# BAYESIAN MODELING OF ANOMALIES DUE TO KNOWN AND UNKNOWN CAUSES

by

**Yanna Shen**

Bachelor of Science, Northeastern University, China, 2000

Master of Science, Northeastern University, China, 2003

Submitted to the Graduate Faculty of

Intelligent Systems Program in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

INTELLIGENT SYSTEMS PROGRAM

This dissertation was presented

by

Yanna Shen

It was defended on

April 23, 2009

and approved by

Gregory F. Cooper, Associate Professor, Biomedical Informatics and Intelligent Systems

Marek J. Druzdzel, Associate Professor, Information Sciences and Intelligent Systems

Rich Tsui, Research Assistant Professor, Biomedical Informatics and Intelligent Systems

Garrick Wallstrom, Assistant Professor, Biomedical Informatics

Dissertation Advisor: Gregory F. Cooper, Associate Professor, Biomedical Informatics and Intelligent Systems

# BAYESIAN MODELING OF ANOMALIES DUE TO KNOWN AND UNKNOWN CAUSES

Yanna Shen, PhD

University of Pittsburgh, 2009

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 (or systems) need to recognize under uncertainty when they are likely to be experiencing influences outside their realm of knowledge. This dissertation investigates Bayesian modeling of unknown causes of events in the context of disease-outbreak detection.

The dissertation introduces a Bayesian approach that models and detects (1) known diseases (e.g., influenza and anthrax) by using informative prior probabilities, (2) unknown diseases (e.g., a new, highly contagious respiratory virus that has never been seen before) by using relatively non-informative prior probabilities and (3) partially-known diseases (e.g., a disease that has characteristics of an influenza-like illness) by using semi-informative prior probabilities. I 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 dissertation is that it introduces a Bayesian approach for jointly modeling both known and unknown causes of events. Such modeling has broad applicability in artificial intelligence in general and biomedical informatics applications in particular, where the space of known causes of outcomes of interest is seldom complete.

## TABLE OF CONTENTS

PREFACE ........ XVI

1.0 INTRODUCTION ........ 1

1.1 OVERVIEW OF THE PROPOSED METHODOLOGY ........ 4

1.2 THE DISSERTATION HYPOTHESIS ........ 7

1.3 GUIDE FOR THE READER ........ 8

2.0 BACKGROUND ........ 11

2.1 BAYESIAN FRAMEWORK ........ 12

2.2 OVERVIEW OF OUTBREAK-DETECTION ALGORITHMS ........ 13

2.2.1 Non-disease-specific detection methods ........ 14

2.2.1.1 Frequentist approaches. ........ 14

2.2.1.2 Bayesian approaches ........ 18

2.2.2 Disease-specific detection methods ........ 21

2.2.2.1 Frequentist approaches. ........ 21

2.2.2.2 Bayesian approaches ........ 21

2.2.3 Methods combining disease-specific and non-disease-specific detectors... 23

2.3 EVALUATION OF OUTBREAK-DETECTION ALGORITHMS ........ 24

2.4 CHOICE OF A PRIOR DISTRIBUTION ........ 28

2.4.1 Conjugate priors ........ 29

2.4.2 Non-conjugate priors ........ 29

2.4.3 Non-informative priors ........ 30

2.4.4 Why use the Beta distribution as a prior belief? ........ 33

2.4.5 Empirical Bayes methods ........ 36

2.4.6 Mixture priors ........ 37

2.5 UNKNOWN-OBJECT REPRESENTATION ........ 38

2.5.1 Probabilistic graphical models ........ 39

2.5.2 First-order probabilistic languages (FOPLs) ........ 39

2.5.3 Bayesian logic (BLOG) ........ 40

2.5.4 Nonparametric Bayesian logic (NP-BLOG) ........ 43

2.6 ENTITY-BASED MODELS FOR RELATIONAL DATA ........ 44

2.6.1 Relational dataset ........ 45

2.6.2 Entity-relationship models ........ 46

2.6.3 Probabilistic entity-relationship model ........ 49

2.6.4 Plate model ........ 50

3.0 THE EXPERIMENTAL DOMAIN ........ 53

3.1 NOTATION ........ 53

3.2 MD-PANDA ........ 55

3.2.1 The nodes ........ 57

3.2.2 The conditional probability tables ........ 59

3.2.3 Inference ........ 61

4.0 THE BAYESIAN HYBRID DETECTION ALGORITHM ........ 62

4.1 DISEASE MODELING ........ 63

4.1.1 The non-outbreak disease model ..... 64

4.1.2 The disease-specific model (DSM) ..... 66

4.1.3 The unknown-disease model (UDM) ..... 67

4.1.4 The partially-known disease model (PDM) ..... 67

4.2 THE UNIVARIATE BAYESIAN HYBRID DETECTION ALGORITHM 70

4.2.1 The univariate entity-based disease model ..... 70

4.2.2 The conditional probability tables ..... 71

4.2.3 Inference ..... 73

4.3 THE MULTIVARIATE BAYESIAN HYBRID DETECTION ALGORITHM ..... 78

4.3.1 The multivariate entity-based disease model ..... 78

4.3.2 The conditional probability tables ..... 81

4.3.3 Inference ..... 82

5.0 EXPERIMENTAL EVALUATION ..... 87

5.1 CHOICES OF DISEASES AND SYMPTOMS ..... 88

5.2 CREATING DATASETS ..... 89

5.3 EVALUATION METHODS ..... 95

5.3.1 Experimental design ..... 95

5.3.2 Experimental procedures ..... 99

5.4 TECHNICAL STATEMENT OF THE HYPOTHESIS ..... 103

5.5 OVERVIEW OF REPRESENTATIVE RESULTS ..... 104

5.5.1 AMOC curves of the UBH algorithm ..... 105

5.5.2 AMOC curves of the MBH algorithm ..... 112

5.6 STATISTICAL ANALYSIS ........................................................................ 115

5.7 DECISION ANALYSIS ........................................................................... 123

5.8 SENSITIVITY ANALYSIS ........................................................................ 128

5.8.1 Results of the UBH algorithm................................................................ 129

5.8.2 Results of the MBH algorithm................................................................ 136

6.0 CONCLUSION AND FUTURE WORK ........................................................ 142

6.1 CONTRIBUTIONS ................................................................................... 143

6.1.1 Disease modeling.................................................................................. 143

6.1.2 A unifying framework in a hybrid disease detection system.................... 144

6.2 FUTURE WORK........................................................................................ 144

APPENDIX A ................................................................................................. 147

APPENDIX B ................................................................................................. 148

BIBLIOGRAPHY ............................................................................................ 155

# LIST OF TABLES

Table 2.1 Evaluation methods used for outbreak detection algorithms ....................................... 25

Table 2.2 An example dataset that MD-PANDA monitors ......................................................... 46

Table 4.1 The conditional probability table for $P(E_i \mid PD_i)$ ........................................................ 73

Table 4.2 The conditional probability table for $P(E_i^j \mid PD_i)$ ........................................................ 82

Table 5.1 A 2 × 3 table that summarizes the two experiments that each involves the use of a disease-specific model (DSM), an unknown disease model (UDM), and a partially-known disease model (PDM)............................................................................................................ 97

Table 5.2 Detection times that were obtained for models DSM1, UDM1, and PDM1 (in Exp. 1) over 27 experimental configurations × 96 outbreak scenarios. ................................................. 117

Table 5.3 Detection times that were obtained for models DSM2, UDM2, and PDM2 (in Exp. 2) over 54 experimental configurations × 96 outbreak scenarios. ................................................. 117

Table 5.4 *p*-values of three pairwise comparisons at zero and one false alert per month, respectively, when using the UBH algorithm. ............................................................................ 122

Table 5.5 *p*-values of three pairwise comparisons at zero and one false alert per month, respectively, from using the MBH algorithm. ............................................................................ 122

Table 5.6 Mean detection time (in days) at a false alert rate of zero and one per month of all six models over all the experiments when using the UBH algorithm ............................................ 124

Table 5.7 Mean detection time (in days) at a false alert rate of zero and one per month of all six models over all the experiments when using the MBH algorithm............................................ 125

Table 5.8 A summary of the decision analysis results.............................................................. 127

Table A.1 Parameters ($\alpha$, $\beta$) estimated for the non-outbreak disease model and the disease-specific models given the non-outbreak disease, three selected outbreak diseases, and three selected disease symptoms.......................................................................................... 147

Table B.1 A complete list of real ED cases for 2004 and 2005 from a large hospital in Allegheny County, Pennsylvania. ........................................................................................... 148

# LIST OF FIGURES

Figure 1.1 An example that shows modeling of the distribution of a parameter and a point value of the parameter. The curve shows the distribution of a parameter that represents the frequency of cough in the individuals with early stage anthrax; the dot on the X-axis shows the mean of the distribution, which represents the point value of this parameter. .................................................. 5

Figure 1.2 Diagram showing the continuous disease-modeling space. ......................................... 6

Figure 2.1 A high-level Bayesian network representation showing the MD-PANDA model, where the dashed rectangular shapes represent subnetworks: subnetwork *G* contains nodes that represent global features common to all people in the population being monitored, and subnetwork *E* represents an entity-based model that contains every entity (Person) in the population; the bolded arrow between subnetworks shows the direction in which the Bayesian-network arcs are oriented between subnetworks, and the arrows between nodes in subnetwork *E* show probabilistic dependencies between nodes. ......................................................................... 47

Figure 2.2 An entity-relationship (ER) model depicting the structure of the relational dataset for Table 2.1. The entity classes (Person and Disease) are shown as rectangular nodes; the relationship class (Has) is shown as a diamond-shaped node; and the attribute classes (Disease.Disease_state, Person.Chief_complaint) are shown as oval nodes. Dashed edges are used for connecting attribute classes to their corresponding entity classes and the relationship class to its associated entity classes. ........................................................................................ 48

Figure 2.3 A directed acyclic probabilistic entity-relationship (DAPER) model showing that a person's chief complaint depends on the person's disease state................................................ 50

Figure 2.4 Plate notation for the high-level Bayesian network model used by MD-PANDA...... 51

Figure 3.1 The Bayesian network structure for the MD-PANDA model. See the text next for a description of the nodes and the conditional probability tables................................................. 55

Figure 3.2 Plate notation of the MD-PANDA model. The subgraph in the plate (bolded box) repeats *N* times, where *N* is the number of individuals in the population being monitored, and any links that cross a plate boundary are replicated once for each subgraph repetition..................... 56

Figure 4.1 A Bayesian network showing the UBH entity-based disease model. ......................... 71

Figure 4.2 A naïve Bayes model representing *J* evidential features for a specific person *i* in the population. ..................................................................................................................... 79

Figure 4.3 Bayesian network showing the MBH disease model where every person's disease state and evidence states are modeled using a naïve Bayes model.............................................. 80

Figure 4.4 Plate notation of the naïve Bayes model shown in Figure 4.2, where *J* in the corner represents the total *J* evidential features modeled for person *i*.................................................. 80

Figure 4.5 Plate notation of the Bayesian network model shown in Figure 4.3, where *J* on the inner plate represents the total *J* evidential features modeled for person *i*, and *N* on the outer plate represents the total number of population being monitored in the region. .................................. 81

Figure 5.1 The probability density functions of each symptom probability given each disease. . 89

Figure 5.2 An example showing outbreak cases generated using FLOO(1,10), in which 1, 2, 3, 4 outbreak cases were generated on day 1 to day 4, respectively, and 5 outbreak cases were generated per day for day 5 to day 10. ........................................................................ 92

Figure 5.3 (a) shows the AMOC curves of Exp. 1, and (b)-(c) show the AMOC curves of Exp. 2 with $d_v$ = *cryptosporidiosis* in (b) and $d_v$ = *early stage anthrax* in (c)...................................... 106

Figure 5.4 AMOC curves of two experimental configurations for Exp. 2. ............................... 108

Figure 5.5 AMOC curves of two experimental configurations for Exp. 1. ............................... 109

Figure 5.6 AMOC curves of two experimental configurations for Exp. 1. ............................... 110

Figure 5.7 Probability density functions of the probability of cough given outbreak disease *early stage anthrax*, the unknown disease $d_*$, and the partially-known disease $d_{*p}$. The mixture distribution is composed of a mixture of a uniform distribution, the distribution for cryptosporidiosis, and the distribution for inhalation tularemia. ............................................... 112

Figure 5.8 (a) and (d) shows the AMOC curves for Exp. 1, and (b), (c), (e), (f) show the AMOC curves for Exp. 2 with $d_v$ = *early stage anthrax* in (b) and (e), and $d_v$ = *inhalation tularemia* in (c) and (f). The plots in (d)-(f) are different from those in (a)-(c) only in that FLOO($\Omega$, $T$) = FLOO(4,14), and thus, the simulated outbreaks are stronger. .................................................. 114

Figure 5.9 Scatter plot of the standardized within-group residuals versus the within-group fitted values for the fitted model. ......................................................................................... 120

Figure 5.10 Normal plot of within-group standardized residuals for the fitted model. ............. 121

Figure 5.11 A decision tree showing the decision analysis on selecting DSM vs. UDM .......... 126

Figure 5.12 Mean detection time (days) at one false alert per month for DSM, UDM, and PDM in Exp. 1 using the UBH algorithm ......................................................................................... 133

Figure 5.13 Mean detection time (days) at one false alert per month for DSM, UDM, and PDM in Exp. 2 using the UBH algorithm ........................................................................... 133

Figure 5.14 Probability value $q_*$ and $q_{*p}$ relative to the prior probability of the appearance of the unknown (partially-known) disease at one false alert per month using the UBH algorithm...... 134

Figure 5.15 Mean detection time (days) at zero false alerts per month for DSM, UDM, and PDM in Exp. 1 using the UBH algorithm ........................................................................... 134

Figure 5.16 Mean detection time (days) at zero false alerts per month for DSM, UDM, and PDM in Exp. 2 using the UBH algorithm ........................................................................... 135

Figure 5.17 Probability value $q_*$ and $q_{*p}$ relative to the prior probability of the appearance of the unknown (partially-known) disease at zero false alerts per month using the UBH algorithm... 135

Figure 5.18 Mean detection time (days) at one false alert per month for DSM, UDM, and PDM in Exp. 1 using the MBH algorithm........................................................................... 139

Figure 5.19 Mean detection time (days) at one false alert per month for DSM, UDM, and PDM in Exp. 2 using the MBH algorithm........................................................................... 139

Figure 5.20 Probability value $q_*$ and $q_{*p}$ relative to the prior probability of the appearance of the unknown (partially-known) disease at one false alert per month using the MBH algorithm..... 140

Figure 5.21 Mean detection time (days) at zero false alerts per month for DSM, UDM, and PDM in Exp. 1 using the MBH algorithm........................................................................... 140

Figure 5.22 Mean detection time (days) at zero false alerts per month for DSM, UDM, and PDM in Exp. 2 using the MBH algorithm........................................................................... 141

Figure 5.23 Probability value $q_*$ and $q_{*p}$ relative to the prior probability of the appearance of the unknown (partially-known) disease at zero false alerts per month using the MBH algorithm .. 141

# PREFACE

This dissertation could not be finished without the help and support from my advisor, Gregory Cooper. I would like to give my greatest thanks to him for his patient supervision, his enthusiastic support, and many valuable discussions with me in the past five years. I have learned a lot not only from his profound professional knowledge but also from his personality and working attitude.

I would also like to thank my dissertation committee members: Marek Druzdzel, Rich Tsui, and Garrick Wallstrom. I appreciate their openness for help and their insightful and helpful comments on my dissertation work. I am very grateful for the financial support from the Intelligent Systems Program and Department of Biomedical Informatics at the University of Pittsburgh, and support by a grant from the National Science Foundation (NSF IIS-0325581).

Members of the Bayesian Biosurveillance group at the Department of Biomedical Informatics have also been very helpful. John Levander, in particular, has offered lots of assistance in generating real emergency department cases according to my need. I also thank John Dowling for assessing the frequency of 54 symptom states of CDC-A outbreak diseases in the population. Shyam Visweswaran was always willing to offer valuable advice when I ran into difficult problems. I also have received useful suggestions from my fellow students Xia Jiang and Peter Sutovsky.

Finally, I would like to thank my parents and my brother for their love and patience through my time at the University of Pittsburgh. Though they spent very limited time together with me during the past six years, their support has always been a source of energy throughout my doctoral study. My thanks should also be given to my husband, Weining Kang, who has been a source of support and inspiration during the past two years. He has always been ready with a comforting word when things did not go well and with a congratulatory word when things worked out in the end. Without his support, I would not be able to achieve all of this.

# 1.0 INTRODUCTION

*Anomaly detection* refers to the problem of finding anomalous patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, aberrations or contaminants in different application domains. Of these, anomalies and outliers are two terms used most commonly in the context of anomaly detection. Detecting anomalies or outliers in data has been studied in the statistics community as early as the 19th century (Edgeworth 1887). Chandola, et al. (Chandola 2009) provide a comprehensive overview of the research on anomaly detection.

Anomaly detection has some important applications in domains such as disease outbreak detection (Wong 2004), fraud detection (Fawcett 1997), and electronic intrusion detection (Warrender 1999). 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. The causes of anomalies can be divided into two types – those that we can anticipate and those that we cannot. As a result, algorithms within these monitoring systems can be classified into two categories – those that detect anomalies we can anticipate and those that detect anomalies we cannot. In this study I refer to these as *specific detection algorithms* and *non-specific detection algorithms*, respectively.

A specific detection algorithm looks for a pre-defined anomalous pattern in the data. For example, in the context of disease outbreak detection, a specific detection approach would examine health-care data for the onset of a specific disease such as inhalational anthrax. A large-scale airborne release of inhalational anthrax has anticipated spatio-temporal characteristics, such as a specific incubation time and a plume-like spatial distribution (Hogan 2004). As a result, when monitoring for such an outbreak, a detection algorithm should be vigilant in watching for these characteristics. BARD (Hogan 2007) is an example of a specific detection algorithm. It models the effects of an outdoor airborne anthrax release using the Gaussian plume model of atmospheric dispersion and a model of inhalational anthrax.

In contrast, a non-specific detection approach tries to detect any anomalous events, relative to some baseline of “normal” behavior. It searches for any statistically significant anomalous patterns in the data, not just those that fit known causes of anomalies. The WSARE algorithm is an example of a non-specific detection algorithm (Wong 2004). It works by searching for any significant changes in the ratios of different subgroups of the data between a recent period and a period in the more distant past. A non-specific detection algorithm can detect a wide range of anomaly types, but usually at the expense of being less effective at detecting any particular type. Commonly used non-specific detection algorithms are usually implemented using frequentist methods that are based on deriving a *p*-value from recent data. Examples of these approaches include Shewhart’s control charts, CUSUM, and EWMA (Montgomery 1991).

Frequentist approaches are useful tools for anomaly detection and are commonly used in the public health community for detection of disease outbreaks. However, compared with Bayesian approaches, it is difficult to incorporate any prior information that we may have, as for example our prior beliefs about the size, location, or temporal progression of a potential outbreak. Bayesian outbreak detection algorithms have another advantage in that they can be readily used in a decision-analytic framework to compute the expected utility of a decision. For example, the algorithm may calculate the posterior probability that there is a SARS outbreak as 0.80, which strongly suggests that some action is in order. Alternatively, the algorithm may compute the probability as $10^{-12}$, and then, no action would be required.

The objective of this dissertation is to develop an anomaly-detection algorithm that is able to capture anomalies due to both known and unknown causes. As described in this section, both specific and non-specific detection algorithms have their advantages and disadvantages in detecting anomalies due to known or unknown causes. Therefore, I propose a hybrid detection algorithm that combines specific and non-specific detectors in order to capture anomalous events due to both known and unknown causes. I propose a Bayesian method to develop this algorithm because we can (1) incorporate any prior knowledge we may have about the anomalies due to known causes, (2) develop our own non-informative and semi-informative priors to model anomalies due to unknown causes, and (3) readily apply the algorithm in a decision-analytic framework to compute the expected utility of a decision.

Bayesian anomaly detection essentially performs Bayesian inference on anomalous events by combining prior beliefs about model parameters with evidence from data using Bayes' theorem. Two challenges of Bayesian inference are that it can be difficult to specify the prior distribution and that the required computations are difficult.

In this dissertation, I describe a Bayesian anomaly-detection algorithm in the context of disease outbreak detection and describe methods to develop my own algorithm to address these two challenges. Because the fundamental ideas are general, they could be used to develop richer Bayesian detection models and detection algorithms than those that currently exist.

## 1.1 OVERVIEW OF THE PROPOSED METHODOLOGY

The question of how to specify a prior distribution turns out to be a disease-modeling problem in the context of disease-outbreak detection. Modeling non-outbreak diseases (i.e., all the diseases that people can have in the absence of any infectious disease outbreak in the population) is relatively straightforward, at least conceptually. Given the large amount of electronically available data, such as emergency department visits, over-the-counter medication sales, and ambulatory care visit records, we can apply traditional machine-learning techniques to learn the parameters of the non-outbreak disease model. However, due to the lack of actual outbreaks of diseases under surveillance, for many diseases it is difficult to learn the outbreak models using machine-learning techniques. Even though some outbreak disease-specific models could be learned, such as influenza outbreaks for which we have ample training data, there are many outbreak diseases for which we have little or no training data. Also, an outbreak could be due to a new infectious disease for which we obviously would not have a specific model. I define such a disease as an *unknown* outbreak disease since it is the one that we do not yet know about. There are some diseases that are known to us but not explicitly modeled, as for example Brucellosis. I define unknown outbreak diseases and unmodeled (but known) outbreak diseases as *non-specific* outbreak diseases[1]. Therefore, non-specific outbreak disease modeling is an important and significant challenge.

The disease-modeling problem in this dissertation describes how to model the probability of a symptom in a disease, as for example the probability that a patient will have a cough given that he or she has respiratory anthrax. Common disease-modeling problems model this parameter

[1] For convenience, I will assume that a non-specific outbreak disease is an unknown disease, unless stated otherwise.

(probability) as a known probability value (Cooper 2006), as for example based on expert assessment. In contrast, this dissertation models the *distributions* of parameters that represent frequencies of the population, as for example the frequency of cough in the individuals with anthrax, whereas the true parameters (frequencies) are unknown. Figure 1.1 shows an example of modeling the distribution of a parameter and a point value of the parameter, where the parameter represents the frequency of cough in the individuals with early stage anthrax. Modeling the distributions of parameters allows us to express our ignorance about how diseases will present themselves and to develop disease models that flexibly fit a variety of symptom patterns in the population.

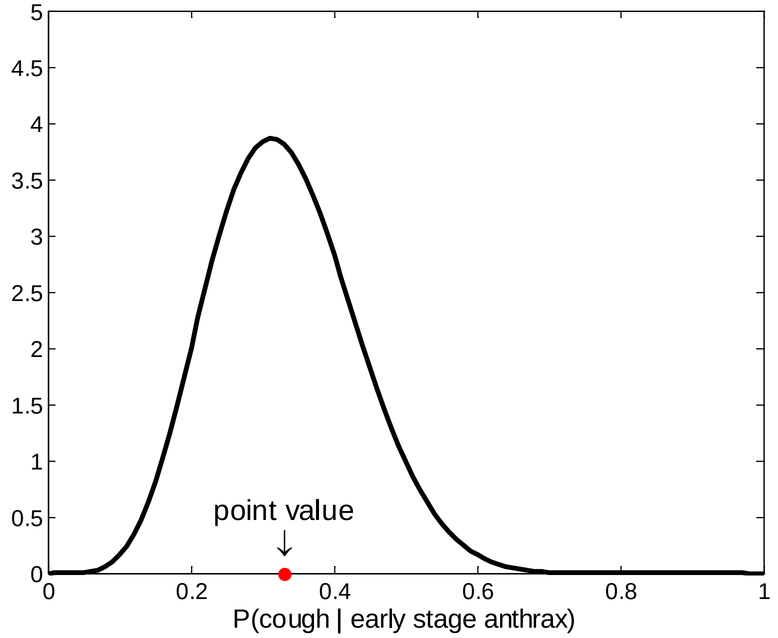


**Figure 1.1** An example that shows modeling of the distribution of a parameter and a point value of the parameter. The curve shows the distribution of a parameter that represents the frequency of cough in the individuals with early stage anthrax; the dot on the X-axis shows the mean of the distribution, which represents the point value of this parameter.

In this dissertation I explore the problem of modeling unknown outbreak diseases. In particular, I will consider disease modeling as a continuous modeling space, from disease-specific to partially-known-disease to unknown-disease modeling (see Figure 1.2). I will explore key areas of this modeling space with an outbreak-detection system (Cooper 2006) that uses Bayesian methods for detecting CDC Category A outbreak diseases (CDC).

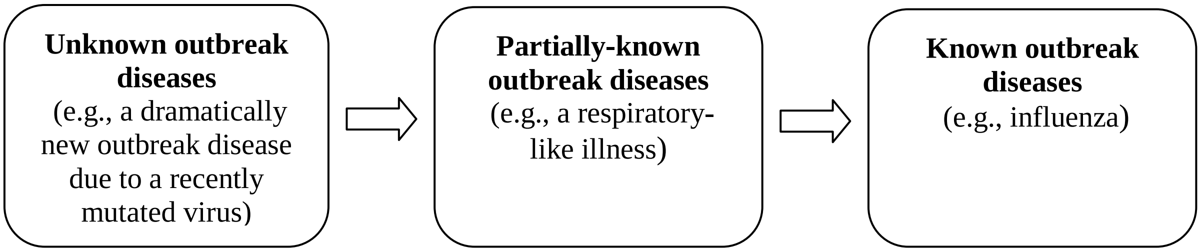

**Unknown outbreak diseases**
(e.g., a dramatically new outbreak disease due to a recently mutated virus)
**Partially-known outbreak diseases**
(e.g., a respiratory-like illness)
**Known outbreak diseases**
(e.g., influenza)

**Figure 1.2** Diagram showing the continuous disease-modeling space.

A *disease-specific model* (DSM) models a disease we are aware of and want to detect. It incorporates our prior knowledge of this disease, such as prior beliefs about the incubation time of the disease. An example of such a disease is influenza, since it is a disease we know about and care to detect early. Other examples are the CDC Category A diseases, which include anthrax, botulism, plague, smallpox, tularemia, and viral hemorrhagic fevers.

