A New Prior for Bayesian Anomaly Detection

Application to Biosurveillance

Y. Shen[1]; G. F. Cooper[2]
[1]Lister Hill National Center for Biomedical Communications, National Institute of Health, Bethesda, Maryland, USA;
[2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Keywords
Anomaly detection, biosurveillance, Bayesian methods, prior probability distributions, efficient inference

Summary

Objectives: Bayesian anomaly detection computes posterior probabilities of anomalous events by combining prior beliefs and evidence from data. However, the specification of prior probabilities can be challenging. This paper describes a Bayesian prior in the context of disease outbreak detection. The goal is to provide a meaningful, easy-to-use prior that yields a posterior probability of an outbreak that performs at least as well as a standard frequentist approach. If this goal is achieved, the resulting posterior could be usefully incorporated into a decision analysis about how to act in light of a possible disease outbreak.

Methods: This paper describes a Bayesian method for anomaly detection that combines learning from data with a semi-informative prior probability over patterns of anomalous events. A univariate version of the algorithm is presented here for ease of illustration of the essential ideas. The paper describes the algorithm in the context of disease-outbreak detection, but it is general and can be used in other anomaly detection applications. For this application, the semi-informative prior specifies that an increased count over baseline is expected for the variable being monitored, such as the number of respiratory chief complaints per day at a given emergency department. The semi-informative prior is derived based on the baseline prior, which is estimated from using historical data.

Results: The evaluation reported here used semi-synthetic data to evaluate the detection performance of the proposed Bayesian method and a control chart method, which is a standard frequentist algorithm that is closest to the Bayesian method in terms of the type of data it uses. The disease-outbreak detection performance of the Bayesian method was statistically significantly better than that of the control chart method when proper baseline periods were used to estimate the baseline behavior to avoid seasonal effects. When using longer baseline periods, the Bayesian method performed as well as the control chart method. The time complexity of the Bayesian algorithm is linear in the number of the observed events being monitored, due to a novel, closed-form derivation that is introduced in the paper.

Conclusions: This paper introduces a novel prior probability for Bayesian outbreak detection that is expressive, easy-to-apply, computationally efficient, and performs as well or better than a standard frequentist method.

Correspondence to:
Y. Shen
Lister Hill National Center for Biomedical Communications
Building 38A, 9N912A
National Institute of Health
Bethesda, Maryland 20894
USA
E-mail: yanna.shen@nih.gov

Methods Inf Med 2010; 49: 44–53
doi: 10.3414/ME09-01-0008
received: February 3, 2009
accepted: July 6, 2009
prepublished: December 21, 2009

1. Introduction

Detection of anomalous events in data is a research area with important applications in domains such as disease-outbreak detection [1], clinical treatment monitoring [2], fraud detection [3], and intrusion detection [4]. 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 the Shewhart control chart method [5], CuSum [6], and EWMA [7], use frequentist statistical techniques that derive statistics, such as *p* values.

In this paper we introduce a Bayesian anomaly-detection algorithm for monitoring broad patterns of anomalies that manifest increased rates of some event (e.g., emergency department (ED) patients with a *fever* symptom)[a]. The algorithm operates on a univariate time series of counts of an event X being monitored. We call this method the Bayesian univariate (BU) algorithm. The BU algorithm models the expected baseline rate of X during the previous 24-hour period using historical data that are assumed to contain no anomalous patterns. It uses the baseline rate to derive a semi-informative prior that represents the expected increased rates of X if there is an anomaly. This prior characterizes anomalous patterns of X. The models for non-anomalous and anomalous patterns of X are used to derive a posterior probability of an anomaly.

This paper describes an example of the BU algorithm in the context of disease outbreak detection. The example algorithm operates on a univariate time series of ED

[a] Decreased rates can be handled in an analogous fashion.

chief complaint data. A chief complaint is a short phrase that describes the primary reason a patient came to the ED (e.g., "severe headache"). Such data are often captured electronically in real time when a triage nurse sees a patient. The example application of the algorithm that is presented here takes as input the number of ED *respiratory* chief complaints (e.g., cough, shortness of breath, sputum production, etc.) in the previous 24-hour period, and it outputs a posterior probability of a respiratory disease outbreak being present during that period. It could be run each hour, for example, to provide an ongoing assessment of the presence of such an outbreak. The algorithm itself can be generalized to other applications in the area of anomaly detection, as for example monitoring whether a medication used to treat a given disease has changed relative to its use in the recent past.

One technical contribution of this paper is the derivation of a closed-form solution to the posterior probability of an anomaly, given the prior probability distribution that we introduce. The time complexity of calculating this posterior probability is just linear in the number of the observed events X during the period being monitored.

We compared the detection performance of the BU algorithm with a standard control chart algorithm, which also uses recent data to detect an anomaly. We hy pothesized that BU would perform disease outbreak detection at least as well as the control chart method. If so, the BU algorithm will have the advantage that it generates a posterior probability that can be applied directly in decision analyses, unlike what is possible with the control chart method and other frequentist approaches. For example, an analysis could be done of whether it is warranted for public health officials to investigate a potential disease outbreak.

The remainder of this paper is organized as follows. In the next section, we describe the main issues of the BU algorithm. We then present an example application of the algorithm, describe the experiments we performed, and show the results. We provide additional discussion of the algorithm and suggest several ways of extending it. We also describe previously published related work. We conclude with a brief summary.

2. Background

We model the expected baseline rate of event X during the previous 24-hour period using a Beta distribution. The Beta distribution is a continuous probability distribution that is parameterized by two positive shape parameters. The Beta distribution has been used for a wide variety of applications because it can take a very diverse set of shapes [8].

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.

There is a long history of using the Beta distribution to represent belief about a relative frequency. In the 19th century G. F. Hardy [9] and W. A. Whitworth [10] proposed quantifying prior beliefs with Beta distributions. In addition, a Beta(1, 1), called the Bayes-Laplace prior, is often applied as a non-informative prior of some rate or proportion of interest [11].

In order to estimate the two parameters of the Beta distribution, we applied the method of moment matching, which is a popular means of parametric density estimation [12–14]. It works by matching the first two sample moments (e.g., the mean and variance) to the corresponding population moments and solving the resulting equations for the parameters to be estimated [15].

