# A Bayesian Biosurveillance Method that Models Unknown Outbreak Diseases

Yanna Shen[1] and Gregory F. Cooper[1]

[1] Department of Biomedical Informatics and the Intelligent Systems Program
M-183 VALE, 200 Meyran Ave
University of Pittsburgh, Pittsburgh PA 15260
{shenyn, gfc}@cbmi.pitt.edu

**Abstract.** Algorithms for detecting anomalous events can be divided into those that are designed to detect specific diseases and those that are non-specific in what they detect. Specific detection methods determine if patterns in the data are consistent with known outbreak diseases, as for example influenza. These methods are usually Bayesian. Non-specific detection methods attempt broadly to detect deviations from some model of the non-outbreak situation, regardless of which disease might be causing the deviation. Many frequentist outbreak detection methods are non-specific. In this paper, we introduce a Bayesian approach for detecting both specific and non-specific disease outbreaks, and we report a preliminary study of the approach.

**Keywords:** anomaly detection, biosurveillance, Bayesian methods

## 1 Introduction

Detection of anomalous events in data has important applications in domains such as disease outbreak detection [1], fraud detection [2] and 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 from baseline behavior. These anomalous events can be divided into two types – those that we know about and those that are unexpected. As a result, algorithms within these monitoring systems can be classified into two categories that we will refer to as specific detection algorithms and non-specific detection algorithms. A robust detection system would use a combination of detection algorithms from both of these categories.

Specific detection algorithms look for pre-defined anomalous patterns in the data. For example, in the context of disease-outbreak detection, a specific detection approach might examine health-care data for the onset of a particular disease, such as inhalational anthrax. In contrast, a non-specific detection approach would try to detect any anomalous events that are missed by the specific detectors. By combining these two approaches we might be able to obtain a hybrid approach that detects anticipated diseases well, while having the non-specific approach serve as a "safety net" that is able to detect unanticipated (and possibly never before seen) diseases. We call this combined approach a *safety-net algorithm*. In this paper, we describe a Bayesian safety-net algorithm for detecting disease outbreaks. While the analysis in this paper is basic, we can apply the fundamental ideas to develop much richer Bayesian safety-net models and detection algorithms.

## 2 Methodology

This section introduces an example model and describes how we use the model for outbreak detection. Due to space limitations, we are only able to present an outline of the complete method.

Let $d_0$ represent all the diseases that Emergency Department (ED) patients can have in the absence of any disease outbreak in the population. Let $d_k$ denote one specific outbreak disease that we know about. If we assume that there are $n$ types of outbreak diseases, then $1 \leq k \leq n$. Finally, we use $d^*$ to represent any unknown or unexpected outbreak disease.

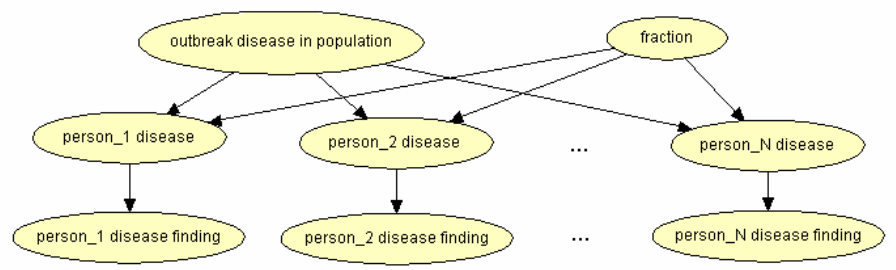


**Fig. 1.** A Bayesian network showing the population-wide disease model

The disease detection model we use is a population-wide Bayesian network model, as shown in Fig. 1, which represents all the people in the population (not just the ED patients). Let the total number of the population be $N$. Let $i$ represent the index of a specific person in the population, where $1 \leq i \leq N$. We use *pop_dx* to represent the values that node *outbreak disease in population* can take. Then *pop_dx* could be *NOB* (no outbreak), $d_k$ (outbreak of known disease $d_k$) or $d^*$ (outbreak of unknown disease $d^*$). The node *fraction* represents the fraction of the total population who have the outbreak disease (either $d_k$ or $d^*$) and visit the ED. We use $ps_i\_dx$ to denote the values that node *person_i disease* can take, which represents the possible diseases that person $i$ can have given *pop_dx*. For the people who do not come to the ED, we assign them the disease state called *none*. For a patient who comes to the ED, his or her disease state is a latent (hidden) variable.

When *pop_dx* = $d_k$ (or $d^*$), a specific person $i$ could have disease $d_0$, $d_k$, $d^*$ or *none*. The probability of person $i$ having $d_k$ (or $d^*$) is equal to the value of the *fraction* node by the construction of that node. If *pop_dx* = *NOB*, a specific person $i$ either has $d_0$ or he/she did not come to the ED, and therefore *pop_dx* = *none*. The probability that the person has $d_0$ is estimated from past ED data.