In contrast, an *unknown-disease model* (UDM) models a disease about which we have almost no knowledge. Hence, I propose a non-informative prior for modeling unknown outbreak diseases. I anticipate that the use of non-informative priors will enable the UDMs to capture disease outbreaks due to diseases that are unanticipated or unmodeled.

A *partially-known-disease model* (PDM), on the other hand, models an anticipated but partially-known disease, as for example a disease that has characteristics of a respiratory-like illness. In this dissertation, I propose a semi-informative prior for modeling partially-known diseases. In particular, a mixture of priors is proposed in order to model outbreak diseases that may have disease characteristics similar to several known outbreak diseases. For example, a mixture of priors that has components of several known respiratory diseases, such as anthrax, plague, and inhalation tularemia, could be used for modeling a respiratory-like illness.

In addition, there is a fourth category of disease model that represents an outbreak disease we know about but do not care to detect. If we want specifically to have a system that detects inhalational anthrax, then we do not want the detection system to send out alerts on other diseases that manifest similar patient symptoms, such as influenza. In this case, we would model influenza in order to *avoid* having the detection system alert on it.

This dissertation addresses the disease modeling problem of the first three categories of models just described and introduces a Bayesian detection algorithm that is able to incorporate all three types of models. I call the proposed Bayesian hybrid detection algorithm the BH algorithm.

## 1.2 THE DISSERTATION HYPOTHESIS

In this dissertation research, I develop a hybrid detection system that combines models of known and unknown outbreak diseases together in order to capture outbreaks due to known diseases and due to unanticipated or unknown diseases. In most real-world problems, we do not know the underlying outbreak disease that is causing the ongoing disease outbreak. Since modeling both known and unknown outbreak diseases allows us to represent a wide range of disease presentations than modeling known outbreak diseases only, the hypothesis of this dissertation is that

> *Modeling both known and unknown outbreak diseases in a hybrid system can lead to better expected disease outbreak detection performance than modeling known outbreak diseases only.*

The dissertation makes this general statement more precise, so that it can be tested, as I did.

Two scenarios occur in the real world when there is an outbreak occurring due to some disease *d*. One scenario is that disease *d* is known to us and has been modeled as a DSM disease. Then one might conjecture that incorporating UDM or PDM in the detection system (which model the possibility of unknown or partially-known disease, respectively) would detract the system from detecting disease *d*. Another scenario is that disease *d* is unexpected and not explicitly modeled in DSM. Then one could conjecture that incorporating UDM or PDM models in the detection system might help in the detection of disease *d*, relative to not modeling *d* in DSM. This dissertation will describe how I constructed these two scenarios in order to test the above hypothesis.

Note that there are numerous ways of modeling unknown outbreak diseases or partially-known outbreak diseases in UDMs or PDMs. This dissertation focuses on a basic way of modeling unknown diseases in order to investigate whether incorporating UDMs or PDMs in a detection system can improve the system's detection performance relative to a system that only incorporates DSMs.

## 1.3 GUIDE FOR THE READER

The dissertation is organized as follows. Chapter 2 contains the background of the dissertation research. Since the proposed algorithm is a Bayesian hybrid detection algorithm that combines specific and non-specific detectors, I first provide a Bayesian framework that combines the results of these detectors. Then I provide an overview of some commonly used specific and non-specific detection algorithms and their evaluation methods. Additionally, I review some representative research on prior distributions and some previous research on modeling unknown objects. Finally, entity-relationship methods are briefly reviewed, because this dissertation builds on a previously developed outbreak disease detection system that uses an entity relationship method.

Chapter 3 first introduces notation used for the remainder of the dissertation. It then introduces the experimental domain of the proposed Bayesian hybrid detection algorithm (the BH algorithm), the MD-PANDA model. In addition, this chapter provides a brief overview of how the MD-PANDA detection system monitors and detects CDC Category A diseases.

Chapter 4 describes the main issues of the proposed BH algorithm, which includes a univariate version and a multivariate version. In particular, the BH algorithm extends MD-PANDA to (1) model unknown and partially-known outbreak diseases, and (2) model the distributions over the probabilities of a person's symptoms, given different diseases that a person could have. Additionally, Chapter 4 includes methods for performing Bayesian inference.

Chapter 5 presents the experimental methods for evaluating the univariate and multivariate versions of the BH algorithm. In particular, I construct two experimental scenarios when there is an outbreak occurring due to some disease *d*, as described in Section 1.2. In addition, experimental results are presented and discussed for the univariate and multivariate versions of the algorithm. This chapter also provides statistical analyses and sensitivity analyses of the experimental results and performs decision analysis regarding whether and in what situations it would be beneficial to model an unknown disease in the detection system. Finally, Chapter 6 contains conclusions and suggestions for future research.

# 2.0 BACKGROUND

This chapter provides background knowledge that relates to the proposed dissertation research. The proposed Bayesian hybrid outbreak-detection algorithm (the BH algorithm) is a Bayesian approach that combines disease-specific and non-disease-specific detectors. I thus begin in Section 2.1 by describing a basic Bayesian framework of outbreak detection that combines the results of a suite of specific and non-specific detectors. Section 2.2 provides a brief overview of some commonly used outbreak detection algorithms, including both disease-specific detection algorithms and non-disease-specific detection algorithms, and Section 2.3 provides a brief overview of the evaluation methods of the outbreak detection algorithms. Section 2.4 provides background on choice of prior distributions in Bayesian inference, where I particularly review previous work on non-informative prior distributions. Since unknown-disease modeling will be explored in this dissertation, Section 2.5 includes some prior research on languages for representing unknown objects. Finally, Section 2.6 provides a brief overview about entity-relationship models as background for introducing the disease domain in which the BH algorithm is evaluated.

## 2.1 BAYESIAN FRAMEWORK

Let *H* be a hypothesis and *E* denote some available evidence. We are often interested in knowing the posterior probability of *H* in light of *E*, that is *P*(*H* | *E*). Assume we can estimate the likelihood *P*(*E* | *H*). Frequently such likelihoods are derived from a model that represents the probability that *H* generates *E*. A Bayesian approach requires the specification of a prior probability of *H*, namely *P*(*H*), which is our belief in *H* before seeing evidence *E*. Equation 2.1 shows the well-known use of the Bayes' rule to derive *P*(*H* | *E*).

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{\sum_{H' \in S} P(E \mid H')P(H')}, \tag{2.1}$$

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

The hypotheses in *S* can be at different levels of abstraction. Consider an anomaly detection application in which we are monitoring population evidence *E* for new outbreaks of disease. Such evidence might include the symptoms of patients who have recently visited emergency departments in a given region. Suppose *S* includes some set of disease-specific disease outbreaks (e.g., outbreaks due to inhalational anthrax, SARS, and West Nile virus). Another hypothesis in *S* represents the absence of any disease outbreak in the population. Traditionally a Bayesian diagnostic system contains only hypotheses for specific disease outbreaks and for the non-outbreak condition. However, I propose to also represent all the diseases (known and unknown) that are not being modeled by a given set of disease-specific disease outbreaks. For example, such an outbreak disease could be smallpox, if smallpox is not modeled in the set of disease-specific disease outbreaks. As another example, such a disease could be a new infectious disease that has never been seen before.

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

## 2.2 OVERVIEW OF OUTBREAK-DETECTION ALGORITHMS

This dissertation describes the Bayesian anomaly-detection algorithm in the context of disease-outbreak detection. Disease-outbreak detection is an important application domain in anomaly detection, which is also referred to as “biosurveillance”. The term “biosurveillance” denotes disease surveillance practiced by public health organizations and many other organizations that monitor for disease, such as hospitals, agribusinesses, and zoos (Wagner 2006). Electronic biosurveillance refers to the systematic collection and automated analysis of electronically available data with the intent of detecting outbreaks of disease rapidly (Mandl 2004). These electronic data come from emergency department (ED) visits, over-the-counter medication sales, ambulatory care visit records, and other sources. The goal of disease outbreak-detection algorithms is to detect outbreaks early while exhibiting few false alerts.

This section provides a brief overview of some commonly used outbreak-detection algorithms, which include both specific and non-specific detection approaches. Each detection approach is categorized as either a frequentist method or a Bayesian method. I focus in more detail on reviewing Bayesian methods, since the anomaly-detection algorithm I developed uses a Bayesian approach.

Existing disease outbreak-detection algorithms can be further categorized as temporal detection algorithms, spatial detection algorithms, and spatio-temporal detection algorithms. Temporal methods operate on aggregate data that are measured only with respect to time in order to detect unusual temporal patterns. Spatial methods involve accumulating data over some time interval, removing the time information, and then searching for areas of unusually high incidences of events. Spatio-temporal methods use spatial and temporal data to look for areas of unusually high incidences of events. Because the detection algorithm in my dissertation research is a temporal method, in this section I only review temporal detection algorithms. Examples of the spatial and the spatial-temporal detection algorithms include the spatial scan statistic (Kulldorff 1997) and the spatio-temporal scan statistic (Kulldorff 2001). Sonesson et al. (Sonesson 2003) and Buckeridge (Buckeridge 2005; Buckeridge 2007) provide good reviews of these types of methods.

### 2.2.1 Non-disease-specific detection methods

This section provides an overview of commonly used non-disease-specific detection methods.

**2.2.1.1 Frequentist approaches** Many non-specific detection algorithms use frequentist statistical techniques. Most of these algorithms operate on a univariate time series of aggregate counts of some event. That is, data only contain one piece of information per time step, as for example the number of patients with respiratory symptoms appearing at an emergency department per day. The majority of univariate algorithms model non-outbreak background activity first and then detect outbreaks by examining marked deviations from that background activity.

The frequentist univariate algorithms include methods from statistical quality control (Hutwagner 2003), regression (Serfling 1963), time series models (Reis 2003), and wavelets (Goldenberg 2002; Zhang 2003). These methods are reviewed in (Wong 2004) and (Moore 2003). Some univariate detection methods described in the following sections, such as EWMA (Lowry 1992) and CUSUM (Crosier 1988), could be extended to multivariate versions.

1) Statistical quality control

Statistical quality control is an effective method of monitoring a process through some statistic of a production process. Many of these techniques have been used in the field of disease surveillance.

A *Control Chart* is one of the simplest and the most commonly used statistical quality control methods for surveillance (Lawson 2005). The observations in the background activity are assumed to follow a normal distribution. The control chart consists of a center line, which is drawn as the process mean, and it also includes upper and lower control limits that indicate the threshold at which the process output is considered statistically unlikely (Banks 1989). If any point falls outside either the upper or the lower control limits, the process is considered to be out of control. Technically, the control chart method is not a temporal algorithm, because it only uses the information about the process contained in the current observation and the decision depends solely on the current observation. It has been shown that the control chart method performs relatively poorly for small and moderate shifts.

*CUSUM* works by maintaining a cumulative sum of deviations from the mean of a univariate count. If the cumulative sum exceeds a threshold, the process is considered to be out of control and an alert is raised. It has been shown to be more effective than a control chart at detecting small shifts from the mean of a process. For large shifts, Frisén and Wessman (Frisen 1999) showed that the CUSUM method converges to the control chart method.

*EWMA* (Exponentially Weighted Moving Average) utilizes all information contained in every observation in the process. It is an average of all the observations that have multiplying factors, which decrease exponentially to give different weights to different data points. Specifically, the weighting of each older data point decreases exponentially, giving more importance to recent observations while still not discarding older observations entirely. If the moving average exceeds some upper control limit, the process is considered to be out of control. According to (Montgomery 1991), EWMA is better at detecting small shifts in a process than the control chart method, and it has a performance similar to the CUSUM algorithm. It is also highly insensitive to the normality assumption.

2) Regression

Regression is a technique that is applied widely in biosurveillance to model background activity while accounting for seasonal trends and day-of-week effects. It forecasts the value for the current day. If the observed value is significantly different from the forecast value, then an alert is raised.

*The Serfling method* is a cyclical (often annual) linear regression method to model temporal patterns of disease (Serfling 1963). The model parameters are obtained through regression on a training set of non-outbreak periods using the standard least squares method.

*The generalized linear mixed model* (GLMM) extends the original logistic regression model by including random effects in the predictors (Lazarus 2002). Logistic regression is a regression model for binomially-distributed dependent variables (Hosmer 1989). Lazarus et al.

developed a system for monitoring ambulatory-care encounter records, and they used a general linear mixed model to estimate the daily counts for each census tract for each syndrome. In developing GLMM, Lazarus first estimated the model parameters from historical data. Then, the GLMM was applied to data of ambulatory-care encounter records from the current day in order to calculate the probability $p_{it}$ of an encounter with a diagnosis in the syndrome monitored for that census tract $i$ and day $t$. Once $p_{it}$ is determined, the expectation of the number of ambulatory-care encounters for census tract $i$ and day $t$ can be calculated. The method calculates the probability of seeing an extreme number of counts on day $t$ assuming that the number of the observed counts on day $t$ is binomially distributed.

3) Time series models

The regression models assume independence over the sequence of observations, thereby making it difficult to model observations that are correlated to each other in time. Techniques in time series analysis have been developed to handle data with correlation between data points in time, seasonality, cyclic components, and non-stationarity, which makes time series models applicable to the task of disease surveillance.

*ARIMA* (AutoRegressive Integrated Moving Average) was first introduced by Box and Jenkins (Box 1976). It makes no assumption of independent data and it is able to describe historical visit rates and account for temporal dependency, secular trends and seasonal effects. An ARIMA($p$, $d$, $q$) process is obtained by integrating an ARMA($p$, $q$) model, where $d$ is a positive integer that controls the level of differencing (if $d$ = 0, this model is equivalent to an ARMA model) and $d$ is also called the order of integration. The ARMA($p$, $q$) model consists of two parts: an autoregressive part of order $p$, and a moving average part of order $q$. Reis and Mandl provide a review of the ARIMA model for biosurveillance in (Reis 2003).

4) Wavelets

Wavelets are a popular pre-processing approach to smooth time series data. Wavelet transforms preserve both the temporal and the frequency information of the signal. In the context of disease surveillance, real datasets exhibit features such as noisiness, periodic variations, and long-term trends that do not vary periodically. Wavelets are excellent tools for modeling such time series (Goldenberg 2002). They work by forecasting the current data value (e.g., the expected emergency department visits) from historical data and then comparing the forecast with the actual value. Zhang (Zhang 2003) applied a multiresolution-based predictor in order to forecast the current data value, i.e., independently forecasting by predictors from other resolutions. The expected value for the current value is obtained by summing all the predictions from all resolutions. Some common prediction models, such as the autoregressive model (Goldenberg 2002), can be applied to each individual resolution for prediction.

5) Change-point statistics

The change-point statistics method (Carlstein 1988) is a popular tool in the statistics literature for detecting when an underlying process changes in terms of the mean or other measures of location. This technique considers a sequence of independent random variables $\{X_i: 1 \leq i \leq n\}$ having cdf $F$ for $i \leq \theta n$ and cdf $G$ otherwise, that is, a time series of signals in which each observation is generated independently from some fixed but unknown distribution $F$ until a certain unknown time $T = \theta n$, after which they are instead generated independently from some fixed but unknown distribution $G$. The objective is to estimate the change-point $\theta \in (0,1)$.

**2.2.1.2 Bayesian approaches** Bayesian non-disease-specific approaches for anomaly detection include hidden Markov models (Rabiner 1989), Kalman filters (Hamilton 1994), Bayes sequential change-point techniques (Shiryaev 1978), dynamic linear models (West 1989), and Bayesian clustering (Banfield 1993).

*Hidden Markov models* (HMMs) represent a sequence of observations that emanate from a chain of unobserved discrete variables (Rabiner 1989). LeStrat and Carrat (LeStrat 1999) applied HMMs to detect an influenza-like illness (ILI) from a univariate time series of ILI data. They 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. (Rath 2003) analyzed the same dataset and showed that better detection accuracy can be achieved by modeling the epidemic rates with a Gaussian distribution and the non-epidemic rates with an exponential distribution. They both used the EM algorithm (Dempster 1977) on historical data to obtain the maximum likelihood estimates for the model parameters.

*Kalman filter* (Harvey 1981) is similar to the HMM approach, but in this method the hidden state is a continuous variable. According to (Buckeridge 2005), adaptive Kalman filters have been used to predict expected values of sales of over-the-counter pharmaceuticals for each day. It has also been applied for monitoring AIDS surveillance data (Stroup 1995).

*Bayes sequential change-point technique* Shiryaev (Shiryaev 1978) formulated the problem of optimal sequential detection of the change point in a Bayesian framework by putting a geometric prior distribution on the unknown change point. Baron (Baron 2002) proposed a hierarchical Bayesian change-point model for influenza outbreaks. Non-outbreak and outbreak phases of influenza are modeled as Gaussian distributions with different parameters. These parameters were estimated from historical data. Prior probabilities of a change point depend on (random) factors that affect the spread of influenza.

*Dynamic linear models* (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 arrival of the first observation of the time series. Nobre et al. (Nobre 2001) modeled the stochastic trend and the seasonal effect of a time series of reported number of cases of hepatitis A and malaria for the United States using the linear growth models described in (West 1989). They used a non-informative prior, a normal distribution with mean zero and a large variance, as the initial prior for the vector of model parameters that are related to the seasonal trend and the stochastic trend. Palomo (Palomo 2005) described a similar method but using initial priors specified by experts.

*Bayesian clustering* considers probability models for partitions of a set of *m* elements, that is, it determines the component membership of these elements. Models are specified in terms of the conditional probability of either joining an already existing cluster or forming a new one. The model-based clustering (MBC) approach (Dasgupta 1998) is an example of the Bayesian clustering method. The MBC method was originally proposed by Banfield and Raftery (Banfield 1993) for clustering of *d*-dimensional data based on a mixture of Gaussian distributions. Dasgupta and Raftery (Dasgupta 1998) later extended this approach and used the EM algorithm for estimating parameters. In model-based clustering, clusters are formed based on the posterior probability of component membership for each data point. Given independent observations $X_1$, $X_2$, …, $X_m$, the likelihood for a mixture model with $K$ components is

$$P\left(X_1, X_2, \cdots, X_m \mid K, \boldsymbol{\theta}\right) = \prod_{i=1}^{m} \sum_{j=1}^{K} \tau_j f_j\left(X_i \mid \theta_j\right),$$

where $\tau_j$ is the probability that an observation belongs to the $j$th component, with $\tau_j \geq 0$ and $\sum_{j=1}^{K} \tau_j = 1$, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ are component-specific parameters.

The MBC method based on multivariate Gaussian mixture models has been used in applications such as identification of flaws in textiles from images (Campbell 1997; Campbell 1999), detection of minefields and seismic faults (Dasgupta 1998), and classification of astronomical data (Mukherjee 1998).

### 2.2.2 Disease-specific detection methods

This section provides an overview of commonly used disease-specific detection methods.

**2.2.2.1 Frequentist approaches** To my knowledge, there is no frequentist detection algorithm designed to monitor for a specific disease, such as inhalational tularemia, although frequentist algorithms have been applied to detect particular syndromes, such as respiratory illness.

**2.2.2.2 Bayesian approaches** In the context of disease outbreak detection, disease-specific detection algorithms are designed to look for anticipated anomalous patterns in data with the purpose of monitoring for a particular disease outbreak. This section provides a brief overview of two disease-specific detection algorithms: BARD (Hogan 2007) and PANDA (Cooper 2004), which were both designed to monitor an outbreak of inhalational anthrax due to aerosol releases of *B. anthracis* spores.

*BARD* (Bayesian Aerosol Release Detector) analyzes both medical surveillance data and meteorological data for early detection and characterization of outdoor releases of *B. anthracis*. The approach is general and could be applied to outbreaks due to other biological agents that can be disseminated by outdoor aerosol release. BARD performs a combined analysis of meteorological data and medical surveillance data using models of both. The system computes the posterior probability of a release given these data. It also computes a posterior distribution over the release location, quantity, and time. In addition, BARD is used for simulating outbreaks due to inhalational anthrax. Simulated anthrax cases generated by BARD were used in (Cooper 2004).

*PANDA* (Population-wide ANomaly Detection and Assessment) uses a multivariate Bayesian approach to biosurveillance. In particular, a causal Bayesian network is used to model an outbreak due to the windborne spread of anthrax. Unlike BARD, however, PANDA does not contain a meteorological model. Each person in the population being monitored for an outbreak is explicitly modeled using a subnetwork. Specifically, each person in the population is represented using a six-node network structure that includes disease status, patient symptoms, and other personal information, while avoiding any information that could personally identify the individual (e.g., name, social security number, and home street address). The primary clinical information about each person is whether he or she presented to the ED with a respiratory chief complaint (e.g., a cough). The subnetworks are connected through a common set of nodes that represent the disease outbreak conditions, such as the hypothesized location and time of release of anthrax spores.

Given current data about individuals in the population, inferences are performed on a Bayesian network to derive the posterior probabilities of outbreak diseases in the population. Since the resulting Bayesian network requires millions of nodes to model a medium-sized U.S. metropolitan population, PANDA uses several optimization methods to keep the model size manageable and the inference time tractable. In particular, if the subnetworks representing one or more individuals are identical in structure and parameters, they are modeled using a single subpopulation subnetwork, which is called an *equivalence class*. Furthermore, PANDA uses an incremental updating method that only updates the network state based on new information about an individual in the population.

Cooper et al. (Cooper 2006) further extended PANDA and developed an outbreak detection system called MD-PANDA (Multiple-Disease PANDA) that was applied to monitor for outbreaks of CDC Category A diseases (anthrax, botulism, plague, small pox, tularemia, and viral hemorrhagic fevers) and several other diseases, namely, influenza, cryptosporidiosis, hepatitis A and asthma. MD-PANDA takes as input a time series of 54 possible emergency department chief complaints, and each hour it outputs the posterior probability of the outbreak diseases mentioned above.

According to the Bayesian non-disease-specific and disease-specific methods described above, choices of prior distributions can generally be divided into three categories: (1) estimating the prior from historical data such as hidden Markov models, Bayes sequential change-point techniques, and Bayesian clustering techniques, (2) using non-informative priors, then sequentially updating the priors when new data becomes available, as in dynamic linear models, and (3) assessing the priors using expert knowledge such as PANDA and BARD detection methods.

### 2.2.3 Methods combining disease-specific and non-disease-specific detectors

As described in Chapter 1, a disease-specific detection method and a non-disease-specific detection method each has its own advantages and disadvantages in detecting different patterns of anomalous events.

A specific detection method is designed to monitor for a specific type of disease outbreak, thus making it better at detecting that specific type of outbreak or a similar one.

Specific detection methods are readily implemented using Bayesian methods (e.g., the specific detection algorithms described in Section 2.2.2.2), which allow us to incorporate expert assessments and any prior knowledge we may have. The probabilistic outputs of the specific detection algorithms can be further used in the public health community to guide the response and actions of the public health officers. However, specific detection methods, as the name implies, may not be as effective at detecting an outbreak disease that is unanticipated or unknown.

Non-specific detection methods are designed for monitoring a wide range of anomalous patterns in data, but they may be less effective at detecting any particular disease, relative to explicitly modeling that disease. Most non-specific detection algorithms are implemented using frequentist approaches, as described in Section 2.2.1.1. Bayesian approaches have been developed that can also be applied to perform non-specific anomaly detection as described in Section 2.2.1.2.

To my knowledge, there is no detection system that combines specific and non-specific detectors. The proposed BH algorithm is a Bayesian approach that combines specific and non-specific detectors such that the algorithm is vigilant in watching for anomalous events due to both known and unknown causes. This dissertation describes the BH algorithm in the context of disease-outbreak detection.

## 2.3 EVALUATION OF OUTBREAK-DETECTION ALGORITHMS

Outbreak detection algorithms are designed to detect outbreaks rapidly by monitoring routinely collected data and raising an alert upon discovery of any significant deviations from the norm.

These algorithms are intended to exhibit few false alerts and to detect outbreaks early. Therefore, evaluations on the effectiveness and timeliness of the algorithms are important. Many metrics have been used to evaluate the performances of outbreak detection algorithms (Table 2.1).

**Table 2.1** Evaluation methods used for outbreak detection algorithms

| Evaluation methods | Description |
|---|---|
| ARL0 | The expected time until the first false alarm |
| ARL1 | The expected delay before detecting an outbreak when there is an outbreak at the initiation of surveillance |
| CED(*t*) | The expected delay before detecting an outbreak when the outbreak occurs at time point *t* |
| PSD(*d*,*t*) | The probability that the outbreak is detected with a delay no longer than *d* |
| PV(*t*) | The probability that the outbreak occurs when an alarm is triggered |
| Sensitivity | The probability of an alarm given that an outbreak occurs |
| False positive rate | The probability of an alarm given that an outbreak does not occur |
| ROC curve | The curve defined by plotting sensitivity versus false positive rate |
| AMOC curve | The curve defined by plotting the expected time to detection (since the outbreak began) versus false positive rate |
| FROC curve | The curve defined by plotting the fraction of outbreak locations detected versus false positive rate |

There are a group of quality assurance measures that have been used to characterize the behavior of outbreak detection algorithms (Sonesson 2003). Evaluation studies of change-point detection algorithms tend to employ these metrics.

