled from the distribution $Beta(\alpha_k, \beta_k)$ to determine the probability p_k of the symptom S appearing in each of the n_k cases. We then sampled from $Binomial(n_k, p_k)$ to determine the number of the outbreak cases having disease d_k and symptom S.

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

In order to evaluate the BH algorithm using different scales of disease-outbreak scenarios, 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, each disease, and each symptom that we selected, we generated 96 datasets, as described above. We thus generated 3 (FLOO settings) × 3 (diseases) × 3 (symptoms) × 96 (outbreak scenarios) = 2592 datasets.

Experimental Methods: Let d_u and d_v be two distinct outbreak diseases. Table 1 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., unknown) in Exp. 2. DSM and UDM here represent two versions of the detection system that are constructed by including a DSM and an UDM model, respectively.

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 did 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.

Table 1 A 2 × 2 table that summarizes the experiments.

	DSM	**UDM**
Exp. 1 (d_u is modeled)	Model d_0, d_u. Simulate outbreak cases from d_u.	Model d_0, d_u, d_*. Simulate outbreak cases from d_u.
Exp. 2 (d_u is not modeled)	Model d_0, d_v. Simulate outbreak cases from d_u.	Model d_0, d_v, d_*. Simulate outbreak cases from d_u.

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 also did not model d_u.

If the above conjectures proved true, the experiments would provide support that modeling an unknown disease (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 1, we computed the likelihood ratio LR using Eq. 1. For the UDM model in Exp. 1, the sum in Eq. 1 is taken over 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 of OD consists only of the term d_u, and for DSM in Exp. 2, the sum of OD consists only of d_v. In this paper, due to space limitations, we only report experimental results for UDM when using a uniform prior over the appearance of the outbreak diseases being modeled.

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 in that time series, namely, $M = 12$.

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 1, 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. If $q = 1$, then modeling d_* will likely be helpful. If $q = 0$, however, modeling d_* will be useless and possibly harmful by allowing more chances for a false positive alert.

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, respectively. Let E_{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 frequently cited as an upper bound on a tolerable rate. Let E_{DSM2} be the average detection time of DSM2 over all the experiments described above at a false positive rate of one per month. Let $E_{DSM} = (1 - q) \times E_{DSM1} + q \times E_{DSM2}$. Define E_{UDM} analogously.

Results: In order to determine the false positive rates under various detection thresholds, we ran the BH 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 disease outbreaks. For each model, we selected the threshold r that yielded one false positive per month. Threshold r was applied to the output likelihood 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 mean detection time of all four disease models over all the experiments, as shown in Table 2.

Table 2 Mean detection time (in days) of all four disease models over all the experiments.

	DSM	**UDM**
Exp. 1 (d_u is modeled)	6.05	6.18
Exp. 2 (d_u is not modeled)	7.14	6.38

At one false alert per month, modeling d_* does not significantly degrade detection performance of UDM in Exp. 1. In Exp. 2, UDM detects the ongoing outbreak 18 hours faster than DSM as expected. The results support our conjectures above. Statistical analyses were conducted in [3], which also supports our conjectures. By solving $(1 - q) \times E_{UDM1} + q \times E_{UDM2} < (1 - q) \times E_{DSM1} + q \times E_{DSM2}$, we obtain $q > 0.14$. Under the assumptions introduced, this result indicates that if the probability is greater than 0.14 of an outbreak being due to an unknown disease, then including d_* in the model is expected to decrease the detection time at a false-positive rate of one alert per month.

Discussion and Future Work

It seems plausible that there are disease-outbreak monitoring situations in which if there is an outbreak then the probability exceeds 0.14 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 this paper could be beneficial. More generally, this paper has introduced a new Bayesian approach for detecting events due to causes of any type for which we have little knowledge.

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.

There are numerous ways of specifying non-informativeness when modeling unknown diseases. For example, we are studying 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 non-informative [3]. We believe the investigation of non-informative and semi-informative priors holds significant promise in biomedical domains, where the causes of events may as yet be undiscovered.

Acknowledgements

This research was funded by a grant from the National Science Foundation (NSF IIS-0325581).

References

[1] LeStrat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in Medicine. 1999;18:3463-78.

[2] Hogan WR, Cooper GF, Wallstrom GL, Wagner MM, Depinay J-M. 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.

[3] Shen Y, Bayesian modeling of anomalies due to known and unknown causes of events. Doctoral Dissertation. Intelligent Systems Program, University of Pittsburgh, 2009.

[4] Castillo E, Colosimo BM. An introduction to Bayesian inference in process monitoring, control and optimization. In: *Bayesian process monitoring, control and optimization*. Boca Raton: Chapman and Hall 2007:3-46.

[5] Press SJ. Subjective and Objective Bayesian Statistics (2nd ed.). New York: John Wiley & Sons 2003.

[6] Tuyl F, Gerlach R, Mengersen K. Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. Bayesian Analysis. 2009;4(1):151-8.

[7] Goodman SN. Toward evidence-based medical statistics 2: The Bayes factor. Annals of Internal Medicine. 1999;130(12):1005-13.

[8] Cooper GF. A Bayesian method for learning belief networks that contain hidden variables. Journal of Intelligent Information Systems. 1995;4:71-88.

[9] Neill DB, Moore AW, Cooper GF. A Bayesian spatial scan statistic. Advances in Neural Information Processing Systems. 2006;18:1003-10.