Given the disease state of person $i$, as represented by $ps_i\_dx$, we use $ps_i\_fd$ to model the state of a binary symptom of that person. The symptom state of a person is modeled using a Bernoulli distribution. It is possible to model more than one symptom, but for simplicity of presentation, we restrict this paper to an example that contains only one symptom. In particular, we consider the symptom states as being *cough* or *no cough*. For the patients that come to the ED, we define $P(ps_i\_df = cough \mid ps_i\_dx = NOB) = p_0$, $P(ps_i\_df = cough \mid ps_i\_dx = d_k) = p_k$ and $P(ps_i\_df = cough \mid ps_i\_dx = d^*) = p^*$. In the next section we describe how we model $p_0$, $p_k$ and $p^*$ under uncertainty.

### 2.1 The Disease Model

As stated, the model represents that a person has disease state $d_u$, for $0 \leq u \leq n$. Let $p_u$ denote $P(ps_i\_df = cough \mid ps_i\_dx = d_u)$. We assume that $p_u$ is distributed according to a Beta distribution, namely, $p_u \sim \text{Beta}(\alpha_u, \beta_u)$. We assessed the parameters of these Beta distributions based on real data and expert judgments.

### 2.2 The Safety-Net Model

The example safety-net model introduced in this paper is designed to detect diseases that have a probability of cough, call it $p^*$, that is not equal to $p_0, p_1, \ldots, p_n$. We will represent a distribution over $p^*$. We use $d_-$ ($d_+$) to denote a hypothetical disease that has a distribution of the probability of cough that is a delta function at 0 (1). Correspondingly, $p_- = 0$ and $p_+ = 1$. Consider every pair of known diseases being modeled, $d_u$ and $d_v$. We consider the possibility that $p^*$ is between $p_u$ and $p_v$, for every $u$ and $v$, such that $u, v \in \{-, 0, 1, \cdots n, +\}$. Each such possibility, $p_u < p^* < p_v$, constitutes one instance of the $d^*$ disease hypothesis, which we denote as $d_{uv}$.

For $d_{uv}$ we define the distribution over $p^*$ as follows. We assume that $p^*$ is uniformly distributed between $p_u$ and $p_v$, and that $p_u$ and $p_v$ are distributed as Beta($\alpha_u$, $\beta_u$) and Beta($\alpha_v$, $\beta_v$), respectively. We stochastically sample $p^*$ according to these distributions to obtain a distribution over $p^*$.

### 2.3 Inference

We wish to derive $P(pop\_dx \mid data)$, where *data* denotes the status of the symptom *cough* for every person in the population. We assume the status is either *cough* or *no cough* for people who come to the ED, and that it is always *unknown* for people who do not. We derive $P(pop\_dx \mid data)$ by deriving $P(data \mid pop\_dx)$, assessing $P(pop\_dx)$, and applying the Bayes rule.

We derive $P(data \mid pop\_dx)$ by setting $pop\_dx$ 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 we have distributions over $P(ps_i\_df \mid ps_i\_dx)$, as described in Sections 2.1 and 2.2; thus, inference includes integrating over these distributions.

We applied a variation of the inference method given in [4], which is polynomial time in the number of people who come to the ED. For the detailed description of inference, please see [5].

## 3 Preliminary Evaluation

For simplicity, we will assume that the *fraction* node has the value 0.0001 with probability 1; this assumption is not necessary, although it does reduce computational complexity.

### 3.1 Creating the Datasets

We created 20 datasets (scenarios), assuming a population size of 100,000 people. Each scenario represents data on the population for one given day of interest. In the remainder of this section, we describe how we created a scenario.

We sampled a Poisson distribution with mean $\lambda = 90$ to determine the number of people who came to the ED on the given day without any outbreak disease. For each

of these people, we sampled their cough status using the distributions defined in Section 2.1.

When simulating the presence of outbreak disease $d_k$ in the population, we assumed that the value of the node *fraction* in Fig. 1 is 0.0001. We assumed 100,000 × 0.0001 = 10 people had $d_k$ and came to the ED. For each of these 10 people, we sampled from the distribution of $P(ps_i\_df = cough \mid ps_i\_dx = d_k)$ to determine their cough status. We then combined these 10 cases with the simulated ED cases without outbreak disease in order to create a complete dataset for the scenario.