- The average in-control run length (ARL0) measures the expected time until the first false alarm, which, in another way, estimates false positive alerts per time. However, this measure is only a method for estimating in-control performance and does not address the effectiveness of the algorithm after an outbreak occurs.

- The average out-of-control run length (ARL1) measures the expected delay before detecting an outbreak when there is an outbreak at the initiation of surveillance.
- The conditional expected delay as a function of the change point (CED(*t*)), which measures the expected delay before detecting an outbreak when the outbreak occurs at time point *t*.

All the three methods evaluate the timeliness of an outbreak-detection algorithm, but do not measure the accuracy of an outbreak-detection algorithm.

- The probability of successful detection (PSD(*d*,*t*)) as a function of a limited delay time *d* and a change point *t*, which is the probability that the outbreak is detected with a delay no longer than *d*. In the applications of disease outbreak detection, especially of an infectious disease where only a limited delay time can be tolerated, this measure is particularly useful.
- The predictive value of an alarm (PV(*t*)), as a function of the time of the alarm *t*, is the probability that an outbreak occurs when an alarm is triggered. If an alarm is triggered, various preventive actions should be taken. This measure allows us to be able to choose what actions to take by estimating how much trust to put in an alarm.

Both methods above do not evaluate the probabilities that an outbreak detection algorithm raises a false alert under the non-outbreak condition.

Other common performance measures include the direct reporting of sensitivity and false positive rates (*fpr*). Sensitivity (true positive rate) is the probability of an alarm given an outbreak, and it estimates the chance that a future outbreak will ever be detected. False positive rate is the probability of an alarm given that there is no outbreak, which calculates the expected number of false alerts of an outbreak per unit time.

When calculated over a range of parameter settings for an algorithm, the set of sensitivity/false positive rate values can be plotted to determine the receiver operating characteristic (ROC) curve (Burkom 2003). The ROC curves are commonly used summaries for assessing the tradeoff between sensitivity and *fpr*.

The ability to detect outbreaks in a timely manner is an issue of central importance (Mostashari 2003). While the sensitivity, *fpr* and ROC curves summarize the ability of an algorithm to detect outbreaks, they do not evaluate the timeliness of detection. The Activity Monitoring Operating Characteristic (AMOC) curve (Fawcett 1999), which is adopted from the area of financial transaction, plots the expected time to detection (since the outbreak began) versus the false positive rate. Outbreaks have different lengths and could be undetected by the algorithm. Therefore sometimes median time to detection is used instead of mean time for robustness (Siegrist 2004). One possible approach for missed detections is to use the duration of the outbreak, or the interval between the beginning of the outbreak and likely detection by another means, such as reporting by clinicians or public health officers.

A limitation of all above metrics is that they do not evaluate the ability of an algorithm to identify the geographic location of an outbreak. Free Response Operating Characteristic (FROC) curve (Chakraborty 2002) is a method to evaluate the spatial accuracy of an algorithm, which is defined by plotting the fraction of outbreak locations detected against false positive rate. But this approach does not evaluate the timeliness of detection either.

Since this dissertation reports many of its experimental results using AMOC curves, which were mentioned above, I provide additional detail about them in the remainder of this section. Activity monitoring involves monitoring the behavior of a large population of entities for interesting events requiring action. As mentioned, the AMOC curve plots the expected time to detection (since the outbreak began) versus the false alert rate. I use the phrase *false alert or alarm rate*, instead of the *false positive rate* in the context of disease outbreak detection, to differentiate the former from measures used in the evaluation of traditional classification problems, in which false positive rate is calculated as the proportion of negative instances that were erroneously reported as being positive. In the application of disease outbreak detection, false alert rate is calculated as the expected number of false alerts per unit time. See (Fawcett 1999) for details regarding how to plot AMOC curves.

## 2.4 CHOICE OF A PRIOR DISTRIBUTION

Modeling an unknown disease using Bayesian methods involves specifying the prior distribution of some characteristics that the disease manifests, as for example the visit rates of patients to the ED with the symptom *fever*. The prior distribution described here is a parameter prior distribution, which will be further used to derive the posterior probability of an outbreak due to the unknown disease.

According to (Castillo 2007), three common choices of priors are conjugate priors, non-conjugate priors, and non-informative priors, in which non-informative priors are usually non-conjugate as well.

### 2.4.1 Conjugate priors

Conjugate priors result in posteriors that have the same type of distribution (with different parameters) as the prior distribution, so conjugacy is a closure property which can simplify computations. For example, the Beta distribution is a conjugate prior to the Binomial likelihood, which yields a closed-form posterior of the Beta distribution with updated parameter values. For a large collection of conjugate priors, see (Gelman 1995). In some applications, “tuning” a conjugate prior to reflect the knowledge of the user is a difficult problem, and even when tuning is done, the conjugate priors might not be able to reflect the user’s knowledge well.

### 2.4.2 Non-conjugate priors

Non-conjugate priors result in posteriors that have a different parametric form than the priors, which does not satisfy the closure property. For example, suppose that $y_1, \ldots, y_n$ are independent samples from a Cauchy distribution with unknown center $\theta$ and known scale 1: $p(y_i \mid \theta) \propto 1 / (1 + (y_i - \theta)^2)$. Suppose the prior distribution for $\theta$ is uniform on [0, 1]. Then the posterior probability $P(\theta \mid y_1, \ldots, y_n)$ is calculated as follows:

$$P(\theta \mid y_1, \cdots, y_n) = \frac{P(\theta)P(y_1, \cdots, y_n \mid \theta)}{\int_0^1 P(\theta)P(y_1, \cdots, y_n \mid \theta) d\theta} \propto \frac{\prod_{i=1}^{n} \left[1 + \frac{1}{1+(y_i - \theta)^2}\right]}{\int_0^1 \prod_{i=1}^{n} \left[1 + \frac{1}{1+(y_i - \theta)^2}\right] d\theta}. \qquad (2.2)$$

Solving the integral shown in Equation 2.2 can be difficult and may require approximation methods such as Markov chain Monte Carlo (MCMC) methods (Andrieu 2003).

### 2.4.3 Non-informative priors

Non-informative priors are sometimes called "objective priors", and the resulting analysis is called objective Bayesian analysis. We use these priors to reflect a situation where there is a relative lack of knowledge about a parameter. Non-informative priors are also called "reference priors" in the sense of being a default choice that is used in a situation when one knows little about a parameter. This section provides an overview of non-informative prior distributions because my dissertation research involves unknown-disease modeling using non-informative priors.

According to (Castillo 2007), a uniform distribution over a subrange of a parameter is informative. Such a prior is non-informative only if the parameter has a range that overlaps with that of the uniform distribution. For example, a uniform distribution on the interval [0,1] for the Binomial proportion parameter *p* is non-informative, since parameter *p* is defined on the range [0,1] that coincides with the range of the uniform distribution on the interval [0,1].

*Proper and improper prior distributions*

One practical problem associated with non-informative priors is the requirement that the posterior distribution be proper. According to (Gelman 1995), a prior distribution $\pi(\theta)$ is proper if it does not depend on data and integrates to 1. Often non-informative priors on continuous, unbounded variables are improper (Lawson 2009). That is, the integration of the prior distribution of the random variable $\theta$ over its range (Ω) is not finite:

$$\int_{\Omega} \pi(\theta) d\theta = \infty .$$

Nonetheless, under the right conditions, improper prior distributions can still lead to proper posterior distributions (Gelman 1995), which is the most critical issue.

This dissertation research uses a non-informative prior (a uniform distribution) to model the parameter, namely the probability of a symptom in a disease, and this parameter has a bounded range on the interval [0,1]. Thus, this non-informative prior is proper.

In most cases, non-informative priors in the medical domain are proper, because the parameters to be modeled in the domain are often bounded in particular ranges due to patients' physical limits, as for example a patient's temperature.

Some methods for constructing non-informative priors are briefly reviewed as follows:

1) Laplace and the principle of insufficient reason

Laplace introduced the *principle of insufficient reason*, in which he suggested using a uniform prior to assign equal probability to each point in the parameter space if the parameter space is finite. When the parameter space is continuous, it is natural to apply the principle of insufficient reason to obtain a flat prior, that is, a prior that is equal to a positive constant.

2) Jeffreys' invariance principle

Jeffreys (Jeffreys 1961) initiated the idea of using a formal rule to define a non-informative prior. He used the concept of invariance as a formal rule. If $\pi(\theta)$ is a non-informative prior for $\theta$ that is derived using some rule, then the same rule should lead to

$$\pi(\varphi) = \pi(\theta)\left|\frac{d\theta}{d\varphi}\right|$$

as a non-informative prior for $\varphi$, where $\varphi = h(\theta)$ is a one-to-one transformation. Jeffreys' general principle is that any rule for determining the prior distribution $\pi(\theta)$ should lead to an equivalent result if applied to the transformed parameter; that is, posterior inferences based on $\pi(\varphi)$ will be the same as those based on $\pi(\theta)$. Jeffreys also described a non-informative prior that meets the invariance principle, which is

$$\pi(\theta) \propto I(\theta)^{1/2}, \tag{2.3}$$

where $I(\theta)$ is Fisher's information for the parameter $\theta$ if observing $y = (y_1, \ldots, y_n)$ of iid observations. $I(\theta)$ is represented as follows:

$$I(\theta) = -E\left[\left.\frac{d^2 \log p(y \mid \theta)}{d\theta^2}\right| \theta\right].$$

3) Maximum entropy

If $\Theta = \{\theta_1, \ldots, \theta_n\}$ is finite and $\pi$ is a probability mass function on $\Theta$, then the entropy of $\pi$, which is meant to capture the amount of uncertainty implied by $\pi$, is defined by $E(\pi) = -\sum \pi(i) \log \pi(i)$. Jaynes (Jaynes 1957; Jaynes 1968) is the main developer of entropy-based methods. Priors with larger entropy are regarded as being less informative, and the method of maximum entropy involves selecting the prior that maximizes the entropy $E(\pi)$. If the problem has no further constraints, then the prior with maximum entropy is the uniform prior.

4) The Berger-Bernardo method

Bernardo (Bernardo 1979) proposed a method for constructing priors based on the notion of missing information. Let $X_1^n = (X_1, \ldots, X_n)$ be a vector of $n$ iid random variables and let $K_n\left(\pi(\theta \mid x_1^n), \pi(\theta)\right)$ be the Kullback-Leibler distance between the posterior density and the prior density, where $K_n\left(\pi(\theta \mid x_1^n), \pi(\theta)\right) = \int \pi(\theta \mid x_1^n) \log\left(\pi(\theta \mid x_1^n) / \pi(\theta)\right) d\theta$.

Let $K_n^\pi = E\left(K_n\left(\pi(\theta \mid x_1^n), \pi(\theta)\right)\right)$ be the expected gain in information. Bernardo's idea was to think of $K_n^\pi$ for a large $n$ as a measure of the missing information in the experiment. Loosely, this is the gain in information provided by the experiment. He suggested constructing a prior that maximizes $K_\infty^\pi = \lim_{n\to\infty} K_n^\pi$. However, $K_\infty^\pi$ is usually infinite. Bernardo found the prior $\pi_n$ that maximizes $K_n^\pi$ to circumvent this problem. He then computed the limit of the corresponding sequence of posteriors and defined the Berger-Bernardo prior as the prior that produces the limiting posterior via Bayes' theorem. When there are no nuisance parameters and certain regularity conditions are satisfied, this prior turns out to be the prior shown in Equation 2.3 for continuous parameter spaces and the uniform prior for finite parameter spaces.

Kass and Wasserman (Kass 1996) provide an excellent review of all the above methods for establishing non-informative priors. Besides the above methods, they also discussed methods based on data-translated likelihoods (Box 1973), coverage matching methods (Welch 1963), Zellner's method (Zellner 1977), and decision-theoretic methods (Hartigan 1965), as well as others. They have found that several of the methods rely specifically on asymptotic theory when sample sizes are very large (relative to the number of parameters being estimated). For example, Jeffreys' invariance principle, the Berger-Bernardo rule, and coverage matching methods are all built from asymptotic theory. They concluded that these methods all lead to Jeffreys' invariance principle or some modification of it.

### 2.4.4 Why use the Beta distribution as a prior belief?

The Beta distribution is a continuous probability distribution that is parameterized by two positive shape parameters ($\alpha$ and $\beta$). The Beta distribution has been used for a wide variety of applications because it can flexibly specify a range of forms of distributions from peaked ($\alpha$, $\beta >$ 1) to uniform ($\alpha = \beta = 1$) and from U-shaped ($0 < \alpha$, $\beta < 1$) to skewed or either monotonically decreasing or increasing (Gelman 1995).

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 I do in this dissertation.

There is a long history of using the Beta distribution to represent belief about a relative frequency. In the 19th century G.F. Hardy (Hardy 1889) and W.A. Whitworth (Whitworth 1897) proposed quantifying prior beliefs with Beta distributions.

Neapolitan (Neapolitan 2003) presented two arguments for why we should use Beta distributions. First, consider the two concepts of a *relative frequency* and an *updated density function*. Neapolitan used the term *relative frequency*, as opposed to the term probability, to represent the propensity that an event happens, while the term *probability* is used to refer to a subjective probability (degree of belief). Given a sample, the density function of the relative frequency conditional on data is called the updated density function of the relative frequency.

The first argument that Neapolitan presented is that if we initially consider all numbers in [0, 1] equally likely to be the value of a relative frequency and therefore use the uniform distribution to represent our prior beliefs, it is a mathematical consequence of the theory that the updated density function is Beta (Casella 2002).

The second argument is a theorem proved by Zabell (Zabell 1982), which states that if the assumptions of exchangeability and Johnson's sufficientness postulate hold, then that individual must use the Beta distribution to quantify any prior beliefs about a relative frequency. Zabell's theorem actually is concerning the Dirichlet distribution, which is a generalization of the Beta distribution.

I now present the formal definition of exchangeability in terms of a finite and infinite sequence of random variables, respectively, and Johnson's sufficientness postulate, as below:

*Let D =* $\{X^1, X^2, \ldots, X^M\}$ *be an ordered set (sequence) of random variables, each with the same r alternatives* 1, 2, …, *r. If for every two sets of values d' and d" such that each alternative occurs the same number of times in d' and d", we have* $P(D = d') = P(D = d")$*, the sequence is said to be finite exchangeable.*

*Let* $X^1, X^2, \ldots$ *be an infinite sequence of random variables, each with the same r alternatives* 1, 2, … *r. If for every M, the sequence of the first M variables is finite exchangeable, the sequence if said to infinite exchangeable.*

*Let* $X^1, \ldots, X^M, X^{M+1}$ *be a sequence of random variables, each with the same r alternatives* 1, 2, …, *r, and let D =* $\{X^1, X^2, \ldots, X^M\}$*. Suppose for every set of values d of the variables in D, we have*

$$P(X^{M+1} = k \mid D = d) = g_k\,(s_k, M),$$

*where* $s_k$ *is the number of times k occurs in d. That is, the probability the next variable is equal to k is a function* $g_k$ *only of how many times k occurs in d and the number of variables in D. Then Johnson's sufficientness postulate is said to hold for the sequence.*

Note that the use of the Dirichlet distribution to quantify our prior beliefs on relative frequencies concerns only the case in which we know the number of values of the variable in advance. For example, we know a die can land six ways, and we know a patient either does or does not have anthrax. In this dissertation research, expert judgments are used to quantify the prior beliefs on the relative frequency of incidence of a disease symptom for an outbreak disease, as for example the relative frequency of cough in anthrax, which is modeled as a binary variable. Thus, the Beta distribution, which is a special case of the Dirichlet distribution, can be used in this dissertation research to quantify the prior beliefs.

Using the Beta distribution as a prior for the Binomial likelihood, Tuyl et al. also suggest using the Bayes-Laplace Beta(1, 1) as the natural representation of the prior ignorance (Tuyl 2009). They argued that the posterior predictive distribution resulting from Beta-Binomial (or multinomial) models, when viewed via a hypergeometric-like representation (Weisstein 2009), suggests the uniform prior (Beta(1, 1)) on the proportion parameter for these models. Let $P(y \mid m, x, n)$ be the posterior probability of $y$ successes in $m$ trials given $x$ successes in $n$ trials, and let $P(x \mid n, y, m)$ be defined analogously. They further expressed $P(y \mid m, x, n)$ using a hypergeometric-like representation, which suggests the uniform prior Beta (1, 1), leading to symmetry in $y$ and $x$. That is, Beta(1, 1) prior reflects the fact that $y$ and $x$ have $m + 1$ and $n + 1$ possible values, respectively, and the ratio of $P(y \mid m, x, n)$ and $P(x \mid n, y, m)$ is 1 when $m = n$ so that $P(y \mid m, x, n)$ equals $P(x \mid n, y, m)$. These facts seem reasonable, but follows from the Beta(1, 1) prior only.

### 2.4.5 Empirical Bayes methods

A prior distribution for an unknown parameter $\theta$, $\pi(\theta)$, has its own parameters, usually referred to as hyperparameters. The hyperparameters can be obtained, for example, by assessing expert opinions. According to (Castillo 2007), hyperparameters can also be assumed to be “random variables that themselves have a prior distribution in an additional level of a hierarchy of parameters and prior distributions that can be repeated at several levels up to the highest level in the hierarchy in which all parameters are assumed known”. Empirical Bayes methods use data to estimate the hyperparameters by applying techniques such as maximum likelihood (Lehmann 1998) or the method of moments (Gelman 1995). Many Bayesian detection algorithms reviewed in Section 2.2.1.2, such as the hidden Markov models, Kalman filters, and the model-based clustering method, use empirical Bayes methods to estimate the hyperparameters.

### 2.4.6 Mixture priors

A finite *M*-component mixture prior (Murdoch 2001) is defined as $\pi(\theta) \propto \sum_{i=1}^{M} \tau_i \pi_i(\theta)$, where $\tau_i$ is the mixing parameter, $\pi_i(\theta)$ is the prior distribution for the *i*th component, and $\sum_{i=1}^{M} \tau_i = 1$. Gajewski and Mayo (Gajewski 2006) use a mixture of informative prior distributions, which come from several sources of information, for example, from two (or more) clinicians. In particular, they used a mixture of Beta distributions. Because the Beta distribution is a conjugate prior for the Binomial likelihood, it is easy to show that a Binomial likelihood with a prior of a mixture of Beta distributions has a posterior distribution that is also a mixture of Beta distributions. As described in the section on conjugate priors, this conjugacy results in a closed-form posterior, which simplifies computation.

In this dissertation research, I have used a non-informative prior to model unknown diseases and a mixture of priors to model partially-known diseases. The mixture of priors consists of priors that are informative for known (modeled) diseases and non-informative for conditions we know little about. I thus call this mixture of priors a semi-informative prior.

As described in Chapter 1, the disease modeling problem is essentially a problem of choosing a prior. Accordingly, this dissertation explores the disease modeling space spanning from using non-informative priors to semi-informative priors to informative priors.

## 2.5 UNKNOWN-OBJECT REPRESENTATION

As this dissertation research involves unknown-disease modeling, this section provides a brief review of some prior research related to unknown objects and unknown-object modeling. In particular, several methods are reviewed in this section, including probabilistic graphical models, first-order probabilistic languages (FOPLs), Bayesian Logic (BLOG), and Nonparameteric Bayesian Logic (NP-BLOG).

A general problem in AI is that intelligent systems must represent and reason about objects, but those objects may not be known a priori and may not be directly and uniquely identified by a perceptual process. Many AI applications must deal with unknown objects.

One such application is *population estimation*, as for example estimating the population of a certain animal species in an area by marking animals that are caught in one sweep through the area and looking at the fraction of marked animals in a subsequent sweep (Borchers 2002). The goal of this application is to estimate the number of unknown objects (Milch 2006).

Another application is solving a *multiple-target tracking* problem. One central problem in multiple-target tracking is *data association*, i.e., the problem of determining the correct correspondence between measurements and existing tracked objects, as for example radar blips and hypothesized aircrafts (Bar-Shalom 1988). This application differs from the population estimation problem in that one is no longer interested in the number of individual objects; instead, tracking individual objects is the goal.

### 2.5.1 Probabilistic graphical models

A graph comprises *nodes* connected by *arcs* (or *edges*). Probabilistic graphical models are graphs in which nodes represent random variables, and the arcs represent probabilistic relationships between these variables. Hence, they provide a compact representation of joint probability distributions over all of the random variables (Bishop 2006).

Uncertainty in data association has been addressed using probabilistic graphical models since the 1960s (Sittler 1964). Consider probabilistic models used for multiple-target tracking. They represent targets moving over time and model dependency between observations and the true target positions. If we observe the positions of blips on a radar screen over several time steps, we can use such a model to compute the probability that two particular blips came from the same aircraft. However, as the number of observations increase, performing such computations becomes computationally intractable, because we need to consider all possible mappings between observations and underlying objects.

Standard graphical models are propositional rather than first-order. That is, they do not support quantification over objects. For example, if we want to reason about the disease status of *N* people in the population, we need to repeat the portions of the graph that deal with one person *N* times.

### 2.5.2 First-order probabilistic languages (FOPLs)

There has also been significant work on first-order probabilistic languages (FOPLs) that can represent multiple objects and the relations between them (Gilks 1994; Koller 1998). These languages allow us to define indexed families of random variables. For example, if we want to define the disease status of person *i* on day *t*, we can use disease_status(*i*, *t*). However, most FOPLs assume that the only objects that exist are the ones explicitly defined. This assumption restricts FOPLs to be used in multiple-target tracking applications, where we might want to reason about objects that are not observed at all.

### 2.5.3 Bayesian logic (BLOG)

Milch, et al. (Milch 2007) developed a representation language called Bayesian LOGic (BLOG) that defines probability distributions over outcomes with varying sets of objects and unknown objects. BLOG defines generative models that create first-order model structures by adding objects and setting function values. A model structure corresponds to a possible world, which is constructed by adding objects iteratively and even recursively via *generating functions*. This generative process allows a BLOG model to define a varying and unbounded number of objects.

The unknown object that a BLOG model defines involves number uncertainty, existence uncertainty, and identity uncertainty. I describe these types of uncertainty using examples of a simplified version of the multi-target tracking problem and a citation-matching problem. See (Milch 2007) for a detailed description of these examples.

Example 2.1: *An unknown number of aircraft exist in some volume of airspace. An aircraft's state (position and velocity) at each time step depends on its state at the previous time step. We observe the area with radar: aircraft may appear as identical blips on a radar screen. Each blip gives the approximate position of the aircraft that generated it. However, some blips may be false detections, and some aircraft may not be detected. Which aircraft exists, and what are their trajectories? Are there any aircraft that are not observed?*

Number uncertainty involves dealing with an unknown number of objects. In Example 2.1, the number of aircrafts may vary in possible worlds. In the generative process, number uncertainty corresponds to a number statement with a single generating function, such as the statement #Aircraft ~ NumAircraftPrior() (on line 3 of Figure 2.1) that represents sampling the number of aircrafts in the area.

```
type Author; type Pulication; type Citation;

#Author ~ NumAuthorsPrior();
#Publications(Author = a) ~ NumPubsPrior();

Name(a) ~ NamePrior();
Title(p) ~ TitlePrior();

PubCited(c) ~ Uniform({Publication p});

Text(c) ~ NoisyCitationGrammar(Title(PubCited(c)),
                                           Name(Author(PubCited(c))));
```

**Figure 2.2** BLOG model for the citation-matching problem.

Existence uncertainty is modeled with generating functions. The generative process constructs a world by adding objects whose existence and properties depend on those of objects already created. In such a process, the existence of objects may be determined by many random variables, not just a single population-sized variable. Lines 1-2 of Figure 2.1 represent *origin function declarations*. An origin function must take a single argument of some type (namely Blip in the example). Generative steps that add objects to the world are described by number statements, such as line 7 of Figure 2.1:

```
#Blip(Source = a, Time = t) ~ DetectionCPD(State(a, t));
```

This statement says that for each aircraft a and time step *t*, the process adds some number of blips, and each of these added blips *b* has the property that Source(*b*) = a and Time(*b*) = *t*.

Identity uncertainty is the most relevant to the uncertainty that is addressed in this dissertation. The citation-matching problem described in (Pasula 2003) involves this type of uncertainty. As for each citation, we would like to recover its true title and authors. For example, the following citations from the CiteSeer database probably refer to the same paper:

> Kozierok, Robin, and Maes, Pattie, A Learning Interface Agent for Meeting Scheduling, Proceedings of the 1993 International Workshop on Intelligent user Interfaces, ACM Press, NY.
>
> R. Kozierok and P. Maes. A learning interface agent for scheduling meetings. In W. D. Gray, W. E. Heey, and D. Murray, editors, Proc. of the International Workshop on Intelligent User Interfaces, Orlando FL, New York, 1993. ACM Press.

The BLOG model for this citation-matching problem is shown in Figure 2.2 with the function declaration statements omitted. BLOG resolves the identity uncertainty by first sampling the total number of authors from some distribution (line 2); then for each author a, sample the number of publications by that author (line 3), using the number statements described above. Then it samples author names and publication titles using their respective prior distributions, e.g., Name(a) ~ NamePrior() and Title(p) ~ TitleDist() in line 4 and 5. Next, for each citation, sample the publication cited by choosing uniformly at random from the set of publications. Finally, generate the citation text using a noisy formatting distribution that allows for errors and abbreviations in the title and author names. See (Milch 2007) for a detailed description of this process.

BLOG is a representational language for defining probability models over objects in the world – it just provides a framework for the generative process that constructs the world.