Some researchers have used a Gamma prior as a representation for disease rate in epidemiology, such as the research work by Clayton and Kaldor [16, 17]. In the domain of disease-outbreak detection, Neill et al. [18] used a Gamma-Poisson model for Bayesian disease-outbreak detection. In particular, they assume that the number of disease cases are Poisson-distributed, where the underlying disease rate is modeled using a Gamma prior distribution. They computed the posterior probability of an outbreak to monitor for potential disease outbreaks.

3. Methods

In this section, we describe the BU algorithm for monitoring disease outbreaks that manifest as increased rates of respiratory ED chief complaints. The term ED that is used below refers to one or more emergency departments in the region being monitored. If more than one, then the total patient cases across all EDs are treated as a single pool.

We first introduce the following notation:

Let N_p be the total number of people in the population in a specific region being monitored for an outbreak of respiratory disease.

Let N_1 be the number of people from the monitored region who came to the ED in the most recent 24 hours with a respiratory chief complaint.

Let $N_2 = Np - N_1$ represent the number of people in the monitored region who during the most recent 24-hour period did not visit the ED or who visited the ED with a non-respiratory chief complaint.

We use *OB* to denote the state of an outbreak existing during the most recent 24-hour period in the region being monitored and *NOB* to represent the absence of any disease outbreak during that period. In particular, *OB* here will represent there being a respiratory outbreak in the population. Note that *OB* and *NOB* are mutually exclusive and exhaustive, and thus, $P(OB) + P(NOB) = 1$.

Let θ_{OB} denote the fraction of people (out of the total population N_p) who during the most recent 24-hour period visited the ED with a respiratory chief complaint, when there is a respiratory outbreak in the population. Thus, $0 \leq \theta_{OB} \leq 1$. We denote the density of θ_{OB} as $f(\theta_{OB})$. Note that θ_{OB} represents the total fraction of people with a respiratory chief complaint who visited the ED, including those with the outbreak disease and those without it.

Similarly, let θ_{NOB} represent the fraction of people (out of the total population N_p) who during the most recent 24-hour period visited the ED with a respiratory chief complaint, when there is not a respiratory outbreak in the population. Thus, $0 \leq \theta_{NOB} \leq 1$. We denote the density of θ_{NOB} as $g(\theta_{NOB})$.

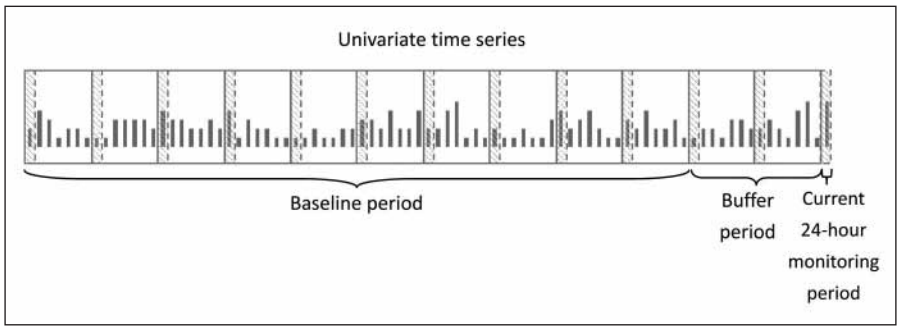

Fig. 1 The sliding buffer concept used in the BU algorithm. Time is displayed on the horizontal axis and the number of respiratory chief complaints per day is displayed on the vertical axis. The small, shaded, horizontal areas correspond to a Monday of each week, which is the current day being monitored. The baseline period is shown as being ten weeks long. The buffer period is two weeks.

3.1 The Prior Distribution

We model θ_{NOB} using a Beta distribution, namely, $g(\theta_{NOB}) \sim \text{Beta}(\theta_{NOB}; \alpha_0, \beta_0)$. We dynamically estimate the parameters α_0 and β_0 from past ED data that are assumed to contain no respiratory disease outbreaks. In particular, we use the sliding buffer concept [19], which separates the past data into two segments: a buffer period and a baseline period, as shown in ▶Figure 1.

The baseline period is used to statistically characterize the patterns of non-outbreak respiratory cases. We will be using the mean and variance of the counts per day during the baseline period to estimate parameters α_0 and β_0. The more recent buffer period is inserted to avoid contamination of the baseline period with a potential outbreak signal during monitoring.

We used the independently developed BARD system to generate the simulated anthrax cases [20]. BARD models in significant detail the dispersion of spores from an aerosolized release of anthrax and the subsequent infection and health-seeking behavior of the exposed population. We evaluated the BU algorithm on datasets that we developed by overlaying simulated anthrax cases onto a background of real ED cases. Each dataset is assumed to contain only one outbreak, where the outbreak duration is assumed to last up to two weeks. We use a buffer period of two weeks to avoid training the baseline model on outbreak cases. In order to accommodate seasonal effects in the data, we use a total of 12 weeks for the baseline period and the buffer period. Thus, we use a baseline period of 10 weeks, as shown in ▶Figure 1.

In particular, we use the same day-of-week data from the past 10 weeks in the baseline period to estimate the mean μ_0 and the variance σ_0^2 of θ_{NOB} for each day of the week, because experience shows that the number of respiratory counts is sensitive to the day of the week. Thus, for example, we will have 10 Mondays on which to estimate the mean and variance for respiratory counts on that day of the week. ▶Figure 1 shows this example, where the current 24-hour monitoring period is currently Monday. Finally, we use the method of moment matching in [15] to estimate parameters α_0 and β_0 from the mean μ_0 and variance σ_0^2 using the equations shown below:

$$\mu_0 = \frac{\alpha_0}{\alpha_0 + \beta_0},$$

$$\alpha_0 + \beta_0 = \frac{\mu_0(1-\mu_0)}{\sigma_0^2} - 1.$$

The BU algorithm is a population-based outbreak-detection algorithm, which models the fraction of people (out of the total population) who visited the ED with a respiratory chief complaint during the last 24 hours. If an outbreak is occurring during this period, we assume that this fraction (θ_{OB}) is at least as large as the fraction in the absence of any disease outbreak (baseline fraction θ_{NOB}), namely, $\theta_{OB} \geq \theta_{NOB}$. As mentioned above, θ_{OB} is capturing the joint rate of both non-outbreak and outbreak cases of respiratory disease; the subscript "*OB*" means the joint rate that includes outbreak cases. In θ_{NOB}, there are no outbreak cases, so this rate only includes non-outbreak ones. This assumption will not hold if a disease outbreak would influence a large number of respiratory patients with non-outbreak diseases to avoid visiting the emergency department. But, such exceptions seem unlikely, especially early in the outbreak, which is when we most want to detect it and yet its presence is not yet generally known.