### 3.2 Experimental Setup

Let $d_u$ and $d_v$ be any two of the eight CDC Category A diseases [6], $d_u \neq d_v$. Table 1 shows the experiments for one pair of $d_u$ and $d_v$, where $d_u$ is the simulated outbreak disease. In experiment B1 (B2), there is an explicit modeling of disease $d_u$ ($d_v$), while in experiment A1 and A2 we also include safety-net disease $d^*$ in the model. In each of the four experiments, we compute the likelihood ratio (*LR*) $P(data \mid outbreak)$ / $P(data \mid non\text{-}outbreak)$ as given by Eq. 1. In experiment A1 (A2), the sum in Eq. 1 is taken over $d_u$ ($d_v$) *and* $d^*$; in contrast, in experiment B1 (B2) the sum of *pop_dx* consists only of the term $d_u$ ($d_v$). For any given outbreak disease $d_k$ being modeled, we assumed that $P(pop\_dx = d_k \mid outbreak) = P(pop\_dx = d^* \mid outbreak) = 0.5$.

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

| | A | B |
|---|---|---|
| 1 | Model $d_0$, $d_u$, $d^*$.<br>Simulate outbreak cases from $d_u$. | Model $d_0$, $d_u$.<br>Simulate outbreak cases from $d_u$. |
| 2 | Model $d_0$, $d_v$, $d^*$.<br>Simulate outbreak cases from $d_u$. | Model $d_0$, $d_v$.<br>Simulate outbreak cases from $d_u$. |

To investigate the degree to which modeling the safety-net disease $d^*$ has an impact on detection performance, we made $d_u$ and $d_v$ to be all possible pairs of the eight outbreak diseases and carried out 8 × 7 = 56 sets of experiments. In each set, we computed the mean $RR_1$ and the mean $RR_2$ over the 20 scenarios, where for a given scenario $RR_1 = LR(S_{B1}) / LR(S_{A1})$ and $RR_2 = LR(S_{A2}) / LR(S_{B2})$, where $S_{A1} = \{d_u, d^*\}$, $S_{B1} = \{d_u\}$, $S_{A2} = \{d_v, d^*\}$, and $S_{B2} = \{d_v\}$. Our hypothesis is that usually $RR_2 > RR_1$, which supports that modeling $d^*$ is doing more good than harm in detecting outbreak diseases.

$$LR = \frac{\sum_{pop\_dx} P(data \mid pop\_dx) P(pop\_dx \mid outbreak)}{P(data \mid pop\_dx = d_0)} \quad (1)$$

## 4 Results and Discussion

Recall that the distribution of $P(ps_i\_df = cough \mid ps_i\_dx = d_k)$ was assessed by a domain expert, for $1 \leq k \leq 8$. We sorted the eight outbreak diseases by their expectations for $P(ps_i\_df = cough \mid ps_i\_dx = d_k)$ in ascending order. In Tables 2, 3 and 4, each row represents a disease $d_u$ and each column represents a disease $d_v$. We list the eight diseases according to the sorted order from left to right and from up and

down. The closer $d_u$ and $d_v$ are to the diagonal, the closer are their means, and in a sense the closer the two diseases are in their symptomatic presentation.

**Table 2.** The mean $RR_2$ given different $d_u$ (rows) and $d_v$ (columns).

| $d_u$ \ $d_v$ | small pox | cryptospo -ridiosis | early anthrax | late anthrax | asthma | influenza | early plague | late plague |
|---|---|---|---|---|---|---|---|---|
| small pox | - | 0.84 | 0.84 | 1.94 | 2.88 | 2.33 | 4.65 | 6.74 |
| cryptospo -ridiosis | 0.90 | - | 0.84 | 1.66 | 2.27 | 1.98 | 3.42 | 4.62 |
| early anthrax | 1.05 | 0.97 | - | 1.40 | 1.77 | 1.64 | 2.48 | 3.15 |
| late anthrax | 2.13 | 1.90 | 1.38 | - | 0.91 | 0.93 | 1.02 | 1.11 |
| asthma | 2.54 | 2.29 | 1.59 | 0.82 | - | 0.84 | 0.87 | 0.91 |
| influenza | 2.75 | 2.54 | 1.72 | 0.82 | 0.82 | - | 0.88 | 0.92 |
| early plague | 2.76 | 2.51 | 1.70 | 0.80 | 0.78 | 0.81 | - | 0.84 |
| late plague | 3.35 | 3.14 | 2.04 | 0.77 | 0.73 | 0.75 | 0.75 | - |

Table 2 shows the mean $RR_2$ given different combinations of $d_u$ and $d_v$. The mean $RR_2$ tends to increase from the diagonal to the top right and to the bottom left corners. It shows that when there is an unexpected disease $d_u$ present, the greater the difference in presentation of $d_v$ relative to $d_u$, the greater the expected benefit from modeling $d^*$.