However, it does not specify explicitly how to estimate the prior distributions for objects of various types, for example, the publication titles and author names in the above citation-matching problem.

The BH algorithm described in this dissertation uses the MD-PANDA model as an experimental domain to construct a “world” for monitoring the disease-outbreak status in the population. It uses a non-informative prior distribution for modeling an unknown outbreak disease that we have almost no knowledge of and a mixture prior distribution for modeling a partially-known outbreak disease that might manifest some disease symptoms, wherein a mixture component is based on the properties of a known disease. The proposed prior distributions described in the dissertation could be used together with BLOG-like representations or with graphical models (e.g., Bayesian networks) to model under uncertainty the unknown causes of events. In this dissertation, I use Bayesian networks.

In addition, compared with the developed inference methods in probabilistic models (e.g., Bayesian networks), much work remains to be done on BLOG. Besides rejection sampling and likelihood weighting algorithms, BLOG needs more efficient and practical inference methods to perform inference on real-world problems.

### 2.5.4 Nonparametric Bayesian logic (NP-BLOG)

As an alternative to explicitly defining a prior distribution over some unknown object, one could use the nonparametric version of BLOG proposed by Carbonetto et al. (Carbonetto 2005), which incorporates Dirichlet process mixture models.

As for the citation-matching problem described in Section 2.4.3, the NP-BLOG model follows a similar generative approach as the BLOG model, the key difference being that it samples collections of unknown objects from Dirichlet processes, and it allows for uncertainty in the order of authors in publications. See (Carbonetto 2005) for a detailed description of the NP-BLOG model for the citation-matching problem.

However, Carbonetto et al. did not specify explicitly how to estimate the prior distributions for objects, such as NamePrior and TitlePrior in Figure 2.2. They postponed the parameterizations to future work.

Both BLOG models and NP-BLOG models have left out one piece of puzzle: how to specify the prior distributions for objects of various types. For the citation-matching problem, Pasula et al. (Pasula 2003) obtained state-of-the-art accuracy using reasonably simple prior distributions for publication titles and author names, estimated from BibTeX files and U.S. Census data. For the prior distributions for the numbers of objects, such as the numbers of authors and publications, Pasula et al. (Pasula 2003) simply used a log-normal distribution, which has a very large variance.

However, in the disease outbreak detection domain, it is sometimes unreasonable for the user to obtain “outbreak data” to estimate the prior distribution of the outbreak diseases due to the lack of actual outbreaks of diseases under surveillance. Thus in the absence of training data on which to base the estimate of prior distributions, a user still needs to model known, partially-known, and unknown diseases.

## 2.6 ENTITY-BASED MODELS FOR RELATIONAL DATA

The proposed BH algorithm uses an experimental domain called MD-PANDA, which is represented using a Bayesian network model (Pearl 2000) in this dissertation. MD-PANDA takes as input relational datasets (Codd 1970) and can also be represented using entity-based models, such as entity-relationship (ER) models, probabilistic entity-relationship (PER) models (Heckerman 2004), and plate models (Buntine 1994). This section provides a brief summary of these models for a relational dataset[2].

### 2.6.1 Relational dataset

Before 1970, there were only "flat" databases, in which information was stored in one long text file. Each entry contained multiple pieces of information (*fields*) about a particular object that were grouped together as a *record*.

The relational database (Codd 1970) was born in 1970. Since then, relational databases have grown in popularity to become a standard. A relational database allows one to easily find specific information. It also allows one readily to sort based on any field and generate reports that contain only certain fields from each record. Relational databases use *tables* to store information. The standard fields and records are represented as columns (fields) and rows (records) in a table.

The dataset in Table 2.1 represents an example relational dataset that MD-PANDA takes as input, where you can quickly group patients with a specific disease symptom, such as cough, because of the arrangement of data in columns.

[2] In what follows, I make no distinction between a database and a dataset.

**Table 2.2** An example dataset that MD-PANDA monitors

| Person ID | Hospital ID | Admission date | Gender | Chief complaint | [More fields…] |
|---|---|---|---|---|---|
| 12345 | 1 | 01/01/2004 | F | Cough | … |
| 67890 | 3 | 01/01/2004 | M | Abdominal pain | … |
| … | … | … | … | … | … |

### 2.6.2 Entity-relationship models

There is a language called the *Entity-Relationship* (ER) *model*, which is commonly used to represent a relational database structure (Ullman 1997).

When building ER models, we distinguish between entities, relationships, and attributes. According to (Heckerman 2004), an *entity* corresponds to a thing or object that is or may be stored in a database; a *relationship* corresponds to a specific interaction among entities; and an *attribute* corresponds to a variable describing some property of an entity or relationship. In the example dataset shown in Table 2.1, the set of individuals contains a sequence of entities (e.g., {person 12345, person 67890, …}). A reference to a set of entities is called an *entity class*. Similarly, *attribute class* refers to an unspecified collection of like attributes. In this example dataset, Person has an attribute class Person.Chief_complaint.

MD-PANDA models the chief complaint state of every person in the population in order to monitor for a possible disease outbreak, and each individual's chief complaint state is observable. It also models an individual's disease state, which is a latent variable that is not observable (from Table 2.1) but needs to be inferred from other observed variables. A high-level Bayesian network model used by MD-PANDA is shown in Figure 2.2, where subnetwork *E* is an entity-based model that models an individual’s disease state and his or her chief complaint state in the population. Thus, MD-PANDA can be considered as an entity-based model.

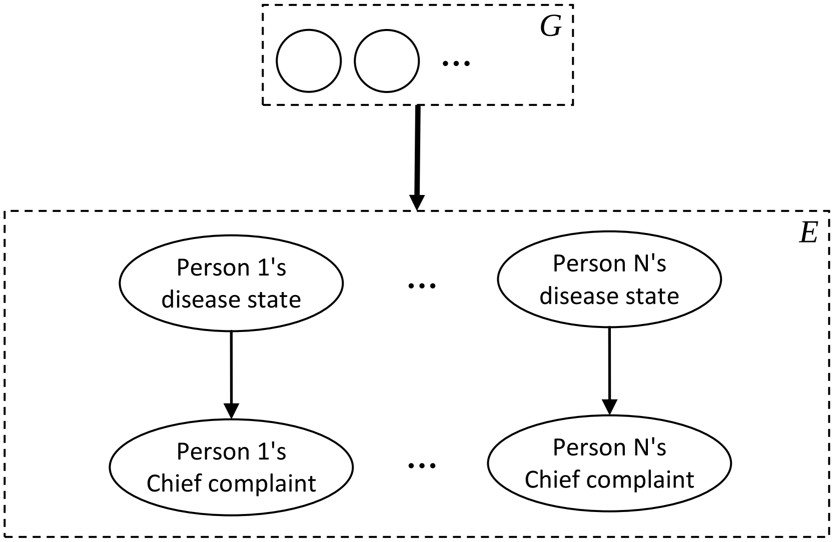


**Figure 2.1** A high-level Bayesian network representation showing the MD-PANDA model, where the dashed rectangular shapes represent subnetworks: subnetwork *G* contains nodes that represent global features common to all people in the population being monitored, and subnetwork *E* represents an entity-based model that contains every entity (Person) in the population; the bolded arrow between subnetworks shows the direction in which the Bayesian-network arcs are oriented between subnetworks, and the arrows between nodes in subnetwork *E* show probabilistic dependencies between nodes.

However, individual’s disease state and chief complaint state can also be explicitly modeled as an entity-relationship pair using an ER model, as shown in Figure 2.3, where we can think of individuals in the population as entities and their chief complaint states as attributes, and individual’s disease as an entity class and his or her disease state as an attribute of this entity class[3]. For example, a possible relationship is the pair (person 12345, disease *cryptosporidiosis*), meaning that person 12345 had disease *cryptosporidiosis*.

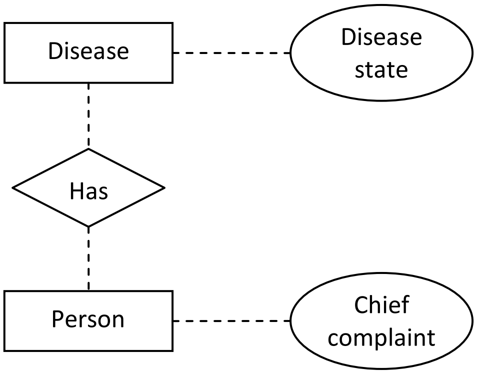


**Figure 2.2** An entity-relationship (ER) model depicting the structure of the relational dataset for Table 2.1. The entity classes (Person and Disease) are shown as rectangular nodes; the relationship class (Has) is shown as a diamond-shaped node; and the attribute classes (Disease.Disease_state, Person.Chief_complaint) are shown as oval nodes. Dashed edges are used for connecting attribute classes to their corresponding entity classes and the relationship class to its associated entity classes.

A relationship class refers to a set of like relationships among entities. In this example, we have the relationship class Has. A relationship class also can have attributes. For example, a relationship class Takes that represents the relationship between entity Student and entity Course can have the attribute Grade. In our example shown in Figure 2.3, the database structure shown

[3] In practice, a user typically models an individual's disease state, such as the stage of one's infection, using a temporal model. However, this dissertation uses a non-temporal model as an experimental domain. Thus, in what follows, I make no distinction between an individual's disease and his or her disease state.

in Table 2.1 does not contain attributes for the relationship class Has[4], therefore the relationship class Has shown in Figure 2.3 is used to easily convey the relationship between entities. Similarly, there is a relationship class Has between entity Person and entity Chief complaint. For simplicity, this relationship is ignored in this figure.

It is important to know that an ER model is a representation of a database structure, such as the database structure shown in Table 2.1, not of a particular database that contains data. That is, an ER model can be built prior to the collection of any data. In contrast, each entity-relationship pair associated with the data has to be explicitly represented in a Bayesian network model. That is, the instantiation of subnetwork *E* can only be developed once we know the exact number of entities (persons) in the database.

### 2.6.3 Probabilistic entity-relationship model

The Probabilistic Entity-Relationship (PER) model is an extension of the ER model as it adds a probabilistic entity-relationship. A specific type of PER model is the directed acyclic probabilistic entity-relationship (DAPER) model (Heckerman 2004), which is an ER model with directed arcs among the attribute classes that represent probabilistic dependencies among corresponding attributes.

Figure 2.4 shows a DAPER model for the example entity and relationship classes of the ER model described above. The DAPER model extends the ER model in Figure 2.3 with the

[4] In practice, the relationship class Has can have the attribute that indicates the time that a person got infected with the disease denoted by the entity Disease. Since this dissertation uses a non-temporal model as an experimental domain, as described above, it is assumed that the relationship class Has does not have attributes.

addition of a solid arc from attribute Disease.Disease_state to attribute Person.Chief_complaint, which represents a probabilistic dependency between the two attributes.

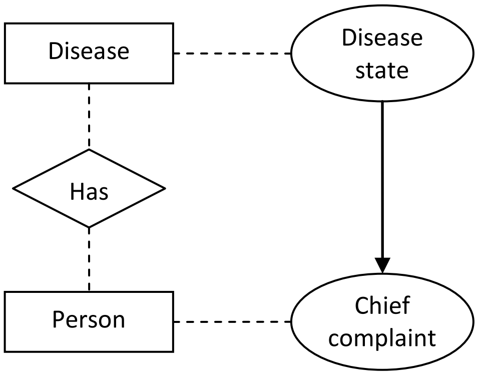


**Figure 2.3** A directed acyclic probabilistic entity-relationship (DAPER) model showing that a person's chief complaint depends on the person's disease state.

Note that the DAPER model shown in Figure 2.4 shows the relationship (conditional independence) between attributes (Disease.Disease_state and Person.Chief_complaint), whereas the ER model shown in Figure 2.3 does not represent the relationship between these two attributes because an ER model is an abstract representation of database structure and this ER model represents the entity-relationship pair based on the database structure shown in Table 2.1.

### 2.6.4 Plate model

A Plate model (Buntine 1994) is a language for compactly representing graphical models in which there are repeated measures. In a plate model, a plate is represented as a large rectangle for labeling an entity class.

Figure 2.5 depicts plate notation of the high-level Bayesian network model shown in Figure 2.2. Instead of drawing each repeated variable (Person's disease state and Person's chief complaint state) individually, a plate or rectangle (shown as bolded rectangle) is used to group variables into a subgraph that repeat together, and a number *N* is drawn on the plate to represent the number of repetitions of the subgraph in the plate, where *N* represents the number of people in the population being monitored. In a plate model, any links that cross a plate boundary are replicated once for each subgraph repetition.

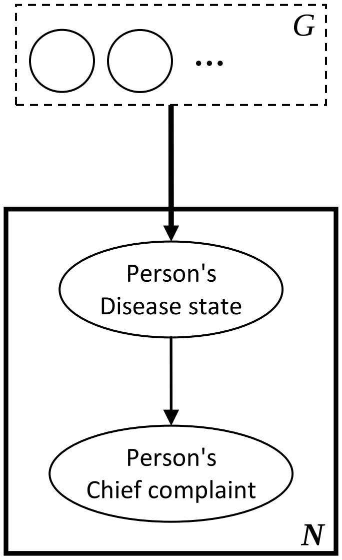


**Figure 2.4** Plate notation for the high-level Bayesian network model used by MD-PANDA.

Unlike in a directed acyclic graph (DAG) model (such as the Bayesian network model shown in Figure 2.2), where an entity-relation pair has to be explicitly modeled, each entity-relation pair in a plate model does not need to be represented explicitly.

Another type of entity-based model is a *Probabilistic Relational Model* (PRMs) (Getoor 2002) that also can be used for representing the entity-relationship encoded in the MD-PANDA model. See (Heckerman 2004) for a detailed review and comparison of PRMs and the models already described above.

Beyond a relational representation, the issue of probabilistic inference in relational representations is also important. Koller and Pfeffer (Koller 1997) have done some preliminary work on performing probabilistic inference using DAPER models. Since a plate model is essentially a Bayesian network model with a different representation notation, inference techniques used by a Bayesian network model can be directly applied on a plate model. Since techniques of probabilistic inference are more developed in a Bayesian network model, this dissertation represents MD-PANDA using a Bayesian network model in what follows.

# 3.0 THE EXPERIMENTAL DOMAIN

The experimental domain for my dissertation research is the disease model used by MD-PANDA. MD-PANDA is a Bayesian detection algorithm that operates on a time series of emergency department (ED) chief complaints for detecting CDC Category A diseases (CDC) (anthrax, botulism, plague, smallpox, tularemia, and viral hemorrhagic fevers) and several other diseases (influenza, cryptosporidiosis, hepatitis A, and asthma). Among these diseases, of which there are ten, anthrax, plague, and viral hemorrhagic fevers each is modeled using its early stage and its late stage, making a total of 13 outbreak diseases. I call these 13 diseases CDC-A$^+$ diseases. Section 3.1 introduces the notation used for the remainder of this dissertation. Section 3.2 provides an introduction to MD-PANDA.

## 3.1 NOTATION

This section introduces the notation used for the remainder of this dissertation. The term ED refers to one or more emergency departments in the region being monitored. If there is more than one, then the total patient cases across all EDs are treated as a single pool.

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

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

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

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

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

Let $j$, $1 \leq j \leq J$, represent the index of a specific disease symptom, where $J$ is the total number of symptoms that are modeled. Note that in the MD-PANDA model, $j$ represents the index of chief complaints. For example, $j = 4$ might represent the binary symptom *cough* that can serve as a chief complaint. MD-PANDA takes as input ED chief complaints (one per patient), while the BH algorithm for my dissertation research takes as input patient disease symptoms, of which there can be more than one per patient.

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

## 3.2 MD-PANDA

MD-PANDA, which is shown in Figure 3.1, is an entity-based Bayesian network model that represents all the people in a given population (not just the ED patients).[5] Recall from Section 2.6.4 that MD-PANDA can also be considered as a plate model. Figure 3.2 shows the plate notation of MD-PANDA model. For a detailed description of the nodes and the conditional probability tables, see the following two sections.

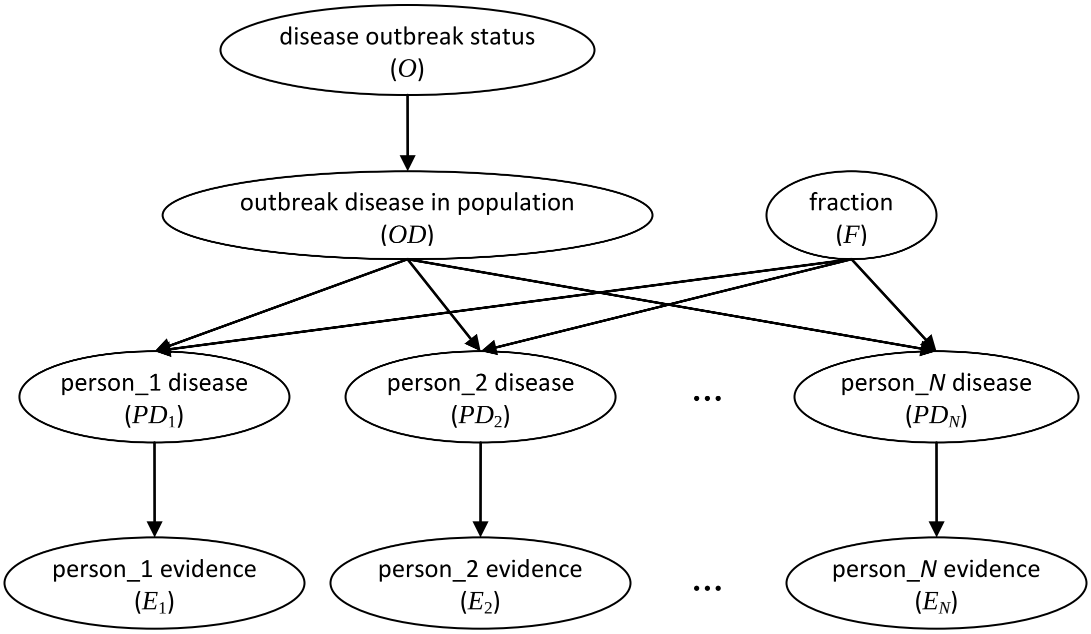


**Figure 3.1** The Bayesian network structure for the MD-PANDA model. See the text next for a description of the nodes and the conditional probability tables.

[5] The particular version of MD-PANDA that models CDC category A diseases has also been called PANDA-CDCA, or PC for short in (Jiang 2008).

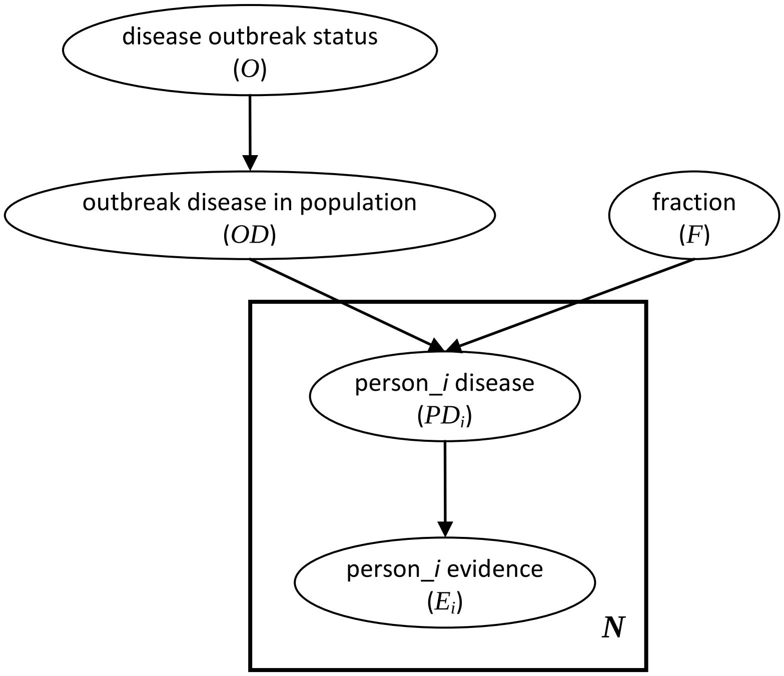


**Figure 3.2** Plate notation of the MD-PANDA model. The subgraph in the plate (bolded box) repeats *N* times, where *N* is the number of individuals in the population being monitored, and any links that cross a plate boundary are replicated once for each subgraph repetition.

MD-PANDA describes a general modeling framework for multiple diseases. In this dissertation, I will be discussing and using the CDCA version (PC) of MD-PANDA, and thus, will only mention PC further.

PC, as described in (Cooper 2006), only examines ED chief complaint data for the most recent 24 hours and does not take into account temporal or spatial information. While temporal and spatial extensions of PC have been developed (Jiang 2008), the algorithm developed in this dissertation research employs the non-temporal non-spatial version of PC as an experimental domain. By doing so, the dissertation can focus more clearly on fundamental issues of modeling both known and unknown diseases, without the added complexity of spatial and temporal modeling, which can be investigated later as extensions. Thus, in this section, I only provide an overview of this version of PC.

### 3.2.1 The nodes

The node *disease outbreak status* represents the outbreak status in the population being monitored 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 $O = OB$, where the ongoing outbreak is due to the outbreak disease denoted by $OD$, otherwise, $O = NOB$.

The node *outbreak disease in population* represents the state of there being an outbreak disease in the population. Let $OD$ denote this node. $OD$ can have the value $d_k$ for $k \geq 1$ (outbreak of known disease $d_k$) or the value *none*, which represents that there is no outbreak disease in the population, among those being modeled.[6] The outbreak diseases $d_k$ that PC models are the CDC category A outbreak diseases listed as follows:

1. early stage anthrax
2. late stage anthrax
3. early stage plague
4. late stage plague
5. smallpox
6. tularemia
7. botulism
8. early stage marburg hemorrhagic fever
9. late stage marburg hemorrhagic fever

[6] PC assumes that different disease outbreaks would not occur simultaneously; however, the model could be extended to allow for multiple disease outbreaks.

The additional diseases modeled in PC are as follows:

1. influenza
2. cryptosporidium
3. hepatitis A
4. asthma

The node *fraction* in Figure 3.1 represents the hypothesized fraction of the total population who has outbreak disease $d_k$ and has visited the ED in the last 24 hours. Let $F$ denote this node. Let $f$ denote an arbitrary value of $F$. For example, $f$ might be $10^{-4}$ or $2 \times 10^{-5}$ or any of a wide range of fractions. As in PC, we model $F$ as a discrete variable for computational convenience.

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 the *person_i disease* node. Although we index over all people in the population, the data we use is de-identified, and thus, we do not know the actual identity of a given person $i$. We use the assignment $PD_i = noED$ to represent that person $i$ did not come to the ED during the most recent 24-hour period. For the patients who came to the ED, a specific patient among them could have a non-outbreak disease $d_0$ or a specific outbreak disease $d_k$ for $1 \leq k \leq K$. That is, $PD_i$ is a random variable that can take on values $d_0$, $d_1$, …, $d_K$.

Given the disease state of person $i$, the *person_i evidence* node is used to model the evidence state of that person, such as the patient's chief complaint. Let $E_i$ represent this node for a specific person $i$. PC currently models 53 possible patient chief complaints. For ED patients who do not have one of those 53 chief complaints, their chief complaints are assumed to have the "catch all" value *other*. For people who did not visit the ED, our convention is to give their chief complaints the value *unknown*. Therefore, $E_i$ = *unknown* and $PD_i$ = *noED* are assigned when a person *i* did not visit the ED during the previous 24-hour period.

As mentioned, PC currently models 13 possible outbreak diseases; thus $K = 13$. In practice, some other disease outbreak might occur, such as a new infectious disease that has never been seen before. Recall from Section 3.1 that we denote an unmodeled or unknown outbreak disease as $d_*$. However, PC is not able to represent such diseases. Chapter 4 describes how I extend PC to model such diseases.

### 3.2.2 The conditional probability tables

This section describes the conditional probability tables (CPTs) in the Bayesian network model in Figure 3.1.

The prior probabilities of variable $O$ are given by $P(O = OB) = 0.05$ and thus $P(O = NOB) = 0.95$. The prior probability of outbreak condition is high due to the incorporation of influenza into the CDC-A$^+$ outbreak diseases being modeled and the frequent occurrence of influenza in any year.

The conditional probabilities of $P(OD = d_k \mid O = OB)$ (for $k \geq 1$) were assessed by an infectious disease expert, Dr. John Dowling at the University of Pittsburgh, based on the literature and his clinical beliefs, where $d_k$ is one of the 13 CDC-A$^+$ diseases. When $O = NOB$, we have $OD$ = *none* with probability 1, that is, $P(OD = none \mid O = NOB) = 1$.

The values of $F$ depend on the temporal progression of disease $d_k$ and the type of disease $d_k$, because an outbreak disease in an earlier stage tends to affect a smaller fraction of population than that disease in a later stage, and some outbreak diseases tend to affect a larger fraction of population than other outbreak diseases. Since PC does not represent temporal progression in this model, there is uncertainty of the disease stage of a potential outbreak disease. Therefore, PC does not model a dependency between $F$ and $OD$. However, in general the disease model in Figure 3.1 could be readily extended to represent a dependency between $F$ and $OD$.