We model θ_{OB} as an increased rate of respiratory chief complaints relative to θ_{NOB}, and thus, we have the prior constraint that $\theta_{OB} \geq \theta_{NOB}$. We assume that we are indifferent to the value of θ_{OB} other than that $\theta_{OB} \geq \theta_{NOB}$. Therefore, θ_{OB} is assumed to have a uniform distribution on the interval $[\theta_{NOB}, 1]$. Because θ_{OB} is only constrained to be greater than θ_{NOB}, and otherwise it is uniform, we call this a semi-informative prior probability on θ_{OB}.

Let $h(\theta_{OB}, \theta_{NOB})$ denote the joint probability density on θ_{OB} and θ_{NOB}, and let $h(\theta_{OB} \mid \theta_{NOB})$ denote their conditional density. By integrating over all possible values of θ_{NOB} ($0 \leq \theta_{NOB} \leq \theta_{OB}$), we can represent the prior distribution of θ_{OB} as follows:

$$f(\theta_{OB}) = \int_0^1 h(\theta_{OB}, \theta_{NOB}) d\theta_{NOB} = \int_0^{\theta_{OB}} h(\theta_{OB}, \theta_{NOB}) d\theta_{NOB},$$

because $h(\theta_{OB}, \theta_{NOB}) = 0$ when $\theta_{NOB} > \theta_{OB}$

$$= \int_0^{\theta_{OB}} h(\theta_{OB} \mid \theta_{NOB}) \cdot g(\theta_{NOB}) d\theta_{NOB} = \int_0^{\theta_{OB}} \frac{1}{1-\theta_{NOB}} g(\theta_{NOB}) d\theta_{NOB}, \quad (1)$$

where the term $1/(1-\theta_{NOB})$ represents the density of a uniform distribution of θ_{OB} on the interval $[\theta_{NOB}, 1]$ given a specific value of θ_{NOB}. ▶Figure 2 shows an example of a prior distribution $g(\theta_{NOB})$ and the derived prior distribution $f(\theta_{OB})$.

Let $B(u,v)$ represent a *Beta function* such that

$$B(u,v) = \int_0^1 t^{u-1}(1-t)^{v-1} dt,$$

which has the following solution:

$$B(u,v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}.$$

Next, let $B(z; u,v)$ represent an *incomplete Beta function* [21] such that

$$B(z;u,v) = \int_0^z t^{u-1}(1-t)^{v-1}\,dt \quad \text{for } 0 \le z \le 1.$$

Since we model θ_{NOB} using a Beta distribution as shown in the equation in ▶Figure 3, Equation 1 can therefore be written as the equation shown in ▶Figure 4, where $B(\theta_{OB}; \alpha_0, \beta_0 - 1)$ is an incomplete Beta function.

3.2 Inference

We wish to derive $P(OB|E)$, where OB represents there being a respiratory-disease outbreak in the population in the last 24 hours, and E denotes evidence, which in our example is the respiratory chief-complaint status of patients who came to the ED during the previous 24 hours. We derive $P(OB|E)$ by deriving $P(E|OB)$, assessing $P(OB)$, and applying Bayes' rule.

We assume that the status of *respiratory* for every person in the population is a sequence of independent Bernoulli trials with proportion parameter θ_{NOB} when there is no outbreak in the population. Given the non-outbreak situation in the last 24 hours, as denoted by NOB, we represent the likelihood $P(E|NOB)$ using the Bernoulli-Beta model as follows.

$$P(E \mid NOB) = \int_0^1 \theta_{NOB}^{N_1}(1-\theta_{NOB})^{N_2} g(\theta_{NOB})\,d\theta_{NOB}, \quad (3)$$

where $g(\theta_{NOB}) \sim Beta(\theta_{NOB}; \alpha_0, \beta_0)$. ▶Equation 3 has the following well known closed-form solution [22]:

$$B(\alpha_0 + N_1, \beta_0 + N_2) \,/\, B(\alpha_0, \beta_0),$$

where B represents a Beta function.

Given a situation where an outbreak is occurring, we assume that we can model whether a person in the population arrives in the ED with a respiratory chief complaint as a Bernoulli trial with proportion parameter θ_{OB}. We then model the chief complaint status of the population as a set of independent, identically distributed Bernoulli trials. We therefore derive $P(E|OB)$ as follows:

$$P(E \mid OB) = \int_0^1 \theta_{OB}^{N_1}(1-\theta_{OB})^{N_2} f(\theta_{OB})\,d\theta_{OB}, \quad (4)$$

where $f(\theta_{OB})$ is the prior distribution shown in ▶Equation 2. The closed-form solution to ▶Equation 4 is derived in the ▶Appendix as being shown in ▶Figure 5.

We compute the posterior probability of a disease outbreak using Bayes' rule as follows:

$$P(OB \mid E) = \frac{P(E \mid OB)P(OB)}{P(E \mid OB)P(OB) + P(E \mid NOB)P(NOB)}. \quad (6)$$

The time complexity for computing $P(E|OB)$ using ▶Equation 5 is $O(N_1)$, where N_1 is the number of people who came to the ED in the most recent 24 hours with a respiratory chief complaint. Computing $P(E|NOB)$ using ▶Equation 3 requires constant time. Therefore, computing the posterior probability $P(OB|E)$ using ▶Equation 6 requires time that is linear in N_1.

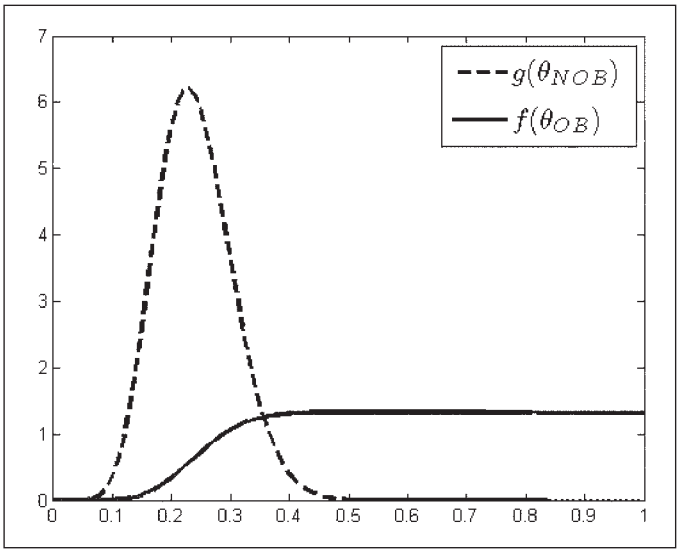

Fig. 2 Plots showing examples of $g(\theta_{NOB})$ and $f(\theta_{OB})$ with the range for θ_{NOB} and θ_{OB} from 0 to 1, as shown in the abscissa

$$g(\theta_{NOB}) = Beta(\theta_{NOB}; \alpha_0, \beta_0) = \frac{1}{B(\alpha_0, \beta_0)}\theta_{NOB}^{\alpha_0 - 1}(1-\theta_{NOB})^{\beta_0 - 1}$$