**Table 3.** The mean $RR_1$ given different $d_u$.

| $d_u$ \ $d_v$ | small pox | cryptospo -ridiosis | early anthrax | late anthrax | asthma | influenza | early plague | late plague |
|---|---|---|---|---|---|---|---|---|
| small pox | - | 1.21 | 1.21 | 1.21 | 1.21 | 1.21 | 1.21 | 1.21 |
| cryptospo -ridiosis | 1.18 | - | 1.18 | 1.18 | 1.18 | 1.18 | 1.18 | 1.18 |
| early anthrax | 1.14 | 1.14 | - | 1.14 | 1.14 | 1.14 | 1.14 | 1.14 |
| late anthrax | 1.15 | 1.15 | 1.15 | - | 1.15 | 1.15 | 1.15 | 1.15 |
| asthma | 1.24 | 1.24 | 1.24 | 1.24 | - | 1.24 | 1.24 | 1.24 |
| influenza | 1.21 | 1.21 | 1.21 | 1.21 | 1.21 | - | 1.21 | 1.21 |
| early plague | 1.23 | 1.23 | 1.23 | 1.23 | 1.23 | 1.23 | - | 1.23 |
| early plague | 1.34 | 1.34 | 1.34 | 1.34 | 1.34 | 1.34 | 1.34 | - |

Table 3 shows the mean $RR_1$. Since there is no disease $d_v$ involved in deriving $RR_1$, every row has the same values. Table 3 shows that $RR_1$ is quite stable at a value only modestly greater than 1, which provides support that when there is no unanticipated disease present, modeling $d^*$ only weakly degrades the detection of $d_u$.

We performed the sign tests to calculate P-values over the null hypothesis $H_o$: $RR_1 > RR_2$ versus the alternative hypothesis $H_a$: $RR_1 \leq RR_2$. Table 4 shows the P-values given different combinations of $d_u$ and $d_v$. Notice that the P-values close to the diagonal are very big, so that we cannot reject the null hypothesis, while the P-values

away from the diagonal are zeros, which rejects the null hypothesis. Table 4 provides support that modeling $d^*$ helps detect unanticipated diseases more than it interferes with detecting known diseases.

**Table 4.** A table shows the P-values given different combinations of $d_u$ and $d_v$.

| $d_v$ / $d_u$ | small pox | cryptospo -ridiosis | early anthrax | late anthrax | asthma | influenza | early plague | late plague |
|---|---|---|---|---|---|---|---|---|
| small pox | - | 0.99 | 0.99 | 0 | 0 | 0 | 0 | 0 |
| cryptospo -ridiosis | 0.99 | - | 1 | 0 | 0 | 0 | 0 | 0 |
| early anthrax | 0.99 | 0.99 | - | 0 | 0 | 0 | 0 | 0 |
| late anthrax | 0 | 0 | 0 | - | 0.99 | 0.99 | 0.99 | 0.99 |
| asthma | 0 | 0 | 0 | 1 | - | 1 | 1 | 1 |
| influenza | 0 | 0 | 0 | 1 | 1 | - | 0.99 | 0.98 |
| early plague | 0 | 0 | 0 | 1 | 1 | 1 | - | 1 |
| late plague | 0 | 0 | 0 | 1 | 1 | 1 | 1 | - |

## 5 Conclusion and Future Work

This paper introduced a Bayesian method for detecting disease outbreaks that combines a specific detection method with a non-specific method. Preliminary results provide support that this hybrid approach helps detect unexpected diseases more than it interferes with detecting known diseases.

We plan to test this approach on real datasets and evaluate its detection performance using other measures, such as AMOC curves [2].

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

## References

1. Wong, W.-K., *Data mining for early disease outbreak detection [Doctoral Dissertation]*. 2004, Carnegie Mellon University: Pittsburgh.
2. Fawcett, T. and F. Provost, *Adaptive fraud detection.* Data Mining and Knowledge Discovery, 1997. **1**(3): p. 291-316.
3. Denning, D., *An intrusion-detection model.* IEEE Transactions on Software Engineering, 1987.
4. Cooper, G.F., *A Bayesian method for learning belief networks that contain hidden variables.* Journal of Intelligent Information Systems, 1995. **4**: p. 71-88.
5. Shen, Y. and G.F. Cooper, *Bayesian disease outbreak detection that includes a model of unknown disease*. Department of Medical Informatics, University of Pittsburgh: Technical Report No. DBMI-07-351, 2007.
6. Hutwagner, L., W. Thompson, and G.M. Seeman, *The bioterrorism preparedness and response early aberration reporting system (EARS).* Journal of Urban Health, 2003. **80**(2, Supplement 1): p. i89-i96.