The values of $f$ of the fraction node $F$ are derived as follows: Let $\mu_0$ denote the mean number of patients who came to the ED each day when there is no disease outbreak in the population, and let $\sigma_0$ denote its standard deviation. Parameters $\mu_0$ and $\sigma_0$ were estimated from real ED data that presumptively contains no disease outbreak. Let $n$ be the number of outbreak cases who visited the ED when there is a disease outbreak in the population. We model the values of $n$ as being the rounded values of $\frac{1}{10}\sigma_0$, $\frac{2}{10}\sigma_0$, $\frac{3}{10}\sigma_0$, $\frac{4}{10}\sigma_0$, $\frac{5}{10}\sigma_0$, $\sigma_0$, $1.5\sigma_0$, $2\sigma_0$, $2.5\sigma_0$, $3\sigma_0$, $3.5\sigma_0$, $4\sigma_0$, $4.5\sigma_0$, and $5\sigma_0$. The 15 values of $f$ were calculated as $n / N$, where $N$ is the total number of individuals in the population who could potentially visit the EDs that are covered in the region, which in this case is estimated to be 400,000. The fraction $F$ is assumed to be uniformly distributed over these 15 discrete values. Thus, $P(F = f) = 1/15$ for all values of $f$[7].

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 a given 24-hour period and (2) has a non-outbreak ED disease. The probability that the person has $d_0$, which is denoted as $\theta$, is estimated from past ED data during which it is assumed no outbreak was occurring. Then $P(PD_i = noED \mid OD = none, F = f) = 1 - \theta$.

When $OD = d_k$ (for $1 \leq k \leq K$), a specific person $i$ could have disease $d_0$, $d_k$, or $noED$. That person cannot have another outbreak disease, because in the current model it is assumed that there is at most one outbreak disease present in the population at any time. The probability

[7] Note that since the interval spacing of the discrete values of $F$ is not the same, assuming a uniform over the discrete version of $P(F = f)$ actually assumes a non-uniform over the continuous probability density function of $P(F = f)$.

of person *i* having $d_k$ is equal to the value of the *fraction* node, *f*, by the construction of that node. Thus, there is 1 - *f* fraction of the total population who do not have $d_k$ (who have $d_0$ or *noED*). It is assumed that a fraction $\theta$ of these people have $d_0$. The probability of person *i* having $d_0$ in light of $d_k$ as an outbreak disease in the population is then equal to $(1 - f)\theta$. Finally, $P(PD_i = noED \mid OD = d_k, F = f) = 1 - f - (1 - f)\theta = (1 - f)(1 - \theta)$.

Let $p_u^j$ represent the probability of a specific person having the *j*th chief complaint given that the person has disease $d_u$, namely $P(E_j = e_j \mid PD_i = d_u)$, where $0 \le u \le 13$ and $e_j$ is one of the 53 possible observed chief complaints for person *i*. $P(E_j = e_j \mid PD_i = d_0)$ is estimated from past ED data that are assumed to contain no disease outbreaks. When person *i* has an outbreak disease, $p_u^j$ is assessed using expert knowledge for $1 \le j \le 53$ and $1 \le u \le 13$. In particular, Dr. John Dowling of the University of Pittsburgh, who is an infectious disease specialist, assessed $p_u^j$ based on the literature and his clinical beliefs (Cooper 2006).

### 3.2.3 Inference

Each hour, PC outputs the posterior probability of each CDC-A$^+$ disease in a population. In order to compute these probabilities, we need to compute $P(OD = none \mid E)$ and $P(OD = d_k \mid E)$ for each outbreak disease $d_k$, where $1 \le k \le 13$, and *E* denotes the observed chief complaint for each person in the population who came to the ED within the past 24 hours. PC performs exact inference to compute these probabilities. Since PC serves as the experimental domain for this dissertation research, and this dissertation research involves more complicated inferences than those in PC, I leave discussion of these issues to Chapter 4.

# 4.0 THE BAYESIAN HYBRID DETECTION ALGORITHM

This chapter describes the main issues of the BH model and algorithm. The BH algorithm combines models of known and unknown outbreak diseases in order to capture outbreaks due to both categories of disease. As described in Chapter 3, PC only models known diseases (13 CDC-A$^+$ diseases) but does not model an unknown disease $d_*$. The BH algorithm extends PC to (1) model an unknown outbreak disease $d_*$ and (2) model the *distributions*[8] over the probabilities of a person's symptoms, given different diseases that a person could have. These extensions make exact inference in the Bayesian network model in Figure 3.1 computationally and conceptually more complicated. The current chapter describes how I performed inference efficiently under these extensions.

The BH algorithm has two versions: a univariate version, which I call the UBH algorithm, and a multivariate version, which I call the MBH algorithm. I begin this chapter by presenting the three categories of disease models in Section 4.1, which include a disease-specific model (DSM), an unknown-disease model (UDM), and a partially-known disease model (PDM). Section 4.2 describes the UBH algorithm, which takes as input a univariate symptom state for every person in the population for the last 24 hours. Section 4.2 describes the MBH algorithm, which takes as input multivariate symptom states for every person in the population. For each

[8] PC models point probabilities, but BH models distributions over those probabilities. Thus, BH models an additional level of uncertainty, relative to PC.

version of the BH algorithm, I describe how to use DSM, UDM, and PDM to model different types of outbreak diseases in the PC domain. I also describe how to perform inference on these Bayesian network models.

## 4.1 DISEASE MODELING

Common Bayesian networks model parameters (probabilities) as being known, and PC uses this approach. PC models the probability of a symptom in an outbreak disease as a known probability value using expert assessment, as for example the probability that a patient will have a cough given that he or she has respiratory anthrax. For the purpose of estimation and assessment in this section, the probability of a symptom in a disease can be viewed as the frequency in a large sample limit of patient cases with that disease.

In contrasts, the BH algorithm models *distributions* over parameters that represent frequencies of the population, as for example the frequency of cough in the individuals with anthrax, whereas the true parameters (frequencies) are unknown. Modeling distributions of parameters allows us to represent uncertainty in how diseases express themselves and to express our own ignorance of how diseases will express themselves – both forms of uncertainty are important.

This section describes how I modeled non-outbreak diseases, known outbreak diseases, unknown outbreak diseases, and partially-known outbreak diseases. In particular, I describe these disease models in terms of the $j$th disease symptom, assuming there are a total of $J$ symptoms ($1 \leq j \leq J$) that are conditionally independent, and the symptom states of each symptom $j$ are binary, as described below. Section 4.2 describes the univariate BH algorithm, where $J = 1$. Section 4.3

describes the multivariate BH algorithm, where $J > 1$, and presents disease models in terms of a total of $J$ disease symptoms, each of which is assumed to be binary.

### 4.1.1 The non-outbreak disease model

As stated, this model represents that a person has a non-outbreak disease $d_0$. Recall from Section 3.2 that $p_0^j$ represents the probability of having symptom $j$ given a person having $d_0$. I assume $p_0^j$ is distributed according to a Beta distribution, namely, $p_0^j \sim \text{Beta}(\alpha_0^j, \beta_0^j)$. Next, I describe how BH models $p_0^j$ from past ED data.

Parameters $\alpha_0^j$ and $\beta_0^j$ were estimated based on real ED data that are assumed to contain no disease outbreaks. Specifically, these parameters were estimated based on real ED reports from a large healthcare system in Pittsburgh from January to December 2002. Patient visit data to the ED were stored in a database since 1990, including dictated and transcribed ED reports and coded ICD-9 discharge diagnoses. Chapman, et al. previously identified a random sample of ED patient cases that either contained one or more respiratory symptoms or did not (Chapman 2004; Chapman 2005; Chu 2007); these two sets were relevant to their study goals. We were able to use their two sets of cases to provide the $\alpha_0^j$ and $\beta_0^j$ parameter estimates needed for this dissertation, as described in the remainder of this section. Chapman, et al. obtained 69 patient cases with respiratory symptoms and 151 cases without, yielding a total of 220 cases. By computing the fraction of respiratory patients in the real ED reports in 2002, they estimated that the fraction of respiratory patients who visit to the ED is approximately 0.08. Thus, for a specific symptom $j$, we can derive the mean probability of symptom $j$ in disease $d_0$ as follows:

$$\begin{aligned}&P(symptom\_j = present \mid person\_i\_disease = d_0)\\&= P(symptom\_j = present, respiratory = present \mid person\_i\_disease = d_0)\\&+ P(symptom\_j = present, respiratory = absent \mid person\_i\_disease = d_0)\\&= P(symptom\_j = present \mid respiratory = present, person\_i\_disease = d_0)P(respiratory = present \mid person\_i\_disease = d_0)\\&+ P(symptom\_j = present \mid respiratory = absent, person\_i\_disease = d_0)P(respiratory = absent \mid person\_i\_disease = d_0),\end{aligned}$$

where $P(respiratory = present \mid person\_i\_disease = d_0) = 0.08$.

Consider the example of deriving $P(cough = present \mid person\_i\_disease = d_0)$. From the 69 respiratory cases there were 10 patients with cough, and from the 151 non-respiratory cases there were no cases with a cough. We derived $P(cough = present \mid person\_i\_disease = d_0)$ as follows:

$$\begin{aligned}&P(cough = present \mid person\_i\_disease = d_0)\\&= P(cough = present \mid respiratory = present, person\_i\_disease = d_0)P(respiratory = present \mid person\_i\_disease = d_0)\\&+ P(cough = present \mid respiratory = absent, person\_i\_disease = d_0)P(respiratory = absent \mid person\_i\_disease = d_0)\\&= \frac{10}{69} \times 0.08 + \frac{0}{151} \times 0.92\\&= 0.01.\end{aligned}$$

Therefore, we can approximate parameters $\alpha_0^j$ and $\beta_0^j$ for the symptom *cough* as shown below, assuming symptom *cough* has a symptom index of $j = 1$:

$$\alpha_0^1 = N_{ED} \cdot P(cough = present \mid person\_i\_disease = d_0),$$

$$\beta_0^1 = N_{ED} - \alpha_0^1,$$

where $N_{ED}$ represents the total number of expected ED patients for a year. I used one-year data of ED patients rather than multi-year data of ED patients to estimate $N_{ED}$ because multi-year data might not capture recent trends in the estimated parameters (e.g., the frequency of cough might change over time), whereas one-year data is likely to capture recent trends and still constitutes a large sample size.

The above method of estimating parameters $\alpha_0^j$ and $\beta_0^j$ is applied to every symptom $j$ of disease $d_0$ that is modeled, where $1 \leq j \leq J$, in order to construct a multivariate model for a non-outbreak disease. For a complete list of parameters, see Appendix A. In the current multivariate version of the BH algorithm (for $J > 1$), this dissertation research assumes that symptom states are conditionally independent given the disease state.

### 4.1.2 The disease-specific model (DSM)

DSM represents that a person has a specific outbreak disease $d_k$ (for $1 \leq k \leq K$). Recall that $p_k^j$ represents the probability of having symptom $j$ given a person having $d_k$. $p_k^j$ is also assumed to have a Beta distribution, namely, $p_k^j \sim \text{Beta}(\alpha_k^j, \beta_k^j)$. Next, I describe how BH models $p_k^j$ using informative priors.

Recall that $p_k^j$ may be viewed as a frequency in the large sample limit of patient cases with disease $d_k$. I assessed parameters $\alpha_k^j$ and $\beta_k^j$ based on expert judgments for $1 \leq k \leq K$. The expert provided his expectation $\mu_k^j$ of $p_k^j$ and an interval assessment $[a_k^j, b_k^j]$ for which he stated a belief that there is a 90% chance that $p_k^j$ is between $a_k^j$ and $b_k^j$. Parameters $\alpha_k^j$ and $\beta_k^j$ were then estimated by solving Equations 4.1 and 4.2 in terms of the Beta density function $\text{Beta}(p_k^j; \alpha_k^j, \beta_k^j)$.

$$\int_{a_k^j}^{b_k^j} Beta\left(p_k^j; \alpha_k^j, \beta_k^j\right) dp_k^j = 0.9. \tag{4.1}$$

$$\mu_k^j = \frac{\alpha_k^j}{\alpha_k^j + \beta_k^j}. \tag{4.2}$$

The above method of estimating parameters $\alpha_k^j$ and $\beta_k^j$ is applied to every symptom $j$ of disease $d_k$, where $1 \leq j \leq J$, in order to construct a multivariate disease-specific model for an outbreak disease $d_k$. For a complete list of parameters, see Appendix A. The multivariate symptom states are assumed to be conditionally independent given the disease state.

### 4.1.3 The unknown-disease model (UDM)

UDM represents that a person has an unknown outbreak disease $d_*$. I model $p_*^j$, which is the probability of symptom $j$ of an unknown disease, using a non-informative prior, where $1 \leq j \leq J$. 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 (Castillo 2007). An example of this was proposed by Bayes himself (Press 2003), who used a uniform [0,1] on the Binomial proportion parameter $p$. Tuyl et al. also suggest using the uniform prior Beta(1, 1), called the Bayes-Laplace prior, on the Binomial proportion parameter $p$ to represent ignorance (Tuyl 2009). I model $p_*^j$ using a uniform distribution over [0,1], namely, $p_*^j$ ~ Beta(1, 1).

When multiple disease symptoms are being modeled, I define an unknown disease $d_*$ as a disease for which we almost completely lack knowledge regarding every disease symptom. Each symptom $j$ is modeled using a non-informative prior described above. All the multiple disease symptoms are modeled as being conditionally independent given the person's disease state.

### 4.1.4 The partially-known disease model (PDM)

PDM represents a partially-known outbreak disease that manifests some disease symptoms similar to one or more modeled known diseases. For example, a partially-known outbreak disease might be modeled as having a cough rate in the population that is similar to the cough rate of a known outbreak disease, such as influenza. Recall from Chapter 1 that this dissertation focuses on a basic way of modeling unknown and partially-known diseases. In particular, I will model a partially-known disease that results in a rate of some disease symptoms in the population that is similar to those of several known diseases that have been modeled. I thus will use a mixture of priors of the known diseases to model a partially-known disease and will represent this partially-known disease with the notation $d_{*p}$.

As described in Section 2.4.5, a mixture of priors is composed of prior distributions that come from several sources of information. A mixture of Beta prior distributions has a closed-form solution for a Beta-Binomial model, which results in a posterior distribution being a mixture of Beta distributions. As the definition of a partially-known disease described above suggests, I intend to use a mixture of priors to model a partially-known disease $d_{*p}$. In particular, a mixture of priors consists of priors that are informative for known (modeled) diseases and non-informative for conditions we know little about, as explained next.

I now present how to model a partially-known disease $d_{*p}$ that has a disease symptom $j$ whose incidence rate is similar in distribution to those of several known diseases. Assuming there are $M$ known diseases being modeled, the method of mixture priors involves a finite mixture model of $M + 1$ components, with each of the $M$ components representing a prior distribution for a known disease and an additional component representing a non-informative prior for conditions that we know little about. Equation 4.3 shows the derived prior distribution $f\left(p_*^j;\boldsymbol{\theta}_*^j\right)$ of the $j$th disease symptom for a partially-known disease $d_{*p}$.

$$f\left(p_*^j;\boldsymbol{\theta}_*^j\right)=\sum_{k=1}^{M+1}\tau_k^j f_k\left(p_k^j;\boldsymbol{\theta}_k^j\right), \tag{4.3}$$

where $0<\tau_k^j<1$ and $\sum_{k=1}^{M+1}\tau_k^j=1$.

In this Equation, when $1\leq k\leq M$, $f_k\left(p_k^j;\boldsymbol{\theta}_k^j\right)$ represents an informative prior distribution for the $j$th symptom of disease $d_k$, and $f_k\left(p_k^j;\boldsymbol{\theta}_k^j\right)$ is assumed to have a Beta distribution as Beta($p_k^j$; $\alpha_k^j$, $\beta_k^j$), where parameters $\alpha_k^j$ and $\beta_k^j$ are estimated using the method described in the disease-specific model in Section 4.1.2 above; when $k = M + 1$, $f_{M+1}\left(p_{M+1}^j;\boldsymbol{\theta}_{M+1}^j\right)$ represents a non-informative prior distribution, for which I use a uniform distribution on [0, 1]. $\tau_k^j$ is the probability that measures the degree to which $d_{*p}$ is similar in distribution to $d_k$ with respect to the *j*th symptom.

In this dissertation research, I assume that we know little about possible similarities in distribution between disease $d_{*p}$ and any disease of the $M + 1$ components. Thus, I assume $\tau_k^j$ is uniformly distributed over the $M + 1$ components for $1 \leq k \leq M + 1$, that is, $\tau_k^j = 1 / (M + 1)$. If, in other applications, we have prior belief about how to weight the disease components (e.g., any partially known disease is expected to be much more like influenza than like any other modeled disease), we could weight the components differentially (e.g., more heavily for influenza).

I now describe an example of modeling a partially-known disease $d_{*p}$ using the method of mixture priors. Suppose the observed evidence contains the status of the symptoms *cough* and *fever* for every person who came to the ED during the most recent 24 hours. Assume there are three known diseases we are modeling, namely, influenza, anthrax, and asthma, and their disease indices are $k = 1$, $k = 2$, and $k = 3$, respectively. The only prior knowledge we wish to model about $d_{*p}$ is that it is a respiratory-like illness that manifests a cough rate in the population that is similar in distribution to those of several diseases, such as influenza, anthrax, and asthma. Moreover, we believe it possible (although not inevitable) that disease $d_{*p}$ could have a rate of cough that is quite different from all of influenza, anthrax, and asthma. Hence, a fourth component is incorporated in the mixture of priors to represent this condition. Assuming *cough* and *fever* have the symptom indices, $j = 1$ and $j = 2$, respectively, we can derive a prior distribution of the probability of cough for $d_{*p}$ as follows:

$$f\left(p_*^1;\boldsymbol{\theta}_*^1\right)=\frac{1}{4}\sum_{k=1}^{3} f_k\left(p_k^1;\boldsymbol{\theta}_k^1\right)+\frac{1}{4}\cdot f_4\left(p_4^1;\boldsymbol{\theta}_4^1\right),$$

where $f_k\left(p_k^j;\boldsymbol{\theta}_k^j\right)$ represents an informative prior derived for the *k*th disease and $f_4\left(p_4^1;\boldsymbol{\theta}_4^1\right)$ represents a non-informative prior, namely Beta(1, 1).

If we have no reason to believe that the fever rate of disease $d_{*p}$ is similar to known disease we are modeling, then we could model the prior distribution of the probability of fever for $d_{*p}$ using a uniform distribution on [0, 1], as described in Section 4.1.3.

## 4.2 THE UNIVARIATE BAYESIAN HYBRID DETECTION ALGORITHM

The univariate Bayesian hybrid detection algorithm (UBH), which I have developed and investigated, uses the PC model, and combines specific and non-specific detectors. It takes as input the total count over the most recent 24 hours of a single emergency department patient symptom, such as cough (Shen 2007). This section describes an example of this algorithm in terms of aggregate counts of the binary symptoms, *cough* vs. *no cough*.

### 4.2.1 The univariate entity-based disease model

Recall from Section 3.2.1, $E_i$ represents the *person_i evidence* node in Figure 3.1. In the UBH algorithm, I represent $E_i$ as a binary symptom (*cough* vs. *no cough*) of person *i*, where $1 \leq i \leq N$.[9] 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 give their evidence state the

[9] See Section 4.3 for a multivariate version of the Bayesian hybrid algorithm that models multiple symptoms of person *i*.

value *unknown*. Figure 4.1 shows the Bayesian network model used for the example UBH algorithm.

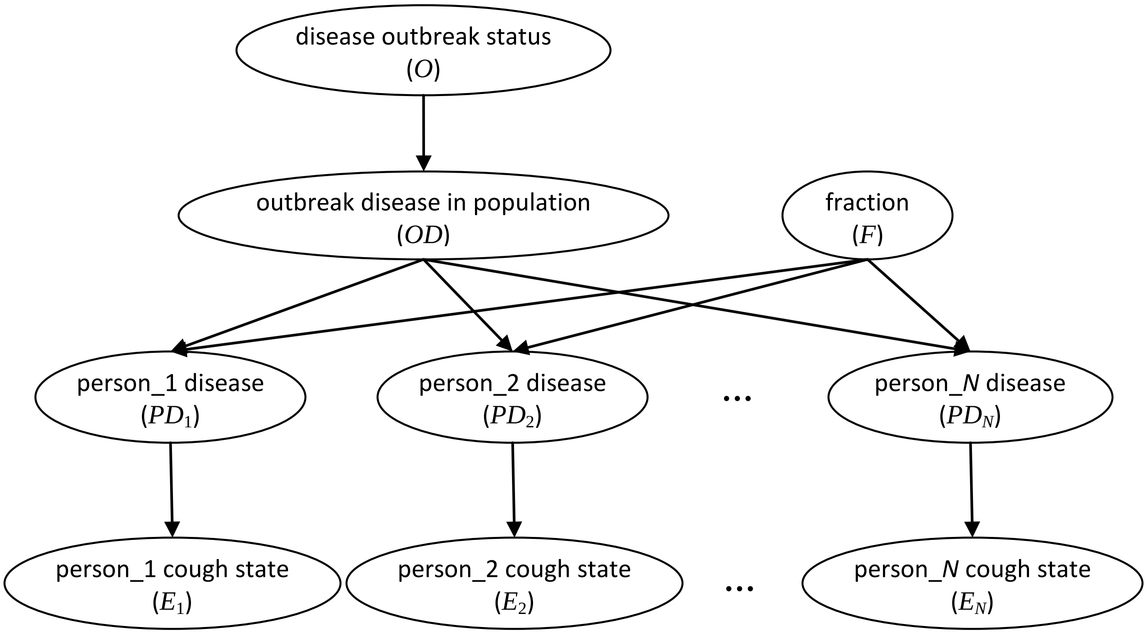


**Figure 4.1** A Bayesian network showing the UBH entity-based disease model.

### 4.2.2 The conditional probability tables

As described in Section 3.2.1, the *disease outbreak status* node represents the outbreak status in the population during the most recent 24-hour period. The node is represented as *O*, where *O* = *OB* or *NOB*. The prior probability of variable *O* is specified as in PC[10], that is, $P(O = OB) = 0.05$, and thus $P(O = NOB) = 0.95$.

[10] The current version of UBH actually derives the likelihood ratio $LR = P(E = e \mid O = OB) / P(E = e \mid O = NOB)$, instead of the posterior probability $P(O = OB \mid E = e)$, as described in Section 4.2.3, in order to remove the need to specify this difficult prior probability. In any event, the evaluation measures that we use are not sensitive to the prior probability.

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 $d_u$ (for $u \in \{1, \cdots, K, *\}$) is occurring in the population. I use expert assessments to determine the probability of $P(OD = d_u \mid O = OB)$, as described in Section 3.2.2. However, the BH algorithm models an unknown outbreak disease $d_*$, and the probability $P(OD = d_* \mid O = OB)$ is difficult to estimate using the literature or expert assessments. Thus, I will perform a sensitivity analysis over various values for $P(OD = d_u \mid O = OB$. See Chapter 5 for a detailed description of specifying these probabilities.

In this chapter I do not describe in detail every node in Figure 4.1, but rather focus on deriving the conditional probability tables of the Bayesian network model in that figure, which is different from the PC model. For a detailed description of the nodes in Figure 4.1, see Section 3.2.1.

If $OD = none$ or $OD = d_k$ (for $1 \le k \le K$), then we can derive $P(PD_i \mid OD = none, F = f)$ or $P(PD_i \mid OD = d_k, F = f)$ in the same fashion as in PC:

$P(PD_i = d_0 \mid OD = none, F = f) = \theta$,

$P(PD_i = noED \mid OD = none, F = f) = 1 - \theta$,

$P(PD_i = d_k \mid OD = d_k, F = f) = f$,

$P(PD_i = d_0 \mid OD = d_k, F = f) = (1 - f)\theta$, and

$P(PD_i = noED \mid OD = d_k, F = f) = (1 - f)(1 - \theta)$.

When $OD = d_*$, we can similarly derive

$P(PD_i = d_* \mid OD = d_*, F = f) = f$,

$P(PD_i = d_0 \mid OD = d_*, F = f) = (1 - f)\theta$, and

$P(PD_i = noED \mid OD = d_*, F = f) = (1 - f)(1 - \theta)$.

Recall that we model the state of a binary symptom $E_i$ for every person in the population using a Bernoulli distribution. Because the univariate version of the BH algorithm only models a single disease symptom, for simplicity, I ignore the superscript that defines the index of a disease symptom that appears in Section 3.2.2. Recall that for the people who came to the ED in the last 24 hours, their evidence states are *cough* or *no cough*, and for people who did not visit the ED, our convention is to give their evidence state the value *unknown*. I define $P(E_i = cough \mid PD_i = d_0) = p_0$, $P(E_i = cough \mid PD_i = d_k) = p_k$ and $P(E_i = cough \mid PD_i = d_*) = p_*$. Table 4.1 describes the conditional probability assignments for $P(E_i \mid PD_i)$. Recall from Section 3.2.2 that PC uses expert knowledge to obtain $P(E_i \mid PD_i)$ in this conditional probability table. The UBH algorithm takes into account the uncertainty of the expert's estimates and models these probabilities as random variables. See Section 4.1 for a detailed description of how to model $p_0$, $p_k$, and $p_*$ in the non-outbreak disease model, in the disease-specific model, and in the unknown-disease model, respectively.

**Table 4.1** The conditional probability table for $P(E_i \mid PD_i)$

| *cough state* ($E_i$) | *ED & $d_0$* | *ED & $d_k$* | *ED & $d_*$* | *noED* |
|---|---|---|---|---|
| *ED & cough* | $p_0$ | $p_k$ | $p_*$ | 0 |
| *ED & no cough* | $1 - p_0$ | $1 - p_k$ | $1 - p_*$ | 0 |
| *unknown* | 0 | 0 | 0 | 1 |

### 4.2.3 Inference

This section describes how to perform inference on the example Bayesian network model in Figure 4.1. The objective of inference is to derive the posterior probability of an outbreak occurring given the observed evidence. That is, we wish to derive $P(O = OB \mid E = e)$, where $e$

denotes the status of the symptom *cough* for every person in the population. I assume the status is either *cough* or *no cough* for people who have come to the ED, and that it was *unknown* for people who have not. I derive $P(O = OB \mid E = e)$ by deriving $P(E = e \mid O = OB)$ and $P(E = e \mid O = NOB)$, assessing $P(O = OB)$, and applying Bayes' rule.