Fig. 3 An equation showing the density function of θ_{NOB}

$$\begin{aligned} f(\theta_{OB}) &= \int_0^{\theta_{OB}} \frac{1}{1-\theta_{NOB}} \frac{1}{B(\alpha_0,\beta_0)} \theta_{NOB}^{\alpha_0-1}(1-\theta_{NOB})^{\beta_0-1}\,d\theta_{NOB} \\ &= \int_0^{\theta_{OB}} \frac{1}{B(\alpha_0,\beta_0)} \theta_{NOB}^{\alpha_0-1}(1-\theta_{NOB})^{(\beta_0-1)-1}\,d\theta_{NOB} \\ &= B(\theta_{OB}; \alpha_0, \beta_0 - 1)/B(\alpha_0, \beta_0) \end{aligned}$$

Fig. 4 An equation (Eq. 2) showing the density function of θ_{OB}

$$P(E \mid OB) = \frac{1}{B(\alpha_0, \beta_0)} \sum_{n=0}^{N_1} \binom{N_1}{n} B(N_2 + 1, n + 1) B(N_1 - n + \alpha_0, N_2 + n + \beta_0)$$

Fig. 5 An equation (Eq. 5) showing the likelihood of evidence *E* when there is a respiratory-disease outbreak in the population in the last 24 hours where *E* represents the respiratory chief-complaint status of patients who came to the ED during the last 24 hours

4. Evaluation

In this section, we describe an example application of the BU algorithm introduced in the previous section. We first describe the experiments used to evaluate the algorithm and then show the experimental results.

4.1 Creating the Datasets

We obtained a background time series of actual chief complaints from a set of EDs in Allegheny County, Pennsylvania in 2001 and 2002, in which the background time series from 2002 was used for testing, and 12-week data prior to each day of 2002 were used for training the baseline model for each day of 2002, i.e., served as the baseline period and the buffer period. We assumed there was no anthrax outbreak occurring during the period that we used to estimate non-outbreak parameters. All personal identifying information was removed from these actual ED cases, and this research was approved by the University of Pittsburgh IRB. We evaluated the BU algorithm on semi-synthetic datasets produced by overlaying the simulated anthrax cases onto the real background time series of ED cases, where the simulated cases of anthrax were produced by the 2004 version of the BARD simulator [23].

BARD uses a Gaussian plume model and weather conditions to estimate the distribution of spore concentrations over a geographic region. Based on the spore concentrations in a given zip code, BARD uses a clinical model to simulate the number of patients who will contract inhalational anthrax over time and present to the ED with respiratory chief complaints. In particular, BARD produces a simulated outbreak *scenario* consisting of a list of simulated anthrax cases, where each case consists of 1) a date-time field when the patient presented to the ED with a respiratory chief complaint, and 2) the patient's home zip code.

BARD was used to generate 96 simulated anthrax-release scenarios, each with a unique combination of release date, wind direction, wind speed, release location, release height, and quantity of spores released. In all scenarios, we assumed that 1.0 kg of anthrax spores was released. For each month in 2002, eight random release times were selected for use by the simulator. We created one semi-synthetic dataset by overlaying the simulated outbreak cases produced by BARD onto the real ED cases starting from the release date. Thus, a total of $8 \times 12 = 96$ different datasets were generated.

The population size that is covered by the EDs, whose data we used in this evaluation, is approximately four hundred thousand, that is, $N_p \approx 400{,}000$.

We close this section with an explanation for why we used simulated outbreak data in these initial experiments, rather than real outbreak data. For instance, we could run the BU algorithm using real influenza outbreak data. However, there is no reliable gold standard regarding the date and time that such outbreaks occur. Thus, there are downsides to evaluating BU using real data. Although there also are limitations in using simulated data, rather than real data, using simulated data allows 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 has frequently been used in research that evaluates biosurveillance algorithms. We believe simulated data provides a useful approach to performing an initial set of experiments. In future work, it would be worthwhile to evaluate these algorithms using real data as well.

4.2 Experimental Methods

We compared the detection performance of the BU algorithm with the detection performance of the control chart (CC) method because both BU and CC 1) take as input data from recent observations, such as ED respiratory chief complaints from the most recent 24-hour period, 2) monitor for an increased respiratory count relative to the baseline count[b], and 3) estimate the characteristics of baseline behavior using the same method described below.

The control chart method typically models the non-anomalous events as a Gaussian distribution [5]. It consists of a center line, which is drawn as the non-anomaly process mean, plus upper and lower control limits that indicate the threshold at which the process output is out of control. According to [5], the control chart methods do not perform well when less than 20 data samples from the in-control process (non-outbreak condition in the context of disease-outbreak detection) are used to estimate the baseline behavior. Let w represent the baseline period (in weeks) that we use to estimate the mean and variance of θ_{NOB}, as described in Section 3.1. In order to perform a balanced comparison between BU and CC, we used a sequence of values for w, namely $w = \{2, 6, 10, 14, 18, 22\}$. For each value of w, we implemented the control chart method and the BU algorithm as follows.

CC takes as input the most recent 24-hour ED respiratory chief-complaint counts, and outputs a Z score[c] for observing that many counts, relative to a mean μ_0 and a standard deviation σ_0. The parame-

[b] Traditional control chart methods that are used in quality control also look for decreased counts of some event X relative to the baseline count of X. In contrast, in the domain of disease outbreak detection, we are often interested in increased counts of X. Thus the control chart method described in this paper only looks for increased counts as described in this section.

[c] If the observed respiratory count is smaller than the mean μ_0, the z-Score is zero; otherwise, the z-Score is $z = (x - \mu_0)/\sigma_0$, where x is the observed respiratory count.