Note that $P(O = OB \mid E = e)$ represents the posterior probability of an outbreak due to any outbreak disease. Recall that the BH detection algorithm models 13 CDC-$\mathrm{A}^+$ diseases and unknown diseases. If we want to know the posterior probability of a particular disease outbreak occurring or an unknown disease outbreak occurring, we then derive $P(OD \mid E = e)$, where $OD = d_k$ (for $1 \leq k \leq K$) represents a specific outbreak disease, and $OD = d_*$ represents an unknown outbreak disease. In this dissertation, the BH algorithm aims to detect disease outbreaks due to any outbreak disease, i.e., deriving $P(O = OB \mid E = e)$, but does not characterize the outbreak to determine whether it is due to $d_k$ or $d_*$.

Computing the posterior probability $P(O = OB \mid E = e)$ involves computing the likelihoods $P(E = e \mid O = OB)$ and $P(E = e \mid O = NOB)$, and specifying the prior probability $P(O = OB)$, as shown below:

$$P(O = OB \mid E = e) = \frac{P(E = e \mid O = OB)P(OB)}{P(E = e \mid O = OB)P(OB) + P(E = e \mid O = NOB)P(NOB)}. \tag{4.4}$$

As described in Section 4.2.2, I derive the likelihood ratio $LR$ as Equation 4.5 to avoid specifying the prior probability $P(O = OB)$, which can be difficult to assess with confidence.

$$LR = \frac{P(E = e \mid O = OB)}{P(E = e \mid O = NOB)}. \tag{4.5}$$

We can derive $LR$ from Equation 4.4 and 4.5, as follows:

$$LR = \frac{P(O = NOB)}{(1/P(O = OB \mid E = e) - 1)P(O = OB)}.$$

Thus, given the prior probabilities $P(O = OB)$ and $P(O = NOB)$, $LR$ is an increasing function of the posterior probability $P(O = OB \mid E = e)$.

The proposed evaluation method described in this dissertation (Chapter 5) determines the expected detection time (at a specific false alert rate) based on the relative order of the probabilistic outputs. Since $LR$ is an increasing function of the posterior probability $P(O = OB \mid E = e)$, computing $LR$ rather than $P(O = OB \mid E = e)$ does not affect the detection performance of the BH algorithm when using this evaluation method. Moreover, the values of the prior probabilities $P(O = OB)$ and $P(O = NOB)$ do not affect this relative order. Thus these prior probabilities do not affect the detection performance of the BH algorithm reported in Chapter 5, but do affect the magnitude of the outputs $LR$ of the BH algorithm.

In order to derive the likelihood ratio $LR$, I first compute the likelihood $P(E = e \mid O = OB)$ as follows:

$$\begin{aligned} P(E = e \mid O = OB) &= \sum_{OD \neq d_0} P(E = e, OD \mid O = OB) \\ &= \sum_{OD \neq d_0} P(E = e \mid OD) P(OD \mid O = OB). \end{aligned} \tag{4.6}$$

The second equality of the above equation holds because $E$ and $OB$ are conditionally independent given $OD$ being an outbreak disease, given the Bayesian network model in Figure 4.1.

I then compute the likelihood $P(E = e \mid O = NOB)$ using the equation below, because $O = NOB$, $OD = none$, and $OD = d_0$ are equivalent events.

$$P(E = e \mid O = NOB) = P(E = e \mid OD = d_0). \tag{4.7}$$

Now the inference problem turns out to be computing the likelihood $P(E = e \mid OD = d_k)$ (for $1 \leq k \leq K$) and $P(E = e \mid OD = d_*)$ in Equation 4.6 and computing the likelihood $P(E = e \mid$

*OD* = $d_0$) in Equation 4.7. For the remainder of this section, let $d_w$ represent any member of {$d_0$, $d_1$, …, $d_K$, $d_*$}, because they are all involved in the same inference procedure.

We can derive *P*(*E* = *e* | *OD* = $d_w$) by performing inference on the Bayesian network in Figure 4.1. Inference is complicated by the fact that we have distributions over *P*($E_i$ = $e_i$ | $PD_i$), as described in Sections 4.1; thus, inference includes integrating over these distributions. I assume that the values of the *fraction* node in Figure 4.1 are distributed over some discrete set of values that are independent of *OD*. Thus we are able to derive *P*(*E* = *e* | *OD* = $d_w$) as follows:

$$P(E = e \mid OD = d_w) = \sum_f P(E = e \mid OD = d_w, F = f)P(F = f). \tag{4.8}$$

By assuming that the *N* cases in the population are independent, we are able to derive *P*(*E* = *e* | *OD* = $d_w$, *F* = *f*) as follows:

$$P(E = e \mid OD = d_w, F = f) = \int_{B_P} \prod_{i=1}^{N} P(E_i = e_i \mid OD = d_w, F = f, B_P)g(B_P)dB_P, \tag{4.9}$$

where $B_P$ is a vector whose values denote the conditional probability assignments associated with *P*($E_i$ = $e_i$ | $PD_i$), and *g*($B_P$) is the probability distribution over $B_P$. Note that $B_P$ is $p_0$ when *OD* = $d_0$, $B_P$ = $p_0$ and $p_k$ when *OD* = $d_k$ (for $1 \le k \le K$), and $B_P$ = $p_0$ and $p_*$ when *OD* = $d_*$.

Since $PD_i$ is hidden, in computing Equation 4.9 we had to sum out all the combinations of $PD_i$, as follows:

$$P(E = e \mid OD = d_w, F = f) = \int_{B_P} \prod_{i=1}^{N} \sum_{PD_i} P(E_i = e_i, PD_i \mid OD = d_w, F = f, B_P)g(B_P)dB_P. \tag{4.10}$$

From the network structure in Figure 4.1, the *person_i cough state* node ($E_i$) and the *outbreak disease in population* node (*OD*) are conditionally independent given the *person_i disease* node ($PD_i$); a similar independence holds between the *person_i cough state* node and the *fraction* node. We can write Equation 4.10 as follows:

$$
\begin{aligned}
&P(E = e \mid OD = d_w, F = f) \\
&= \int_{B_P} \prod_{i=1}^{N} \sum_{PD_i} P(E_i = e_i \mid PD_i, B_P) P(PD_i \mid OD = d_w, F = f) g(B_P) dB_P .
\end{aligned}
\tag{4.11}
$$

We then rearrange the sum in Equation 4.11 to obtain Equation 4.12. In this equation, $P(PD_i \mid OD = d_w, F = f)$ is calculated as described Section 4.2.1. $P(E = e| PD_i, B_P)$ is the probability associated with Table 4.1.

$$
\begin{aligned}
&P(E = e \mid OD = d_w, F = f) \\
&= \sum_{PD_1 \cdots PD_N} \int_{B_P} \prod_{i=1}^{N} P(E_i = e_i \mid PD_i, B_P) P(PD_i \mid OD = d_w, F = f) g(B_P) dB_P .
\end{aligned}
\tag{4.12}
$$

Recall that when $OD \neq none$, $PD_i$ is hidden for the *i*th person who comes to the ED. Let $N_{ED}$ be the number of people who came to the ED. The sums in Equation 4.12 were taken over every possible value of the hidden variable $PD_i$ for every possible person *i* who came to the ED, which resulted in a complexity exponential in $N_{ED}$.

In order to compute the likelihood efficiently, I adapted the inference method given in (Cooper 1995), which performs Bayesian inference using grouping and factorization. I thus call this method the Bayesian grouping and factorization method or the BGF method. In this method, the population is aggregated into groups by their *cough* status. The example application described in Section 4.2.1 contained three groups, namely *ED* & *cough*, *ED* & *no cough*, and *unknown*. I call these three groups $G_c$, $G_n$, and $G_u$ respectively. Note that for the people in $G_{u,}$ their disease status is not hidden because their disease status is assigned to be *noED* as described in Section 3.2.1. Let $u_{ED}$ be the number of the groups excluding $G_u$, namely $u_{ED} = 2$. To compute $P(E \mid OD, f)$, the inference method used a factored approach to efficiently sum over every possible value of the hidden variable $PD_i$ for every possible person *i* in $G_c$ and in $G_n$, with a time complexity that is $O\left(N_{ED}^{\;u_{ED}}\right) = O\left(N_{ED}^{\;2}\right)$ (Cooper 1992; Cooper 1995).

## 4.3 THE MULTIVARIATE BAYESIAN HYBRID DETECTION ALGORITHM

If an outbreak due to disease *d* occurs in the population, patients infected with disease *d* are often expected to exhibit several disease symptoms of disease *d*. Although many evidential features may be predictive of an outbreak, detection algorithms generally monitor each evidential feature separately (as described in Section 4.2), which limits the surveillance system's detection capabilities. This section describes a multivariate Bayesian hybrid (MBH) detection algorithm that takes as input multiple disease symptoms, such as *cough*, *fever,* and *headache*, of every person in the population for the last 24 hours.

### 4.3.1 The multivariate entity-based disease model

The MBH algorithm extends the UBH algorithm to model multiple clinical findings for every patient in the ED. However, extracting specific clinical findings from electronic ED patient reports continues to be a major challenge (McDonald 1997), although steady progress is being made (Chapman 2004; Chapman 2005; Chu 2007). This dissertation assumes that we will obtain a set of clinical findings for each patient in the ED in the foreseeable future. Thus, the multivariate disease model uses such evidence rather than assuming we only will have a single patient chief complaint, which is information that is already readily available.

When *J* evidential features of every person are available, the disease model of the MBH algorithm models them as being conditionally independent, as described below. This disease model is different from the PC model in that PC models a patient chief complaint (one per patient) while the MBH algorithm models multiple disease symptoms per patient.

Recall from Section 4.2.1 that $E_i$ is a variable that represents an evidential feature for person $i$. When $E_i$ contains $J$ symptoms, $E_i^1$, …, $E_i^J$, I apply the naïve Bayes model (Mitchell 1997), as shown in Figure 4.2, where every conditional symptom state of person $i$ ($E_i^j$, for $1 \leq j \leq J$) is modeled using a Bernoulli distribution. Thus, we obtain the Bayesian network model shown in Figure 4.3 for the MBH algorithm, where every subnetwork that models the symptoms of an individual assumes conditional independence. As described in Section 2.6.4, the conditional independence model in Figure 4.2 can be represented using plate notation, as shown in Figure 4.4. Thus, we obtain a multivariate disease model for the MBH algorithm using plate notation, as shown in Figure 4.5.

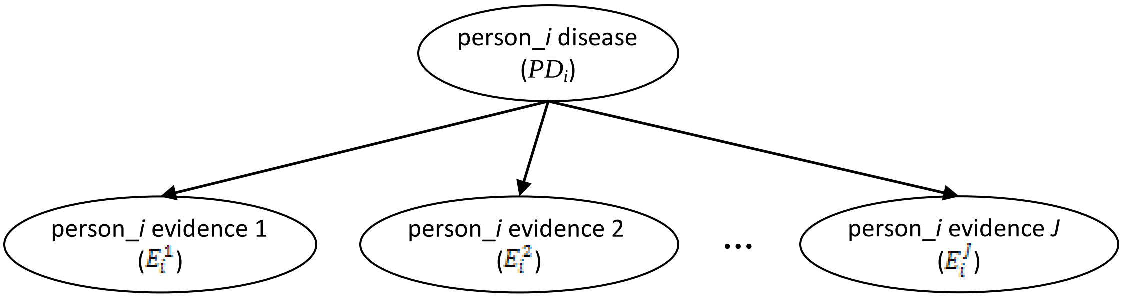


**Figure 4.2** A naïve Bayes model representing *J* evidential features for a specific person *i* in the population.

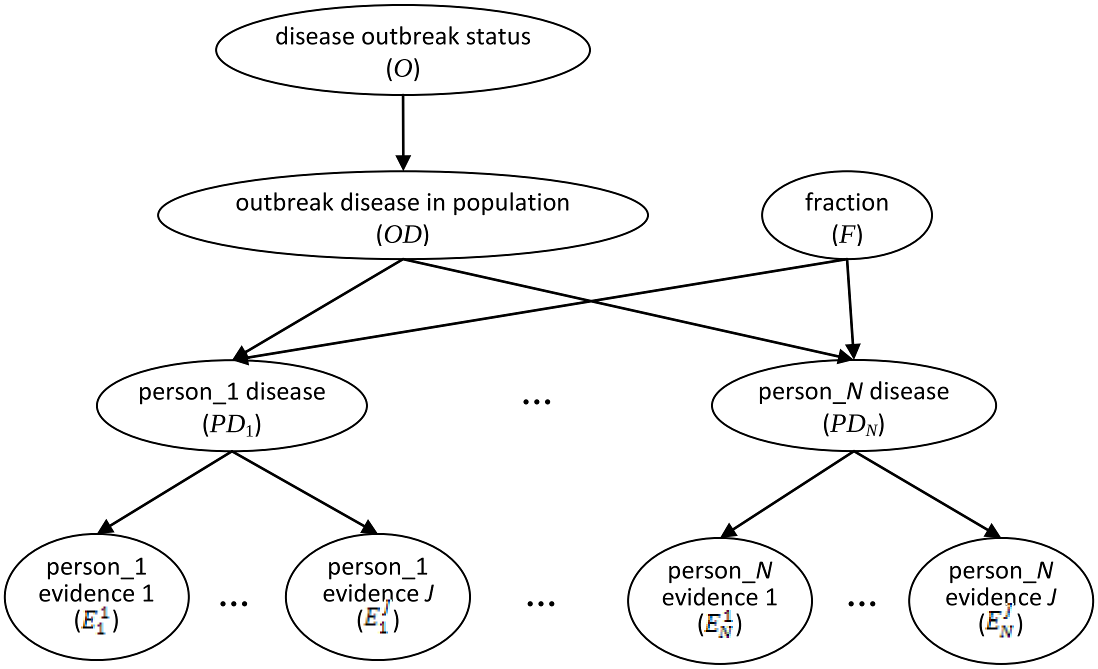


**Figure 4.3** Bayesian network showing the MBH disease model where every person's disease state and evidence states are modeled using a naïve Bayes model.

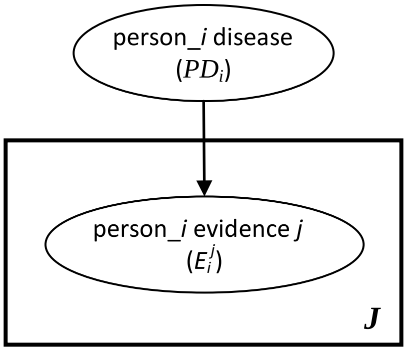


**Figure 4.4** Plate notation of the naïve Bayes model shown in Figure 4.2, where *J* in the corner represents the total *J* evidential features modeled for person *i*.

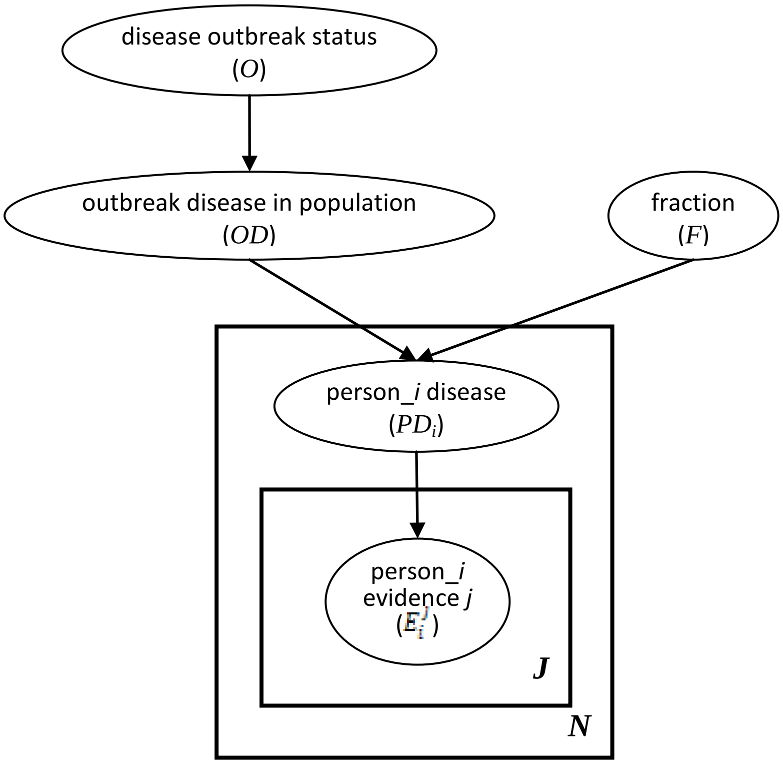


**Figure 4.5** Plate notation of the Bayesian network model shown in Figure 4.3, where *J* on the inner plate represents the total *J* evidential features modeled for person *i*, and *N* on the outer plate represents the total number of population being monitored in the region.

### 4.3.2 The conditional probability tables

In Figure 4.3, nodes *O*, *OD*, *F*, and $PD_1$, … $PD_N$ have the same conditional probability tables as those in Figure 4.1. This section will focus on deriving the conditional probability tables for the multivariate evidential features $E_i^1$, …, $E_i^J$, for a specific person *i* in the population.

As shown in Figure 4.3, a naïve Bayes model assumes that $E_i^1$, …, $E_i^J$ are conditionally independent given the disease state of that person ($PD_i$). Recall that every conditional symptom state of person *i* ($E_i^j$, for $1 \le j \le J$) is modeled using an independent Bernoulli distribution. For a person who came to the ED in the last 24 hours, his (or her) evidence state $E_i^j$ is modeled as having symptom *j* ($e_i^j$) or not having symptom *j* ($\sim e_i^j$). Table 4.2 describes the conditional probability assignments for $P(E_i^j \mid PD_i)$, where $p_0^j$, $p_k^j$, and $p_*^j$ are modeled using the methods described in Section 4.1.

**Table 4.2** The conditional probability table for $P(E_i^j \mid PD_i)$

| *evidence state* | *ED & $d_0$* | *ED & $d_k$* | *ED & $d_*$* | *noED* |
|---|---|---|---|---|
| *ED & $e_i^j$* | $p_0^j$ | $p_k^j$ | $p_*^j$ | 0 |
| *ED & $\sim e_i^j$* | $1 - p_0^j$ | $1 - p_k^j$ | $1 - p_*^j$ | 0 |
| *unknown* | 0 | 0 | 0 | 1 |

### 4.3.3 Inference

As described in Section 4.2.3, we wish to derive the likelihood ratio $LR = P(E = e \mid O = OB) \,/\, P(E = e \mid O = NOB)$. Computing this likelihood ratio involves deriving $P(E = e \mid OD = d_w)$, where $d_w$ is a disease that represents any member of $\{d_0, d_1, \ldots, d_K, d_*\}$. We would like to derive $P(E = e \mid OD = d_w)$ by performing exact inference on the Bayesian network in Figure 4.3.

Recall from Section 4.2.3 that computing $P(E = e \mid OD = d_w)$ involves the following inference procedure.

$$P(E = e \mid OD = d_w) = \sum_f P(E = e \mid OD = d_w, F = f)P(F = f), \qquad (4.13)$$

and $P(E = e \mid OD = d_w, F = f)$ can be derived as follows, as described in Section 4.2.3:

$$\begin{aligned} P(E=e \mid OD=d_w, F=f) &= \int_{M_P} \prod_{i=1}^{N} P(E_i=e_i \mid OD=d_w, F=f, M_P) dM_P \\ &= \int_{M_P} \prod_{i=1}^{N} \sum_{PD_i} P(E_i=e_i \mid PD_i, M_P) P(PD_i \mid OD=d_w, F=f) h(M_P) dM_P \\ &= \sum_{PD_1 \cdots PD_N} \int_{M_P} \prod_{i=1}^{N} P(E_i=e_i \mid PD_i, M_P) P(PD_i \mid OD=d_w, F=f) h(M_P) dM_P, \end{aligned} \tag{4.14}$$

where $M_P$ is a matrix that contains probabilities that are described below, and we assume that our belief about the distributions of these probability parameters are independent, namely, parameter independence.

In the multivariate version, $E_i$ contains multivariate evidential features $E_i^1, \ldots, E_i^J$. Recall that we had distributions over $P(E_i^j = e_i^j \mid PD_i = d_u)$, as described in Sections 4.3.2. Let $p_u^j = P(E_i^j = e_i^j \mid PD_i = d_u)$. Let $M_P$ represent a matrix that contains every $p_u^j$, where $j$ is the row index and $1 \leq j \leq J$, and $u$ is the column index and $u \in \{0,1,\cdots,K,*\}$. Let $B_P^j$ be a vector that represents the $j$th row in $M_P$, i.e., $B_P^j = (p_0^j, p_1^j, \ldots, p_K^j, p_*^j)$. According to the conditional independence assumption of a naïve Bayes model and the assumption of parameter independence described above, $P(E_i \mid PD_i, M_P)$ and $h(M_P)$ can thus be represented as follows:

$$P(E_i = e_i \mid PD_i, M_P) = \prod_{j=1}^{J} P\left(E_i^j = e_i^j \mid PD_i, B_P^j\right). \tag{4.15}$$

$$h(M_P) = \prod_{j=1}^{J} g\left(B_P^j\right). \tag{4.16}$$

By combining Equation 4.14, 4.15, and 4.16, we derive $P(E = e \mid OD = d_w, F = f)$ as follows:

$$\begin{aligned} & P(E=e \mid OD=d_w, F=f) \\ &= \sum_{PD_1 \cdots PD_N} \int_{B_P^1 \cdots B_P^J} \prod_{i=1}^{N} \left[ \prod_{j=1}^{J} P\left(E_i^j = e_i^j \mid PD_i, B_P^j\right) g\left(B_P^j\right) \right] P(PD_i \mid OD=d_w, F=f) dB_P^1 \cdots dB_P^J . \end{aligned} \tag{4.17}$$

By combining Equation 4.13 and Equation 4.17, we can derive $P(E = e \mid OD = d_w)$ and finally compute the likelihood ratio as described above.

Recall from Section 4.2.2 that when $OD \neq none$ the disease state $PD_i$ is hidden for the *i*th person who came to the ED. Let $N_{ED}$ be the number of people who came to the ED in the most recent 24 hours. If we solve Equation 4.17 using a brute-force method, we have to sum over every possible combination of values of hidden variables $PD_i$ for every possible person $i$ who came to the ED. The time complexity is thus exponential in $N_{ED}$, as described in Section 4.2.3.

Another option is to apply the relatively more efficient exact inference method, the BGF method, described in Section 4.2.3 (that uses grouping and factorization) for the univariate BH algorithm (Cooper 1995). However, when multiple disease symptoms are modeled, this inference method often turns out to be computationally intractable as well. Recall that each person is modeled having $J$ symptoms in this multivariate version, and each symptom state of person $i$ is modeled as being *true* or *false* using a Bernoulli distribution. For example, a person who came to the ED could have symptom states as being *cough* = *true*, *fever* = *true*, and *diarrhea* = *false*. If we use the BGF method, in the worst case, there will be a total number of $u_{ED} = 2^J$ groups. With this method the time required to compute $P(E \mid OD, f)$ is $O\left({N_{ED}}^{u_{ED}}\right)$, which is not computationally feasible when $J$ and therefore $u_{ED}$ is large. In this dissertation, I applied stochastic methods to approximate Equation 4.18, which is equivalent to Equation 4.17.

$$
\begin{aligned}
&P\left(E = e \mid OD = d_w, F = f\right) \\
&= \int\limits_{B_P^1 \cdots B_P^J} \prod_{i=1}^{N} \sum_{PD_i} \left\{ \left[ \prod_{j=1}^{J} P\left(E_i^j = e_i^j \mid PD_i, B_P^j\right) g\left(B_P^j\right) \right] P\left(PD_i \mid OD = d_w, F = f\right) \right\} dB_P^1 \cdots dB_P^J .
\end{aligned}
\tag{4.18}
$$

Recall that for $PD_i \neq noED$ we had distributions over $P(E_i^j = e_i^j \mid PD_i)$ represented as $g(B_P^j)$ in Equation 4.18. Since current techniques allow directly sampling from $g(B_P^j)$, which is a

Beta distribution, I investigated using Monte Carlo integration (Wasserman 2004) to approximate Equation 4.18. Monte Carlo integration is a method of approximating an expectation by the sample mean of a function of sampled random variables. Using sampling methods makes it easy to generalize the patient model (shown in Figure 4.2) beyond the naïve Bayes model.