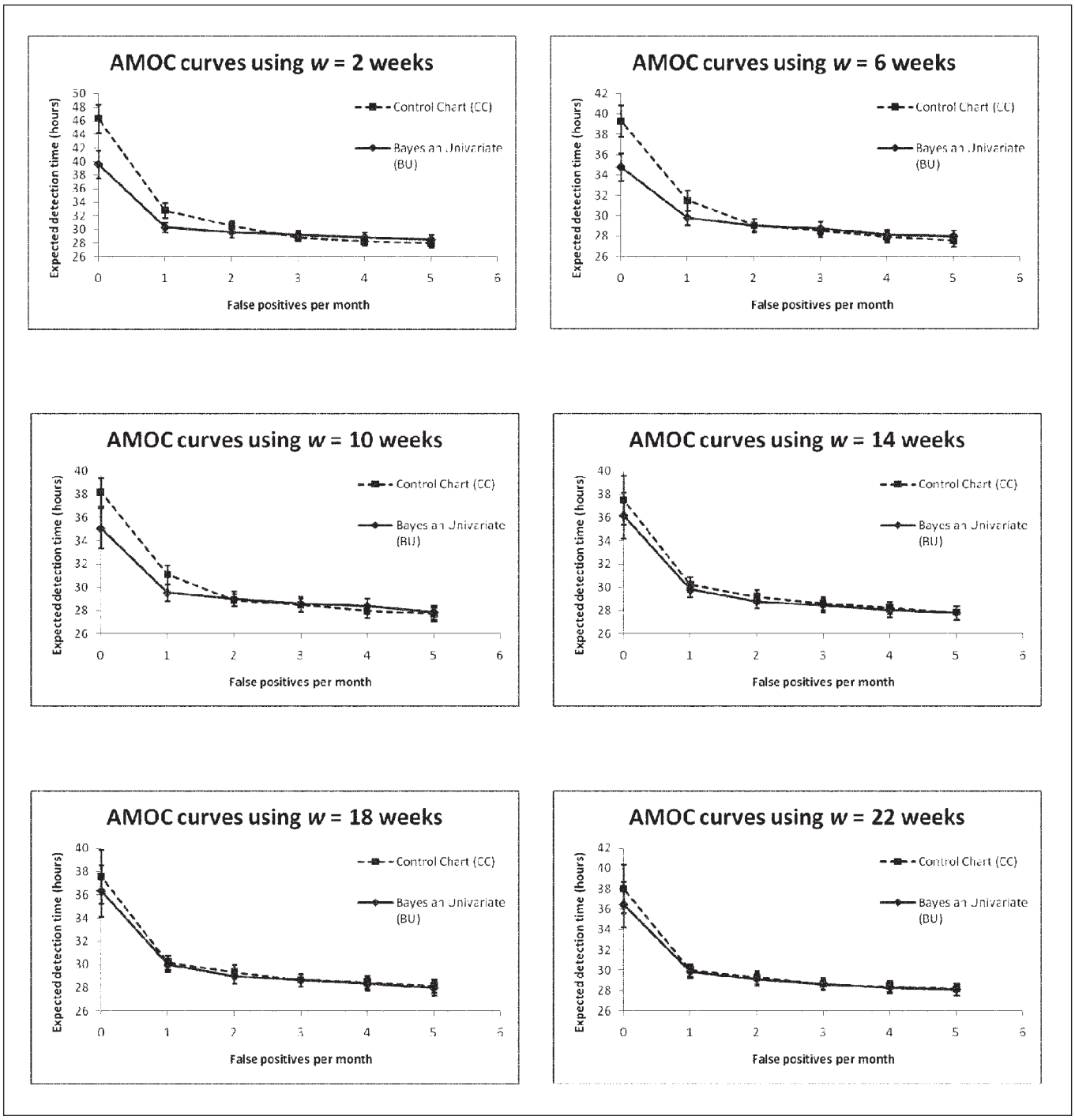

Fig. 6 AMOC curves of BU and CC from using $w = 2, 6, 10, 14, 18$, and 22 weeks, respectively. In each curve of the subplot, the marker on the curve shows the expected detection time under a specific false positive rate, and the bar shows the 95% confidence interval of the expected detection time.

ters μ_0 and σ_0 are estimated from the previous w weeks of ED data, using the sliding buffer method described in Section 3.1.

In applying ▶Equation 6, we use $P(OB) = 0.01$ and $P(NOB) = 0.99$. The particular choice of these prior probabilities does not affect the detection performance of the BU algorithm that is reported in the next section, as long as they are not 0 or 1. However, the choice does affect the absolute magnitude of the posterior probabilities that are output by the BU algorithm.

For each value of w, we plot an AMOC curve for BU and CC. An AMOC curve shows the expected detection time as a function of the false positive rate [24]. The false-positive rate and the detection time of each algorithm are determined using the method described below.

For each of the two algorithms, which we denote as algorithm A, we applied A on

Table 1 Mean difference of the expected detection time (hours) between BU and CC and the 95% confidence interval of that mean difference

	$w = 2$	$w = 6$	$w = 10$	$w = 14$	$w = 18$	$w = 22$
0 false positives per month	6.7 (2.5, 10.9)	4.6 (1.8, 7.4)	3.2 (0.2, 6)	1.4 (-2.7, 5.5)	1.3 (-3.2, 5.8)	1.5 (-3.1, 6.1)
1 false positive per month	2.5 (0.7, 4.3)	1.7 (0.1, 3.3)	1.6 (0.1, 3.1)	0.4 (-1, 1.8)	0.2 (-1, 1.4)	0.2 (-1.1, 1.5)

the background time series of actual ED cases from 2002 that presumptively contain no anthrax disease outbreaks in order to determine its false positive rates under various detection thresholds. In particular, a sequence of thresholds for evaluating A was obtained by sorting the probabilistic outputs from running A on the background time series, with one probability output per day. From these sorted probabilities, we retained the unique ones, and used this sequence of probabilities as detection thresholds for A. Note that the set of thresholds for BU can be different from those for CC due to their different outputs. Using a detection threshold r in the threshold set for A, we can determine its false positive rate under the threshold r, as described below. Every threshold in the threshold set for A was used.

The false positive rate was derived as FP/M, where $M = 12$ months and FP is the number of false positives that occurred using a detection threshold r, when the background ED data (which is real data with no simulated or real outbreaks) was monitored by the given algorithm A during a 12-month period.

We then ran algorithm A on each of the 96 semi-synthetic datasets to determine its expected detection times under the set of thresholds for A. The detection time of algorithm A for a specific semi-synthetic dataset (which is one of the 96 datasets containing simulated anthrax cases) is the time from the simulated anthrax release until the threshold r was crossed by the output of A. The expected detection time of A is the average detection time over the 96 datasets.

4.3 Results

This section shows the AMOC curves for different lengths w of the baseline period. We report results from zero to five false positives per month.

▶Figure 6 shows a set of AMOC curves for BU and CC, in which the baseline period w is 2 to 22 weeks. BU has relatively better detection performance when using baseline periods in which $w = 2$ to 10 weeks.

For CC the baseline periods of $w = 18$ to 22 weeks result in relatively better detection performance, which suggests that a relatively large amount of baseline data needs to be used in order to construct an effective and robust control chart method.

At zero and one false positive per month, BU has an expected detection time that is less than that of CC over all values of w. The maximum detection-time gain of BU over CC is approximately 6.7 hours under zero false positives per month when using a baseline period $w = 2$ weeks. At two to five false positives per month and for all values of w, CC performs as well as BU, as shown in ▶Figure 6.

We performed one-sided paired t-test using the detection times of BU and CC over the 96 outbreak datasets used in the evaluation. We report a statistical analysis of detection times at zero and one false positive per month because one false positive per month is often considered as an upper bound on a tolerable false-positive rate in the domain of disease outbreak detection. Statistical analyses show that, at a significance level of 0.05, BU detects the outbreak statistically significantly faster than CC at zero and one false positive per month when using a baseline period $w = 2$, 6, and 10 weeks. For higher values of w that were evaluated, the detection times of BU and CC do not show statistically significant differences. Overall, the results provide support that BU performs disease-outbreak detection at least as well as the control chart method.

▶Table 1 shows the mean difference of the expected detection time of CC over BU and the 95% confidence interval of the mean difference when using a sequence of values for the length w of the baseline period.

4.4 Discussion

Researchers have estimated that each hour earlier that an anthrax outbreak is detected can save as much as $200M in economic costs [25, 26] and more importantly can save many lives, because antibiotics could be started sooner in individuals who were exposed.

A given algorithm (BU, CC, or other algorithms) may generally decrease the outbreak detection time, but will also generally lead to false positive alerts. Such alerts would have a cost in that public health officials would need to deal with them, which might involve time-consuming investigations. Thus, deciding the appropriate threshold to use is important. Decision theory provides a coherent approach for trading off these costs and benefits to establish the detection threshold to use [27]. Since BU outputs a posterior probability, this output can be directly used as a key component in such a decision analysis.

5. Extensions

In this section, we discuss several possible ways of extending the BU algorithm. Recall that we model the prior distribution of θ_{NOB}, namely $g(\theta_{NOB})$, using the same day-of-week data from the past 12 weeks, which incorporates day-of-week effects. More generally, we can derive a conditional version of $g(\theta_{NOB} \mid PV)$, where PV is a predictor variable that might include day-of-week, season-of-year, holiday, and other effects.

The current BU algorithm monitors for increased counts of some event X relative to the baseline count, and the prior distribution of θ_{OB} was derived to characterize such an anomalous pattern, i.e., $\theta_{OB} \geq \theta_{NOB}$. In some types of anomaly detection, a decrease in the counts of some variable may suggest an anomalous event. For example, in fraud detection, a decrease in credit card purchases

near the card owner's home town may suggest that the card has been stolen. It would be useful to derive a prior distribution of θ_{OB} that models decreased counts of event X relative to the baseline count, i.e., $\theta_{OB} \leq \theta_{NOB}$. Let $\theta'_{NOB} = 1 - \theta_{NOB}$ and $\theta'_{OB} = 1 - \theta_{OB}$, then monitoring for $\theta_{OB} \leq \theta_{NOB}$ is equivalent to monitoring for $\theta'_{OB} \geq \theta'_{NOB}$, for which the closed-form derivation shown in this paper can be directly applied. Then, the case of $\theta_{OB} \leq \theta_{NOB}$ can be derived by transformations of variables. We then can monitor for both anomalous increases and decreases of various variables.

BU algorithm could form the foundation of an analogous multivariate anomaly detection method. The most straightforward extension assumes independence of two (or more) types of evidence E_1 (e.g., respiratory chief complaints) and E_2 (e.g., over the counter sales of cough & cold medications), which yields that $P(E_1, E_2 | OB) = P(E_1 | OB) \times P(E_2 | OB)$, and $P(E_1, E_2 | NOB) = P(E_1 | NOB) \times P(E_2 | NOB)$. Each of the terms on the right side of these equations can be computed using the methods described in this paper. Multivariate extensions of BU that do not assume independence are more challenging and provide an open area of research.

BU could be extended to model the progression of a disease over time in order to create a temporal multivariate outbreak-detection algorithm along the lines of that described in Jiang [28].

In this paper we modeled $f(\theta_{OB})$ and $g(\theta_{NOB})$ as marginal prior probability distributions, as illustrated for example in ▶Figure 2. This approach allowed ▶Equations 3 and 4 to be derived separately. It would be interesting to derive and apply these equations when the distribution of θ_{OB} and θ_{NOB} is modeled jointly as well.

6. Related Work

This section gives a brief overview of some commonly used methods for anomaly detection, which includes frequentist methods and Bayesian methods.

Typical frequentist approaches for anomaly detection include methods from statistical quality control [29], regression [30], time series models [31], and wavelets [32, 33]. These methods are useful tools for anomaly detection and are commonly used in the public-health community for detection of disease outbreaks. However, it is difficult to incorporate any prior information that we may have, for example, our prior beliefs about the typical patterns of respiratory cases in outbreaks of respiratory diseases.

Bayesian approaches have been developed that can be applied to anomaly detection, such as dynamic linear models (DLMs) [34] and hidden Markov models (HMMs) [35]. DLMs are implemented by updating priors to obtain posteriors using a sequential approach for forecasting. To start a DLM modeling process, it is necessary to specify the initial priors before the arrival of the first observation of the time series. Nobre et al. [36] modeled the stochastic trend and the seasonal effect of an epidemiological time series using linear growth models described in [34]. They used a normal distribution with mean zero and a large variance to be the initial priors for the model parameters. LeStrat and Carrat [35] proposed to detect outbreak and non-outbreak phases of influenza by modeling the incidence rates of influenza-like illnesses with HMMs using a mixture of Gaussian distributions. Rath et al. [37] analyzed the same datasets and showed that better detection accuracy can be achieved by modeling the outbreak rates with a Gaussian distribution and the non-outbreak rates with an exponential distribution.