For each symptom *j* and each disease that person *i* could have, I sample *M* times from the Beta distribution of $g(B_P^j)$ to get a total number of *M* sampled values. For each sample, I used the sampled value as the value of $P(E_i^j = e_i^j \mid PD_i)$. Given a value of $P(E_i^j = e_i^j \mid PD_i)$, as for example $P(E_i^j = e_i^j \mid PD_i) = p_{j,d}^{(m)}$, we can use exact inference to efficiently compute $P(E = e \mid OD = d_w, F = f, p_{j,d}^{(m)})$ from the Bayesian network in Figure 4.3 using Equation 4.19, where $p_{j,d}^{(m)}$ is a sampled value for the probability of a patient with symptom *j* given that the patient has disease *d*.

$$\begin{aligned} P\left(E = e \mid OD = d_w, F = f, p_{j,d}^{(m)}\right) &= \prod_{i=1}^{N} \sum_{PD_i} \left[ \prod_{j=1}^{J} P\left(E_i^j = e_i^j \mid PD_i\right) \right] P\left(PD_i \mid OD = d_w, F = f\right) \\ &= \prod_{i=1}^{N} \sum_{PD_i} \left[ \prod_{j=1}^{J} p_{j,d}^{(m)} \right] P\left(PD_i \mid OD = d_w, F = f\right). \end{aligned} \tag{4.19}$$

In Equation 4.19, when $OD = d_0$, $PD_i = d_0$; when $OD = d_k$ (or $d_*$), $PD_i$ can take two values as $PD_i = d_0$ or $d_k$ (or $PD_i = d_0$ or $d_*$). Thus, the summation in this equation takes $O(J)$ time, where *J* is the total number of symptoms. Thus, computing Equation 4.19 requires $O(J \cdot N)$ time, where *N* is the total number of people in the population being monitored; thus, computing it is efficient.

Using Monte Carlo integration over the statistic given by Equation 4.19, we can compute $P(E = e \mid OD = d_w, F = f)$, as shown in Equation 4.20, which requires a time complexity of $O(J \cdot N \cdot M)$.

$$P\left(E = e \mid OD = d_w, F = f\right) \approx \frac{1}{M} \sum_{m=1}^{M} P\left(E = e \mid OD = d_w, F = f, p_{j,d}^{(m)}\right) \tag{4.20}$$

However, Monte Carlo integration might fail to converge, thus not giving an accurate approximation. Thus, I also investigated using importance sampling (Wasserman 2004) to approximate the integral in Equation 4.18.

Importance sampling chooses a proposal distribution *q* from which to simulate a set of random variables, as opposed to directly sampling from the target distribution. A main reason for using importance sampling is the potential to reduce the variance of the approximation by an appropriate choice of the proposal distribution *q*, as samples from *q* may be more "important" for the estimation of the integral in the sense of having greater probability mass. Wasserman (Wasserman 2004) suggests using a proposal distribution that has a larger tail relative to the target distribution. I used a uniform distribution over [0, 1] as a proposal distribution and compared the simulation results with those using the basic Monte Carlo integration. I found that Monte Carlo integration converged well for solving the problem in this dissertation, and I think it is mainly because this problem contains relatively thicker-tailed target distributions (Beta distributions).

# 5.0 EXPERIMENTAL EVALUATION

This chapter describes an experimental evaluation of the Bayesian hybrid detection algorithm. Section 5.1 describes the outbreak diseases and symptoms that were used for creating semi-synthetic datasets. Section 5.2 describes how the experimental data were simulated using the selected diseases and symptoms.

Recall from Chapter 1 that the hypothesis proposed is that modeling both known and unknown outbreak diseases in a hybrid system can lead to better expected disease outbreak detection performance than modeling known outbreak diseases only. Therefore, to test this hypothesis, I constructed a detection system that only models known outbreak diseases, which I call a disease-specific detector, and a hybrid detector that models both known and unknown outbreak diseases. The experimental evaluation compares the disease-specific detector and the hybrid detector in order to evaluate the hypothesis. Section 5.3 describes the experimental methodology, and Section 5.4 provides a detailed, technical statement of the dissertation hypothesis. Section 5.5 presents an overview of the representative experimental results.

Section 5.6 contains results from using a uniform prior over the appearance of outbreak diseases being modeled. In particular, that section contains a statistical analysis for assessing whether the disease-specific detector and the hybrid detector perform significantly differently in terms of expected detection performance, and Section 5.7 contains a decision analysis that investigates the circumstances in which the hybrid detector has a greater expected utility than the disease-specific detector in monitoring for a disease outbreak. Finally, Section 5.8 describes a sensitivity analysis over the prior probability of the appearance of an unknown disease $d_*$ or a partially-known disease $d_{*p}$ being modeled.

## 5.1 CHOICES OF DISEASES AND SYMPTOMS

Recall from Chapter 4 that there are 13 CDC-A$^+$ diseases and 54 possible disease symptoms that we model. In this dissertation research, I chose three diseases from the 13 CDC-A$^+$ diseases for use in the experiments described below. The three diseases are *cryptosporidiosis*, *early stage anthrax*, and *inhalation tularemia*. For each of the three diseases, I model three disease symptoms *cough*, *headache*, and *abdominal pain*. I selected the three diseases and the three symptoms because these diseases and their symptoms contain a wide variety of distributional patterns (over probability of a person's symptom state given that person has a specific outbreak disease) among the 13 total CDC-A$^+$ diseases as described below.

Let $p_k^j$ represent the probability of a person having the $j$th symptom, given that person has disease $d_k$, where $1\leq\ j \leq 3$ and $1\leq\ k \leq 3$. The distributions over $p_k^j$ were estimated by using expert assessment as described in Section 4.1.2. Recall from Section 4.3.2 that $P(E_i^j \mid PD_i = d_k)$ also is denoted $p_k^j$ for a specific person $i$. Figure 5.1a-c plot the distributions over $P(E_i^j = \textit{abdominal pain} \mid PD_i = d_k)$, $P(E_i^j = \textit{cough} \mid PD_i = d_k)$, and $P(E_i^j = \textit{headache} \mid PD_i = d_k)$, respectively, where $d_k$ is either *cryptosporidiosis*, *early stage anthrax*, or *inhalation tularemia*. That is, as discussed in Chapter 3, we do not assume that these probabilities are known precisely, but rather, there is a distribution over them.

The following sections describe how I use these three diseases and their symptoms to create datasets and set up the experiments.

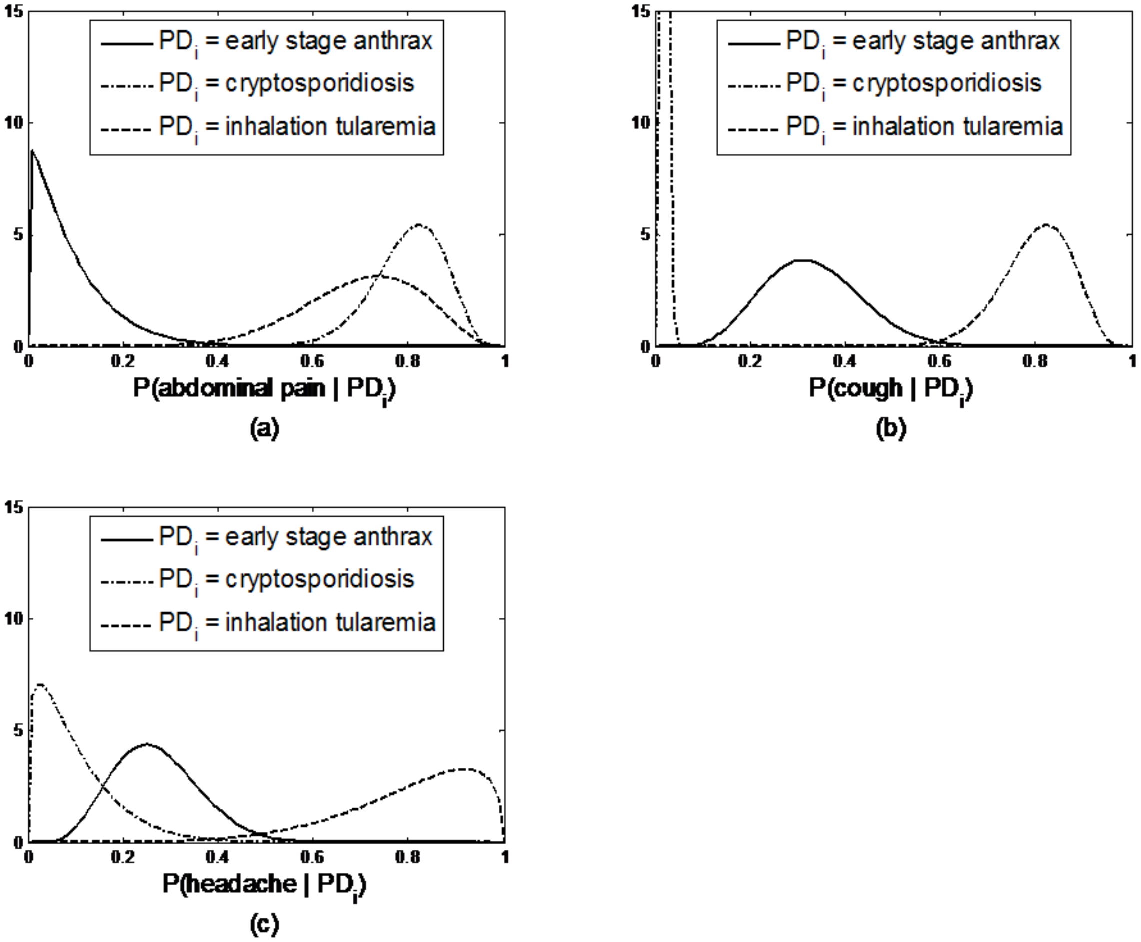


**Figure 5.1** The probability density functions of each symptom probability given each disease.

## 5.2 CREATING DATASETS

I created datasets by using information from real ED events along with simulated outbreak cases produced by a linear outbreak simulator called the FLOO simulator (Neill 2005), which is described in detail below. A time series of real ED cases was used to generate the number of patients who came to the ED in the previous 24-hour period.

I obtained real ED cases for 2004 and 2005 from several hospitals in Allegheny County, Pennsylvania, of which I selected the largest hospital to obtain information for use in my dissertation experiments. The mean number of patients who visited the ED of this hospital per day was about 130, and the population size that is covered by this hospital is approximately 60,000. The time series of real ED cases of the hospital was used to determine the number of people who came to the ED on a given day without any disease outbreak.

The MBH algorithm takes as input evidence from every person in the population from the most recent 24-hour period, where each person has three symptom states (*cough*, *headache*, and *abdominal pain*) that can each either be present or absent, as described in Section 5.1. Every dataset is generated by overlaying the simulated outbreak cases onto a time series of background cases. Recall that the three symptom states are modeled as conditionally independent. Thus, assuming this state of independence, I simulate the symptom states of background cases and outbreak cases. The background time series was generated using the time series of real ED cases, where I use $n_0$ to represent the number of ED cases for a specific day. On any given day (on or after midnight that day and before midnight the next day), I sampled from Beta($\alpha_0^j$, $\beta_0^j$) to determine the probability ($p_0^j$) of a person having symptom $j$, given that person had disease $d_0$ as described in the non-outbreak disease model in Section 4.1.1, where $1 \leq j \leq 3$. I then sampled from Beta($n_0$, $p_0^j$) to determine the number of people having that symptom when there was no disease outbreak in the population on that day, where $n_0$ is the number of people who in reality came to the ED on that day. I did this for each of the three symptoms I selected. Thus, for example, a possible ED patient case that might be generated by this process is (*cough* = present, *headache* = absent, *abdominal pain* = absent). A total of $n_0$ such cases would be generated for the current day. These generated cases with simulated symptom states are called *background cases* for that day. Note that I only created a single time series of background cases for all the experiments described below. Every dataset (outbreak scenario) was created by overlaying the simulated outbreak cases (as described below) onto this time series of background cases.

From among the three outbreak diseases (*cryptosporidiosis*, *early stage anthrax*, and *inhalation tularemia*) being used in the evaluation, $d_k$ represents the specific outbreak disease that is occurring. The simulated outbreak cases with disease $d_k$ were generated using the "Fictional Linear Onset Outbreak" (or "FLOO") simulator described in (Neill 2005), where $d_k$ (for $1 \leq k \leq 3$) represents a specific outbreak disease out of the three diseases I selected. A simulated FLOO($\Delta$,$T$) outbreak has duration $T$. It generates $t\Delta$ cases on day $t$ of the outbreak ($0 < t \leq T/2$), and then generates $T\Delta/2$ cases per day for the remainder of the outbreak. Figure 5.2 shows an example of outbreak cases generated using FLOO(1, 10), in which the maximum number of outbreak cases generated is calculated as $T\Delta/2 = 10\times 1 / 2 = 5$, as shown in this figure.

Let $n_k$ be the simulated outbreak cases generated by FLOO($\Delta$,$T$) per day. I sampled from the Beta distribution, Beta($\alpha_k^j$, $\beta_k^j$), to determine the probability ($p_k^j$) of a person having symptom $j$, given that person had disease $d_k$, as described in the disease-specific model in Section 4.1.2. I then sampled from Beta($n_k$, $p_k^j$) to determine the number of the outbreak cases having disease $d_k$ with symptom $j$, where $1\leq k \leq 3$ and $1 \leq j \leq 3$. Thus, for example, a possible outbreak patient case having disease $d_k$ that might be generated by this process is (*cough* = present, *headache* = present, *abdominal pain* = absent). A total of $n_k$ such cases would be generated for the current day.

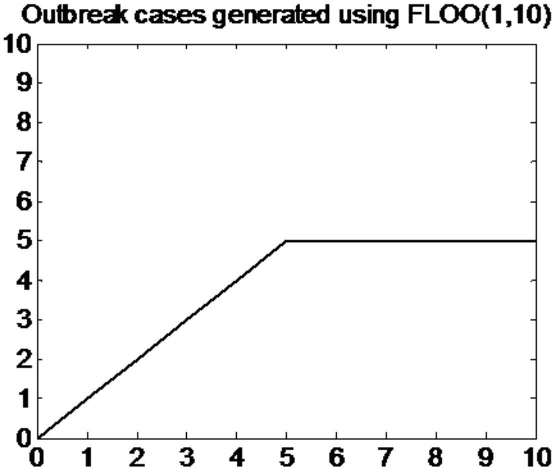


**Figure 5.2** An example showing outbreak cases generated using FLOO(1,10), in which 1, 2, 3, 4 outbreak cases were generated on day 1 to day 4, respectively, and 5 outbreak cases were generated per day for day 5 to day 10.

I applied the method described in (Cooper 2004) to randomly generate the onset dates of the simulated outbreak due to disease $d_k$. That is, I randomly selected 8 dates from each of the 12 consecutive months in 2005 as the starting dates in which simulated outbreaks due to disease $d_k$ were created. I created one dataset by overlaying the simulated outbreak cases produced by FLOO($\Delta$,$T$) onto the background ED cases starting from the onset date to $T$-1 days thereafter. I thus created 8 × 12 = 96 datasets (scenarios) of outbreaks due to disease $d_k$.

In order to evaluate the BH algorithm's detection performance using different scales of disease-outbreak scenarios, I generated outbreak cases using three sets of FLOO parameters. To specify the FLOO parameters, I first estimated the standard deviation ($\sigma_0$) of the number of ED patients per day from real ED data for 2004; $\sigma_0 \approx 8$. I then created three disease-outbreak scenarios, FLOO(1, 10), FLOO(2, 14) and FLOO(4, 14), in which the maximum outbreak cases overlaid per day correspond to approximately 0.6$\sigma_0$ (low severity), 1.8$\sigma_0$ (medium severity) and 3.5$\sigma_0$ (high severity), respectively. Using each of these three FLOO parameter settings, I generated 96 outbreak scenarios for a specific disease and three evidential features. The outbreak cases in each of these outbreak scenarios were overlaid onto the single time series of ED background cases to generate a complete scenario.

Recall that the three diseases in the experiments are *cryptosporidiosis*, *early stage anthrax*, and *inhalation tularemia*. As mentioned above, for each disease I used three different FLOO($\Delta$,$T$) settings to inject cases. A total of 3 (FLOO setting possibilities) × 3 (disease possibilities) × 96 (outbreak scenarios) = 864 datasets were created, with each containing three evidential features per patient case (*cough*, *headache*, and *abdominal pain*) that could each be either present or absent.

*Discussion of noise effects of the datasets*

Recall from Section 4.1 that the frequency of a symptom state in the individuals with the non-outbreak disease $d_0$ is assumed to have a Beta distribution, as for example the frequency of cough in the individuals with $d_0$. Using this example, the simulated cough cases with $d_0$ were generated using a Beta-Binomial model. The remainder of this section shows that the variance of the number of cough cases generated from the Beta-Binomial model is greater than the variance of cases that would be generated by the Beta model (except when the total number of ED cases for a specific day is one, in which the two variances are equal). Thus, by design the data contain considerable noise, which arguably makes it more realistic than having less noisy data.

Let $n$ be the total number of ED cases for a specific day; Appendix B contains the actual number of cases used in the experiments. For each of the $n$ cases, I determined its cough status by simulating from the Beta model. Let $p_i$ be a random sample for the $i$th ED case drawn from the Beta model of disease $d_0$, then $p_i$ represents the probability that the $i$th ED case is a cough case. Let $I_i$ be an indicator of the event that the $i$th ED case is a cough case. Then $I_i$ takes values

0 and 1, and the probability of $I_i$ equals 1 is $p_i$. Let $X$ be a random variable that represents the number of cough cases generated from the Beta model. Then $X$ is the sum of $I_i$ for $i$ from 1 to $n$, and $I_1,\ldots, I_n$ are independent. Thus, we derive the variance of $X$ as

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^{n} I_i\right) = n\mathrm{Var}(I_1) = n\left[\mathrm{E}\left(I_1^2\right) - \left(\mathrm{E}(I_1)\right)^2\right] = n\left[\mathrm{E}(I_1) - \left(\mathrm{E}(I_1)\right)^2\right].$$

By law of total expectation (Bertsekas 2002), we have

$$\mathrm{Var}(X) = n\left[\mathrm{E}(I_1) - \left(E(I_1)\right)^2\right] = n\left[\mathrm{E}\left(\mathrm{E}(I_1 \mid p_1)\right) - \left(\mathrm{E}\left(\mathrm{E}(I_1 \mid p_1)\right)\right)^2\right] = n\left[\mathrm{E}(p_1) - \left(\mathrm{E}(p_1)\right)^2\right].$$

Alternatively, let us consider that the number of cough cases with disease $d_0$ is determined by sampling from the Beta-Binomial model. Let $Y$ be a random variable that represents the number of cough cases generated from the Beta-Binomial model. Let $p$ be a random variable that represents the probability of cough and follows a Beta distribution. Thus, by law of total variance (Bertsekas 2002), we can derive the variance of $Y$ as follows:

$$\mathrm{Var}(Y) = \mathrm{E}\left(\mathrm{Var}(Z \mid p)\right) + \mathrm{Var}\left(\mathrm{E}(Z \mid p)\right),$$

where $Z$ is a random variable of the Binomial model with parameter $p$. Then we have $\mathrm{Var}(Z \mid p) = np(1 - p)$ and $\mathrm{E}(Z) = np$. By definition, $p$ has the same distribution as $p_1$. Thus,

$$\mathrm{Var}(Y) = \mathrm{E}\left[np(1 - p)\right] + \mathrm{Var}(np) = n\mathrm{E}\left[p_1 - p_1^2\right] + \mathrm{Var}(np_1).$$

Given $\mathrm{Var}(np_1) = n^2\mathrm{Var}(p_1) = n^2\left[\mathrm{E}\left(p_1^2\right) - \left(\mathrm{E}(p_1)\right)^2\right]$, we can further write $\mathrm{Var}(Y)$ as

$$\begin{aligned}
\mathrm{Var}(Y) &= n\mathrm{E}(p_1) - n\mathrm{E}\left(p_1^2\right) + n^2\left[\mathrm{E}\left(p_1^2\right) - \left(\mathrm{E}(p_1)\right)^2\right] \\
&\geq n\mathrm{E}(p_1) - n\mathrm{E}\left(p_1^2\right) + n\left[\mathrm{E}\left(p_1^2\right) - \left(\mathrm{E}(p_1)\right)^2\right] \quad \text{, since } n \geq 1\text{. The equality holds when } n = 1. \\
&= n\left[\mathrm{E}(p_1) - \left(\mathrm{E}(p_1)\right)^2\right].
\end{aligned}$$

The above two equations show that when more than one patient case present to the ED, the number of cough cases generated from the Beta-Binomial model has a greater variance than those generated from the Beta model. Thus, besides the general noise (random) effects associated with random symptom-state-sampling, there is another level of noise effects that result from sampling from using the Beta-Binomial model, which results in a greater variance than sampling from the Beta model.

## 5.3 EVALUATION METHODS

In order to evaluate the disease-specific model (DSM), the unknown disease model (UDM), and the partially-known disease model (PDM) described in Section 4.1, I first constructed outbreak detection systems using each of these models in the framework of the univariate BH (UBH) algorithm and the multivariate BH (MBH) algorithm, respectively, and then evaluated the detection performance of these algorithms, as described below.

### 5.3.1 Experimental design

Recall from Section 1.2 that the hypothesis of this dissertation is that modeling both known and unknown outbreak diseases in a hybrid system can lead to better expected disease outbreak detection performance than modeling known outbreak diseases only. In order to evaluate this hypothesis, I constructed two experiments in which an outbreak is occurring due to some disease $d_u$ with the characteristics: (1) disease $d_u$ is known to us and has been modeled in DSM, and (2) disease $d_u$ is unexpected and not explicitly modeled in DSM.

I first describe an example of a disease outbreak due to *early stage anthrax*, namely $d_u$ = *early stage anthrax*. I construct two experiments, as described above, which are denoted as Exp. 1 and Exp. 2, respectively. In Exp. 1, we assume that early stage anthrax is modeled in DSM. In

UDM, we not only model early stage anthrax but also model an unknown disease $d_*$. Similarly in PDM, we model early stage anthrax and a partially-known disease $d_{*p}$. Since *early stage anthrax* is known to us in Exp. 1, the partially-known disease is modeled using a mixture of prior distributions that is composed of a uniform prior distribution and the three selected outbreak diseases *cryptosporidiosis, early stage anthrax,* and *inhalation tularemia*. In Exp. 1, $d_*$ in UDM ($d_{*p}$ in PDM) is an extra disease that is not actually occurring in the experiment. An extra disease increases the number of false alert disease possibilities. Thus, detection performance might degrade relative to DSM, which only is modeling the actual outbreak disease in Exp. 1. Nonetheless, I conjecture that modeling $d_*$ ($d_{*p}$) will not significantly degrade detection performance because $d_*$ and $d_{*p}$ are able to model $d_u$. Exp. 1 investigates the extent to which this conjecture holds.

In Exp. 2, we assume that we do not know the outbreak disease that is causing an ongoing disease outbreak. In Exp. 2, suppose DSM models only *cryptosporidiosis* as a possible outbreak disease. Suppose also that in addition *cryptosporidiosis*, UDM also models an unknown disease $d_*$, and PDM models a partially-known disease $d_{*p}$ using a mixture of prior distributions that is composed of a uniform prior distribution and the priors of two of the three selected outbreak diseases that we know about and have modeled, namely, *cryptosporidiosis* and *inhalation tularemia*. Suppose that the simulated outbreak is due to *early stage anthrax*. Since DSM does not model *early stage anthrax*, it may be difficult for it to detect it. In contrast, UDM contains the "catch all" disease $d_*$ that can match a wide variety of disease presentations, including *early stage anthrax*. Thus, UDM's detection performance might surpass that of DSM. By similar reasoning, PDM's detection performance might surpass that of DSM as well. Exp. 2 investigates whether these results occur.

Table 5.1 summarizes the two experiments constructed for DSM, UDM, and PDM, in which the two experiments are denoted as Exp. 1 and Exp. 2, respectively. In this table, both experiments have simulated outbreaks due to disease $d_u$, which is one of the three selected diseases (*cryptosporidiosis*, *early stage anthrax*, and *inhalation tularemia*) and is determined using leave-one-out experiments as described below. Disease $d_u$ is being modeled in DSM, PDM, and UDM in Exp. 1. In contrast in Exp. 2 $d_u$ is not being modeled in any of these three. DSM models disease $d_u$ in Exp. 1, while it models another (known) disease $d_v$ in Exp. 2, where $d_v \neq d_u$. In each of the two experiments, UDM models an unknown disease $d_*$ using the method described in Section 4.1.3, and PDM models a partially-known disease $d_{*p}$ using the method described in Section 4.1.4.

See Section 4.1 for the general methods for modeling non-outbreak diseases and for modeling outbreak diseases that are known, unknown, and partially-known. In the remainder of this section, I describe how I model these diseases in the specific experiments summarized by Table 5.1.

**Table 5.1** A 2 × 3 table that summarizes the two experiments that each involves the use of a disease-specific model (DSM), an unknown disease model (UDM), and a partially-known disease model (PDM).

| | **DSM** | **UDM** | **PDM** |
|---|---|---|---|
| Exp. 1 | Model $d_0$, $d_u$.<br>Simulate outbreak cases from $d_u$. | Model $d_0$, $d_u$, $d_*$.<br>Simulate outbreak cases from $d_u$. | Model $d_0$, $d_u$, $d_{*p}$.<br>Simulate outbreak cases from $d_u$. |
| Exp. 2 | Model $d_0$, $d_v$.<br>Simulate outbreak cases from $d_u$. | Model $d_0$, $d_v$, $d_*$.<br>Simulate outbreak cases from $d_u$. | Model $d_0$, $d_v$, $d_{*p}$.<br>Simulate outbreak cases from $d_u$. |

In each model (DSM, UDM, and PDM) of Exp. 1 and Exp. 2, we model a non-outbreak disease $d_0$ using past ED data that are assumed to contain no disease outbreaks. We also need to model a specific known outbreak disease $d_u$ (or $d_v$), which is modeled using expert judgment. See Section 4.1.1 and Section 4.1.2 for a detailed description of modeling of $d_0$ and $d_u$ (or $d_v$), respectively.

UDM in Exp. 1 and Exp. 2 needs to model an unknown outbreak disease $d_*$ using a non-informative prior distribution, for which I use a uniform distribution. See Section 4.1.3 for a detailed description of the modeling of the $d_*$ disease.