BU models the parameter prior probability of a non-anomaly using a Beta distribution. In contrast to previous Bayesian methods, including those discussed above, it derives the parameter prior distribution of an anomaly from that of the non-anomaly in a semi-informative manner. This prior probability distribution is semi-informative in the sense that the rate of a monitored event during an anomalous period is constrained to be greater than the rate during a non-anomalous period. Beyond that constraint, the anomaly prior probability is non-informative in the sense of following a uniform probability distribution.

7. Summary

This paper introduced a Bayesian anomaly-detection algorithm called BU. The BU algorithm takes as input a univariate time series of some event, as for example respiratory ED chief complaints. It then outputs a posterior probability of an anomalous event occurring. In performing anomaly detection, the algorithm uses a semi-informative prior that models an increased count over baseline. The semi-informative prior is derived based on the baseline prior, which is estimated from using historical data. We developed a computationally efficient method for using this prior to derive the posterior probability of an anomaly.

We described the algorithm in the context of disease outbreak detection. In a study using simulated anthrax outbreaks, the algorithm performed statistically significantly better than a control chart method when baseline periods were small enough to avoid seasonal effects. Importantly, the algorithm outputs a posterior probability, which can be used in decision analyses about how to act when confronted with a potential outbreak.

The algorithm is general and can be applied in many anomaly-detection application areas beyond disease outbreak detection. Also, as discussed above, there are a number of ways in which the algorithm can be extended, which appear promising.

Acknowledgments

This research was funded by a grant from the National Science Foundation (NSF IIS-0325581). We thank Bill Hogan for applying his BARD system to generate simulated cases of anthrax that were used in the evaluation, and we thank John Levander for his assistance in using those cases.

References

1. Wong W-K. Data mining for early disease outbreak detection. Doctoral Dissertation. Carnegie Mellon University, Pittsburgh. 2004.
2. Hauskrecht M, Valko M, Kveton B, Visweswaran S, Cooper GF, editors. Evidence-based anomaly detection in clinical domains. In: Proceedings of the Fall Symposium of the American Medical Informatics Association; 2007. pp 319–332.

3. Fawcett T, Provost F. Adaptive fraud detection. Data Mining and Knowledge Discovery 1997; 1 (3): 291–316.
4. Denning D. An intrusion-detection model. IEEE Transactions on Software Engineering. 1987; 13 (2): 222–232.
5. Shewhart WA. Economic control of quality of manufactured product. New York: D. Van Nostrand Company 1931. (A 1981 reprint is available from the American Society for Quality Control.)
6. Page ES. Continuous inspection schemes. Biometrika 1954; 41: 100–115.
7. Roberts SW. Control chart tests based on geometric moving averages. Technometrics 1959; 1: 239–250.
8. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman & Hall; 1995.
9. Hardy GF. Insurance record. 1889; 8. (Reprinted in Transactions of Actuaries, 1920.)
10. Whitworth WA. Exercise in choice and chance. 1897. (Reprinted by Hafner, New York, 1965.)
11. 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): 151–158.
12. Altun Y, Smola A. Unifying divergence minimization and statistical inference via convex duality. In: Proceedings of the 19th Annual Conference on Learning Theory. 2006. pp 139–153.
13. Barndorff-Nielsen O. Information and Exponential families in statistical theory. New York: Wiley; 1978.
14. Song L, Zhang X, Smola A, Gretton A, Scholkopf B. Tailoring density estimation via reproducing kernel moment matching. In: Proceedings of the 25th International Conference on Machine Learning 2008; 307: 992–999.
15. Gelman A, Carlin JB, Stern HS, Rubin DB. Appendix A: Standard probability distributions. Bayesian data analysis. London: Chapman & Hall; 1995. p 481.
16. Clayton DG, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987; 43: 671–681.
17. Mollie A. Bayesian and empirical Bayes approaches to disease mapping. In: Lawson ABea, editor. Disease Mapping and Risk Assessment for Public Health. Chichester: Wiley; 1999.
18. Neill DB, Moore AW, Cooper GF. A Bayesian spatial scan statistic. Advances in Neural Information Processing Systems 2006; 18: 1003–1010.
19. Burkom HS, Elbert Y, Feldman A, Lin J. Role of data aggregation in biosurveillance detection strategies with applications from ESSENCE. MMWR Morbility and Mortality Weekly Report 2004; Sep 24 (53 Suppl): 67–73.
20. 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: 5225–5252.
21. Weisstein EW. "Incomplete Beta Function." From MathWorld – A Wolfram Web Resource. 2003. Available from: http://mathworld.wolfram.com/IncompleteBetaFunction.html.
22. Casella G, Berger LR. Statistical Inference (second edition). Australia; Pacific Grove, CA: Thomson Learning; 2002.
23. Hogan WR, Cooper GF, Wagner MM. A Bayesian anthrax aerosol release detector. RODS Technical Report 2004.
24. Fawcett T, Provost F. Activity monitoring: Noticing interesting changes in behavior. In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining 1999. pp 53–62.
25. Kaufmann A, Meltzer M, Schmid G. The economic impact of a bioterrorist attack: Are prevention and postattack intervention programs justifiable? Emerging Infectious Diseases 1997; 3 (2): 83–94.
26. Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA, et al. The emerging science of very early detection of disease outbreaks. Journal of Public Health Management Practice 2001; 7 (6): 51–59.
27. Edwards W, Miles RF, von Winterfeldt D, editors. Advances in Decision Analysis. Cambridge University Press; 2007.
28. Jiang X. A Bayesian network model for spatio-temporal event surveillance. Doctoral Dissertation. University of Pittsburgh, Pittsburgh. 2008.
29. Hutwagner LC, Thompson W, Seeman GM. The bioterrorism preparedness and response early aberration reporting system (EARS). Journal of Urban Health 2003; 80 (2, Supplement 1): i89–i96.
30. Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public Health Reports 1963; 78: 494–506.
31. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. BMC Medical Informatics and Decision Making 2003; 3 (2).
32. Goldenberg A, Shmueli G, Caruana RA. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. In: Proceedings of National Academy of Sciences 2002; 99 (8): 5237–5240.
33. Zhang J, Tsui FC, Wagner MM, Hogan WR. Detection of outbreaks from time series data using wavelet transform. AMIA Annual Symposium Proceedings 2003. pp 748–752.
34. West M, Harrison J. Bayesian forecasting and dynamic models. New York: Springer-Verlag; 1989.
35. LeStrat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in Medicine 1999; 18: 3463–3478.
36. Nobre FF, Monteiro ABS, Telles PR, Williamson GD. Dynamic linear models and SARIMA: a comparison of their forecasting performance in epidemiology. Statistics in Medicine 2001; 20: 3051–3069.
37. Rath T, Carreras M, Sebastiani P. Automated detection of influenza epidemics with hidden Markov models. Proceedings of the Fifth International Symposium on Intelligent Data Analysis 2003; 2810: 521–532.
38. Weisstein EW. "Binomial Theorem." From MathWorld – A Wolfram Web Resource. 2006. Available from: http://mathworld.wolfram.com/BinomialTheorem.html.

Appendix: Derivation of $P(E|OB)$

By substituting Equation 1 into Equation 4, we obtain the following:

$$P(E \mid OB) = \int_0^1 {\theta_{OB}}^{N_1} (1-\theta_{OB})^{N_2} \int_0^{\theta_{OB}} \frac{1}{1-\theta_{NOB}} g(\theta_{NOB})\, d\theta_{NOB} d\theta_{OB} \,. \tag{7}$$

By changing the order of integration and substituting $g(\theta_{NOB})$ using its density function $Beta(\theta_{NOB}; \alpha_0, \beta_0)$, we obtain Equation 8.

$$\begin{aligned} P(E \mid OB) &= \frac{1}{B(\alpha_0,\beta_0)} \int_0^1 \frac{1}{1-\theta_{NOB}} {\theta_{NOB}}^{\alpha_0-1} (1-\theta_{NOB})^{\beta_0-1} \int_{\theta_{NOB}}^1 {\theta_{OB}}^{N_1} (1-\theta_{OB})^{N_2} \, d\theta_{OB} d\theta_{NOB} \\ &= \frac{1}{B(\alpha_0,\beta_0)} \int_0^1 {\theta_{NOB}}^{\alpha_0-1} (1-\theta_{NOB})^{\beta_0-2} \int_{\theta_{NOB}}^1 {\theta_{OB}}^{N_1} (1-\theta_{OB})^{N_2} \, d\theta_{OB} d\theta_{NOB}. \end{aligned} \tag{8}$$

Note that the inner integral in Equation 8 integrates all possible values of θ_{OB} as $\theta_{NOB} \le \theta_{OB} \le 1$. In order to express this inner integral in terms of a Beta function, we introduce a variable γ, where $0 \le \gamma \le 1$. Let $\theta_{OB} = (1-\theta_{NOB})\gamma + \theta_{NOB}$, then the inner integral in Equation 8 can be calculated as follows by using variable substitution.

$$\begin{aligned} \int_{\theta_{NOB}}^1 {\theta_{OB}}^{N_1} (1-\theta_{OB})^{N_2} \, d\theta_{OB} &= \int_0^1 [(1-\theta_{NOB})\gamma + \theta_{NOB}]^{N_1} \{1-[(1-\theta_{NOB})\gamma + \theta_{NOB}]\}^{N_2} \, d[(1-\theta_{NOB})\gamma + \theta_{NOB}] \\ &= \int_0^1 [(1-\theta_{NOB})\gamma + \theta_{NOB}]^{N_1} [1-\theta_{NOB} - (1-\theta_{NOB})\gamma]^{N_2} (1-\theta_{NOB}) d\gamma \\ &= \int_0^1 [(1-\theta_{NOB})\gamma + \theta_{NOB}]^{N_1} (1-\theta_{NOB})^{N_2+1} (1-\gamma)^{N_2} \, d\gamma. \end{aligned} \tag{9}$$

By using the Binomial theorem [38], the term $[(1-\theta_{NOB})\gamma + \theta_{NOB}]^{N_1}$ in Equation 9 can be represented as $\sum_{n=0}^{N_1} \binom{N_1}{n} (1-\theta_{NOB})^n \gamma^n \theta_{NOB}^{N_1-n}$ since N_1 is a positive integer. Therefore, Equation 9 can be further written as Equation 10.

$$\int_{\theta_{NOB}}^1 {\theta_{OB}}^{N_1} (1-\theta_{OB})^{N_2} \, d\theta_{OB} = \int_0^1 (1-\theta_{NOB})^{N_2+1} (1-\gamma)^{N_2} \sum_{n=0}^{N_1} \binom{N_1}{n} (1-\theta_{NOB})^n \gamma^n \theta_{NOB}^{N_1-n} d\gamma. \tag{10}$$

By swapping the order of integration and summation in Equation 10, we obtain

$$\begin{aligned} \int_{\theta_{NOB}}^1 {\theta_{OB}}^{N_1} (1-\theta_{OB})^{N_2} \, d\theta_{OB} &= \sum_{n=0}^{N_1} \binom{N_1}{n} \theta_{NOB}^{N_1-n} (1-\theta_{NOB})^{N_2+n+1} \int_0^1 \gamma^n (1-\gamma)^{N_2} \, d\gamma \\ &= \sum_{n=0}^{N_1} \binom{N_1}{n} \theta_{NOB}^{N_1-n} (1-\theta_{NOB})^{N_2+n+1} B(n+1, N_2+1). \end{aligned} \tag{11}$$

Finally by substituting Equation 11 into Equation 8 and swapping the order of integration and summation, we write $P(E|OB)$ as follows:

$$\begin{aligned} P(E \mid OB) &= \frac{1}{B(\alpha_0,\beta_0)} \int_0^1 {\theta_{NOB}}^{\alpha_0-1} (1-\theta_{NOB})^{\beta_0-2} \left[\sum_{n=0}^{N_1} \binom{N_1}{n} \theta_{NOB}^{N_1-n} (1-\theta_{NOB})^{N_2+n+1} B(n+1, N_2+1) \right] d\theta_{NOB} \\ &= \frac{1}{B(\alpha_0,\beta_0)} \sum_{n=0}^{N_1} \binom{N_1}{n} B(n+1, N_2+1) \int_0^1 {\theta_{NOB}}^{N_1+\alpha_0-n-1} (1-\theta_{NOB})^{N_2+\beta_0+n-1} \, d\theta_{NOB} \\ &= \frac{1}{B(\alpha_0,\beta_0)} \sum_{n=0}^{N_1} \binom{N_1}{n} B(n+1, N_2+1) B(N_1+\alpha_0-n, N_2+\beta_0+n). \end{aligned}$$