Recall from Section 4.1.4 that I use a mixture of priors to model a partially-known disease $d_{*p}$ in terms of symptom $j$ in the PDM models, and $M$ represents the number of known diseases that have been modeled in the mixture. Using the three selected diseases (*cryptosporidiosis*, *early stage anthrax*, and *inhalation tularemia*) I carry out leave-one-out experiments, in which the disease that is left out is the disease $d_u$ that will be simulated to cause an outbreak in Exp. 1 and in Exp. 2. Exp. 1 simulates a scenario in which the ongoing outbreak is caused by a disease that we know about and have modeled, thus we have $M = 3$ in Exp. 1; Exp. 2 simulates a scenario in which the ongoing outbreak is due to an unknown disease $d_u$, therefore, we have $M = 3 – 1 = 2$ in Exp. 2.

Thus there are 3 + 1 = 4 components in the mixture model in Exp. 1 and 2 + 1 = 3 components in the mixture model in Exp. 2. In each experiment, the additional component represents the condition that we know little about ($d_{*p}$ regarding symptom $j$), and we use a uniform distribution on [0, 1] to model this component of the mixture. Recall from Section 4.1.4 that each component in the mixture model has a prior that represents possible similarities in distribution between disease $d_{*p}$ and any disease of the four components. I assume a uniform prior probability over these components. Thus, the symptom *j* of the partially-known disease $d_{*p}$ is modeled as follows:

$$f\left(p_*^j;\boldsymbol{\theta}_*^j\right)=\frac{1}{M+1}\sum_{k=1}^{M} f_k\left(p_k^j;\boldsymbol{\theta}_k^j\right)+\frac{1}{M+1}\cdot U_{[0,1]}\ , \tag{5.1}$$

where $f_k\left(p_k^j;\boldsymbol{\theta}_k^j\right)$ is a Beta distribution, which is an informative prior derived for the *k*th disease, as described in Section 4.1.2, $U_{[0,1]}$ represents a uniform prior distribution, $M$ = 3 in Exp.1, and $M$ = 2 in Exp. 2.

### 5.3.2 Experimental procedures

In each of the 2 × 3 = 6 detection systems constructed in Exp. 1 and Exp. 2 in Table 5.1, I compute the likelihood ratio $LR = P(E = e \mid O = OB) / P(E = e \mid O = NOB)$ using the following equation, as described in Section 4.2.3.

$$LR=\frac{\sum_{OD\neq d_0} P(E=e\mid OD)P(OD\mid O=OB)}{P(E=e\mid OD=d_0)}. \tag{5.2}$$

In the UDM model in Exp. 1, the sum in Equation 5.2 is taken over $d_u$ and $d_*$, and in the UDM model in Exp. 2, the sum in Equation 5.2 is taken over $d_v$ and $d_*$; in contrast, in the DSM model in Exp. 1, the sum of *OD* consists only of the term $d_u$, and in the DSM model in Exp. 2, the sum of *OD* consists only of $d_v$. The PDM model applies the same strategy as the UDM model when using Equation 5.2, but sums over $d_{*p}$ and $d_u$ in Exp. 1 (or $d_{*p}$ and $d_v$ in Exp. 2).

I performed sensitivity analysis and used a sequence of probability values for $P(OD = d_* \mid O = OB)$ and $P(OD = d_{*p} \mid O = OB)$. Section 5.8 contains a detailed description of this analysis. To convey the basic approach, I restrict this chapter to an example that uses a single prior probability for $P(OD = d_* \mid O = OB)$. For any given outbreak disease $d_u$ (or $d_v$) being modeled, I assume a uniform prior over the specific disease $d_u$ (or $d_v$) and $d_*$, which yields that $P(OD = d_u \mid O = OB) = P(OD = d_* \mid O = OB) = 0.5$ or $P(OD = d_v \mid O = OB) = P(OD = d_* \mid O = OB) = 0.5$. Similarly, this uniform prior was applied to the specific disease $d_u$ (or $d_v$) and $d_{*p}$ for the PDM model described in this chapter.

Recall from Section 4.2.3 and Section 4.3.3 that there is a *fraction* node that needs to be involved in computing the likelihood ratio in Equation 5.2. I constructed the *fraction* node in Figure 4.1 (for the UBH algorithm) and in Figure 4.3 (for the MBH algorithm) using a sequence of uniformly distributed values as $f = i/N$ for $i = 1$ to 10, where $N$ is the total number of people in the population. Recall that the standard deviation ($\sigma_0$) of the mean number of ED patients per day from real ED data for 2004 is approximately $\sigma_0 = 8$. By recognizing a small number of outbreak cases (from one case to $1\sigma_0$ cases, namely 1, 2, …, 8), the detection system can detect disease outbreaks early. For each value $f$ among the ten, I compute $P(E = e \mid OD = d_u, F = f)$ and $P(E = e \mid OD = d_v, F = f)$ as shown in Equation 4.9. I then compute $P(E = e \mid OD = d_u)$, $P(E = e \mid OD = d_v)$, and $P(E = e \mid OD = d_0)$ for the UBH algorithm, as described in Section 4.2.3 and compute the above probabilities for the MBH algorithm, as described in Section 4.3.3. I finally compute the likelihood ratio using Equation 5.2.

I ran the UBH algorithm and the MBH algorithm using the 2 × 3 experimental setup given in Table 5.1, as described next.

**UBH algorithm:** I ran the UBH algorithm on the datasets described in Section 5.2 assuming a given dataset only contains a single evidential feature that is either *abdominal pain*, *cough*, or *headache*. In a given dataset, a simulated patient who came to the ED could either have that symptom or not. The UBH algorithm took as input such data during the most recent 24-hours of simulated time.

Using the three selected diseases that are described in Section 5.1, I did leave-one-out experiments. The disease that is left out is the disease $d_u$ that will be simulated to cause an outbreak in Exp. 1 and in Exp. 2. Recall from Section 5.1 that I also selected three symptoms (*abdominal pain*, *cough*, or *headache*) to be the evidential features, and the UBH algorithm is run on one feature (out of the three features) at a time. I created 96 semi-synthetic datasets using each FLOO($\Delta$,$T$) setting, of which there are three parameter settings for $\Delta$ and $T$, as described in Section 5.2. In Exp. 1, I thus created 3 (disease possibilities for $d_u$) × 3 (symptom possibilities) × 3 (FLOO setting possibilities) = 27 experimental configurations. I then ran the UBH algorithm a total of 3 (disease models: DSM, UDM, and PDM) × 27 (experimental configurations) = 81 times, and there are 96 outbreak scenarios to run each time. In Exp. 2, I created 3 (disease possibilities for $d_u$) × 2 (disease possibilities for $d_v$) × 3 (symptom possibilities) × 3 (FLOO setting possibilities) = 54 experimental configurations and ran the UBH algorithm a total of 3 (disease models: DSM, UDM, and PDM) × 54 (experimental configurations) = 162 times, and there are 96 outbreak scenarios to run each time.

**MBH algorithm:** I ran the MBH algorithm on the multivariate datasets described in Section 5.2. Given the three diseases, I did leave-one-out experiments as described above. As opposed to the UBH algorithm, the MBH algorithm takes as input the three evidential features (*abdominal pain*, *cough*, or *headache*) for each patient case during the most recent 24-hours of simulated time. In Exp. 1, I thus created 3 (disease possibilities for $d_u$) × 3 (FLOO setting possibilities) = 9 experimental configurations and ran the MBH algorithm a total of 3 (disease models: DSM, UDM, and PDM) × 9 (experimental configurations) = 27 times, and there are 96 outbreak scenarios to run each time. In Exp. 2, I created 3 (disease possibilities for $d_u$) × 2 (disease possibilities for $d_v$) × 3 (FLOO setting possibilities) = 18 experimental configurations and ran the MBH algorithm a total of 3 (disease models: DSM, UDM, and PDM) × 18 (experimental configurations) = 54 times, and there are 96 outbreak scenarios to run each time.

Thus, for a specific experimental configuration described above, there are 96 outbreak scenarios to run for a specific disease model (DSM, UDM, or PDM) of a specific detection algorithm (UBH or MBH).

Given the output of the likelihood ratio of one outbreak scenario, I determined its detection time and false alert rate for various detection-ratio thresholds. The detection time is the time from the simulated release until the detection-ratio threshold *r* was exceeded. If a detection threshold has never been exceeded using the output likelihood ratios, the detection time was taken as the maximum duration time of the outbreak. For example, if this happens in an outbreak scenario created using FLOO(1,10), then the detection time is taken to be 10 days. The false alert rate is derived as *FP* / *M*, where *FP* is the number of false alerts that occurred using threshold *r* when each experiment monitored the time series of simulated ED cases (12 months) in which there is no (simulated) outbreak, and *M* is length in months in that time series, namely, *M* = 12.

For each algorithm (UBH and MBH), I applied this process to the 96 outbreak scenarios, computed the expected detection time over the 96 outbreak scenarios, and plotted AMOC curves (Fawcett 1999) for DSM, UDM, and PDM in Exp. 1 and for DSM, UDM, and PDM in Exp. 2. As described above, I thus plotted 27, 54, 9, and 18 sets of AMOC curves for the UBH algorithm in Exp. 1, the UBH algorithm in Exp. 2, the MBH algorithm in Exp. 1, and the MBH algorithm in Exp. 2, respectively, where each set contains three AMOC curves that correspond to DSM, UDM, and PDM, respectively.

## 5.4 TECHNICAL STATEMENT OF THE HYPOTHESIS

Chapter 1 provided the following qualitative statement of the dissertation hypothesis:

> *Modeling both known and unknown outbreak diseases in a hybrid system can lead to better expected disease outbreak detection performance than modeling known outbreak diseases only.*

The chapters and sections since Chapter 1 have provided significant detail related to how to evaluate the above hypothesis. Given that detail, it is now possible in the current section to provide a more technical statement of the dissertation hypothesis.

Let event *G* denote the following event: Given that an outbreak is occurring, the outbreak disease is not explicitly being modeled in the detection system. According to Table 5.1, *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 an unknown or a partially-known disease (in the form of $d_*$ or $d_{*p}$) yields a net gain in detecting disease outbreaks. This evaluation is relative to *q*. If $q = 1$, then modeling $d_*$ or $d_{*p}$ will likely be helpful. If $q = 0$, however, modeling $d_*$ or $d_{*p}$ will be useless and possibly harmful by allowing more chances for a false alert alert. At the end of this chapter, based on the experimental results that follow, I describe a decision analysis that derives range of values of *q* for which modeling $d_*$ or $d_{*p}$ yields a net expected gain in detection performance.

I represent models DSM, UDM, and PDM in Exp. 1 as DSM1, UDM1, and PDM1, respectively, and likewise represent models DSM, UDM, and PDM in Exp. 2 as DSM2, UDM2, and PDM2. Let $\mathrm{E}_{\mathrm{DSM1}}$ be the average detection time of DSM1 over all the experiments constructed for the UBH algorithm described in Section 5.1-5.3 at a false alert rate of one per month[11]. Let $E_{DSM2}$ be the average detection time of DSM2 over all the experiments constructed for the UBH algorithm described in Section 5.1-5.3 at a false alert rate of one per month. Let $E_{DSM} = (1 - q) \times E_{DSM1} + q \times E_{DSM2}$. Define $E_{UDM}$ and $E_{PDM}$ analogously.

The dissertation hypothesis can now be stated more precisely as follows:

*There exists a $q < 1$ such that $E_{PDM} < E_{UDM} < E_{DSM}$.*

If the experimental results support the above hypothesis, they will also support the qualitative hypothesis described at the beginning of this section.

## 5.5 OVERVIEW OF REPRESENTATIVE RESULTS

This section presents the AMOC curves of a selection of representative experimental configurations out of all those described in Section 5.3.2. In this section, for brevity I describe an experimental configuration using the following notation: $d_u$, FLOO($\Delta, T$), and the symptom. An example is $d_u$ = *inhalation tularemia*, FLOO($\Delta, T$) = FLOO(4,14), and symptom = *cough*. This example represents an experimental configuration with 96 simulated outbreak scenarios due to disease *inhalation tularemia*, with simulated *inhalation tularemia* cases generated using FLOO simulator with parameter $\Delta = 4$ and $T = 14$, and with each outbreak case being observed present or absent for the symptom *cough*.

If a symptom is specified in an experimental configuration, such as the symptom *cough*, then the univariate BH algorithm (UBH) using a specific model (DSM, UDM, or PDM) is run under that experimental configuration; if a symptom is specified as *multivariate*, then the

[11] A directly analogous hypothesis exists and will be tested for UBH at zero false alerts per month, for MBH at one false alert per month, and for MBH at zero false alerts per month as well, but for brevity these are not described here explicitly.

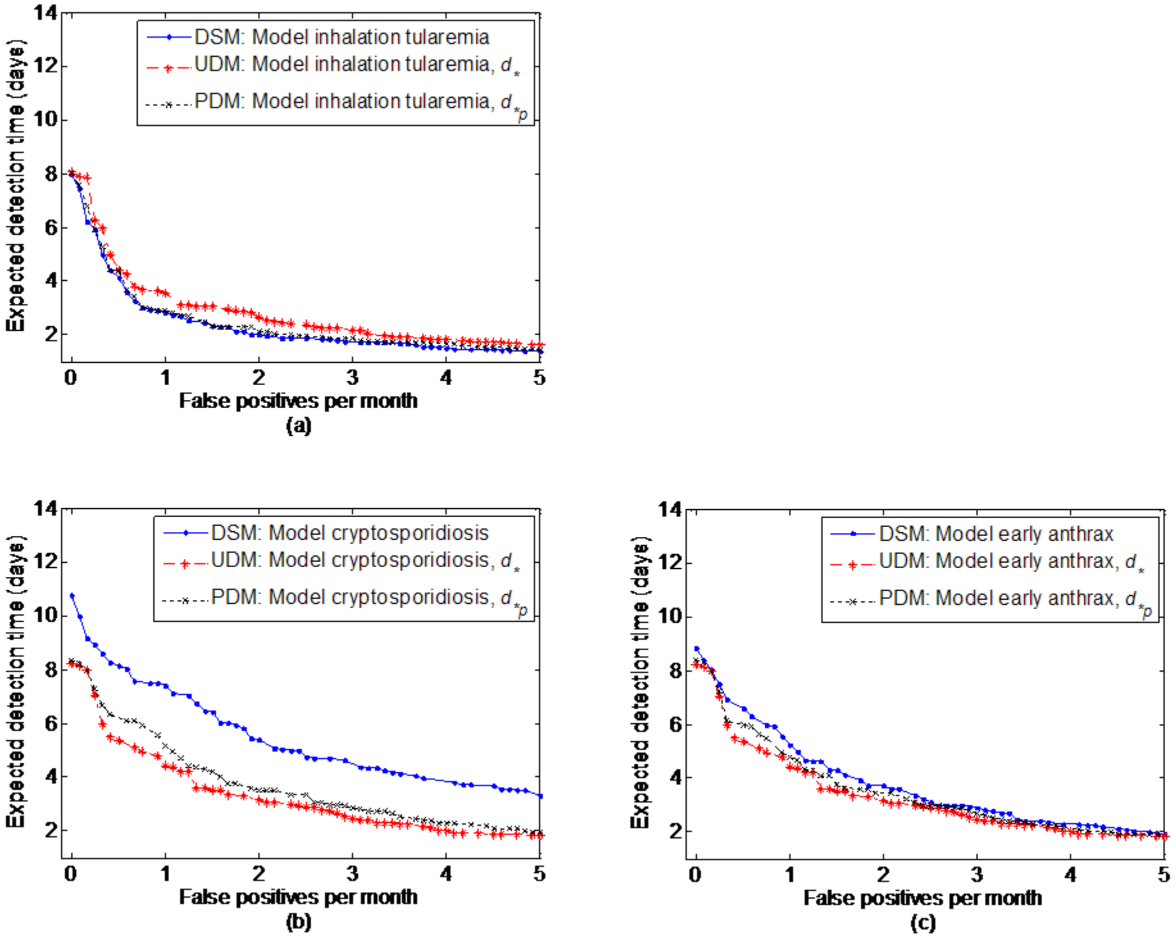
multivariate BH algorithm (MBH) using a specific model (DSM, UDM, or PDM) is run, and the model used is specified in the legend of each figure. For example, an experimental configuration of $d_u$ = *inhalation tularemia*, FLOO($\Delta$,$T$) = FLOO(4,14), and symptom = *multivariate* represents running the MBH algorithm on 96 simulated outbreak scenarios of *inhalation tularemia*, with simulated *inhalation tularemia* cases generated using FLOO(4,14), and with each case symptoms *abdominal pain*, *cough*, and *headache* with a value of present or absent.

### 5.5.1 AMOC curves of the UBH algorithm

Figure 5.3 shows three sets of AMOC curves for the UBH algorithm when the simulated outbreak is due to *inhalation tularemia* and the symptom being modeled is *cough*. The figure caption explains additional details.

Figure 5.3a shows the AMOC curves for Exp. 1 using the above experimental configuration, where the simulated outbreak is due to disease *inhalation tularemia,* which is being modeled. As shown, DSM performs slightly better than UDM and PDM. In particular, at one false alert per month, which is frequently cited as an upper bound of a tolerable rate, UDM has an expected detection time of approximately 3.6 days while DSM and PDM each has an expected detection time of approximately 2.9 days. Thus, DSM and PDM detect a simulated *inhalation tularemia* outbreak about 0.7 days earlier than UDM. At zero false alerts per month, the three models take approximately the same time to detect the disease outbreak. This result supports that when an existing outbreak disease is being modeled (in this situation it is *inhalation tularemia*), then also modeling $d_*$ in UDM (or $d_{*p}$ in PDM) does worsen the performance somewhat, but not dramatically.

**Experimental configuration:**
**$d_u$ = *inhalation tularemia*, FLOO setting = FLOO(4,14), and symptom = *cough***



(a)

(b)

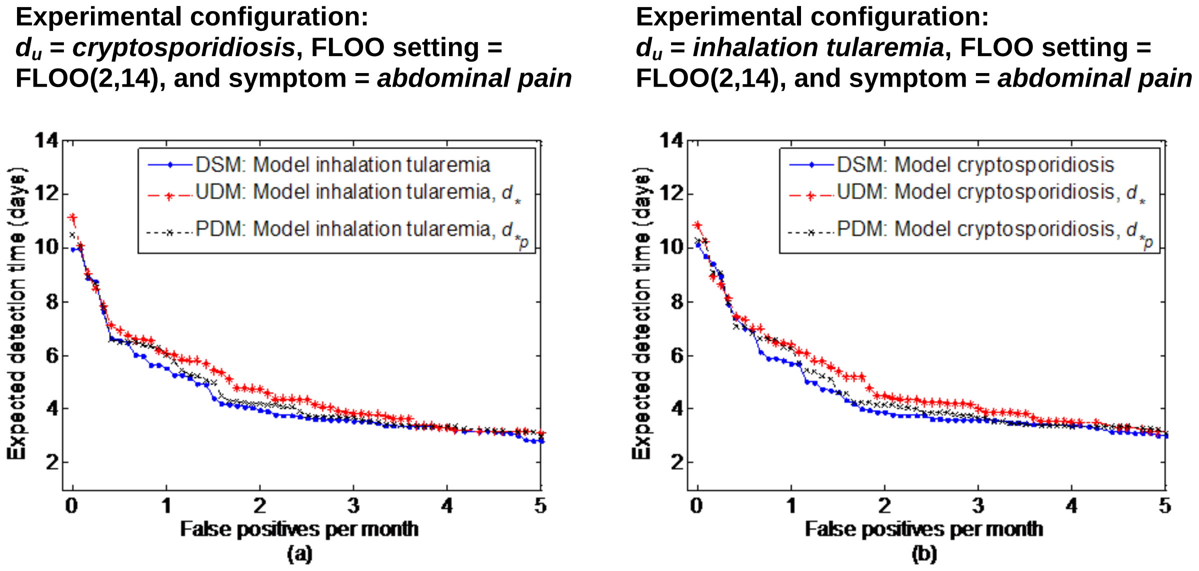
(c)

**Figure 5.3** (a) shows the AMOC curves of Exp. 1, and (b)-(c) show the AMOC curves of Exp. 2 with $d_v$ = *cryptosporidiosis* in (b) and $d_v$ = *early stage anthrax* in (c).

Figure 5.3b and 5.3c show the AMOC curves for Exp. 2, where the outbreak disease *inhalation tularemia* is not being explicitly modeled. In contrast, only *cryptosporidiosis* is explicitly modeled in the experiment that has the results shown in Figure 5.3b. Similarly, only *early stage anthrax* is explicitly modeled in the experiment shown in Figure 5.3c. In Figure 5.3b, at one false alert per month, DSM has an expected detection time of 7.5 days, and PDM has an expected detection time of 5.6 days. UDM, however, has an expected detection time of only 4.8 days, which is 2.7 days earlier than DSM and 0.8 days earlier than PDM. In Figure 5.3c, the gain in detection time of UDM over DSM is not as large as that in Figure 5.3b, due to the fact that the distribution of the probability of the *cough* symptom of *inhalation tularemia* and that of *early stage anthrax* are similar, as shown previously in Figure 5.1. The results in Figure 5.3b and 5.3c support that when the ongoing outbreak disease is not being explicitly modeled, then modeling $d_*$ or $d_{*p}$ can be quite helpful in detecting the unmodeled disease. While Figure 5.3 shows a set of experimental results that support the dissertation hypothesis, there are some exceptions, as shown below.

Figure 5.4a shows the AMOC curves for Exp. 2 using an experimental configuration of $d_u$ = *cryptosporidiosis*, FLOO(2,14), and *abdominal pain*, where an ongoing outbreak is caused by *cryptosporidiosis*, which is not modeled in the detection model, while just disease *inhalation tularemia* is modeled in DSM, UDM, and PDM. The dissertation hypothesis predicts that modeling an unknown disease $d_*$ in UDM and $d_{*p}$ in PDM would help detect the disease outbreak. That is, UDM and PDM are expected to have better expected detection performance than DSM. However, as shown in Figure 5.4a, DSM has a slightly shorter expected detection time than UDM and PDM at the false alert rates of zero and one per month.

Figure 5.4b provides the AMOC curves for Exp. 2 using an experimental configuration of $d_u$ = *inhalation tularemia*, FLOO(2,14), and *abdominal pain*, where an ongoing outbreak is caused by *inhalation tularemia*, which is not modeled in the detection model, while just disease *cryptosporidiosis* is modeled in DSM, UDM, and PDM. As described above, it is expected that UDM and PDM would have better expected detection performance than DSM. However, the results again show that DSM has modestly better expected detection performance than UDM and PDM.



**Experimental configuration:**
***$d_u$* = *cryptosporidiosis*, FLOO setting = FLOO(2,14), and symptom = *abdominal pain***

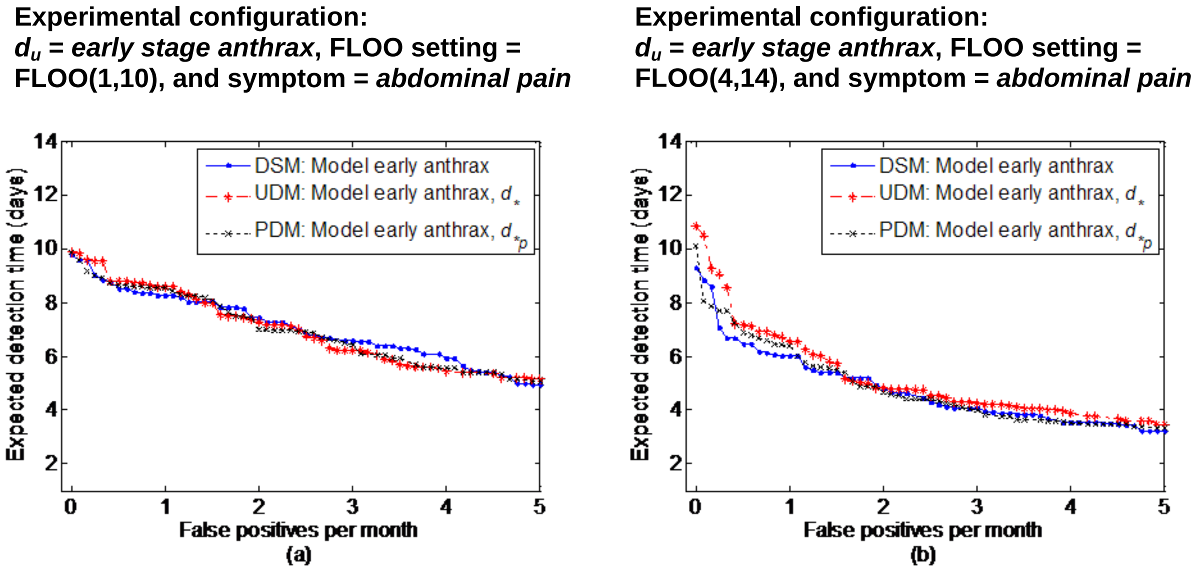
**Experimental configuration:**
***$d_u$* = *inhalation tularemia*, FLOO setting = FLOO(2,14), and symptom = *abdominal pain***

(a)

(b)

**Figure 5.4** AMOC curves of two experimental configurations for Exp. 2.

A possible reason for the results in Figure 5.4a and 5.4b is that *cryptosporidiosis* and *inhalation tularemia* have a similar distribution of the symptom *abdominal pain* (as shown in Figure 5.1a). Thus, even when an ongoing outbreak is due to *cryptosporidiosis*, modeling *inhalation tularemia* only in DSM would lead to a better expected detection performance than modeling it in UDM and PDM. An analogous explanation exists for the exception shown in Figure 5.4b.



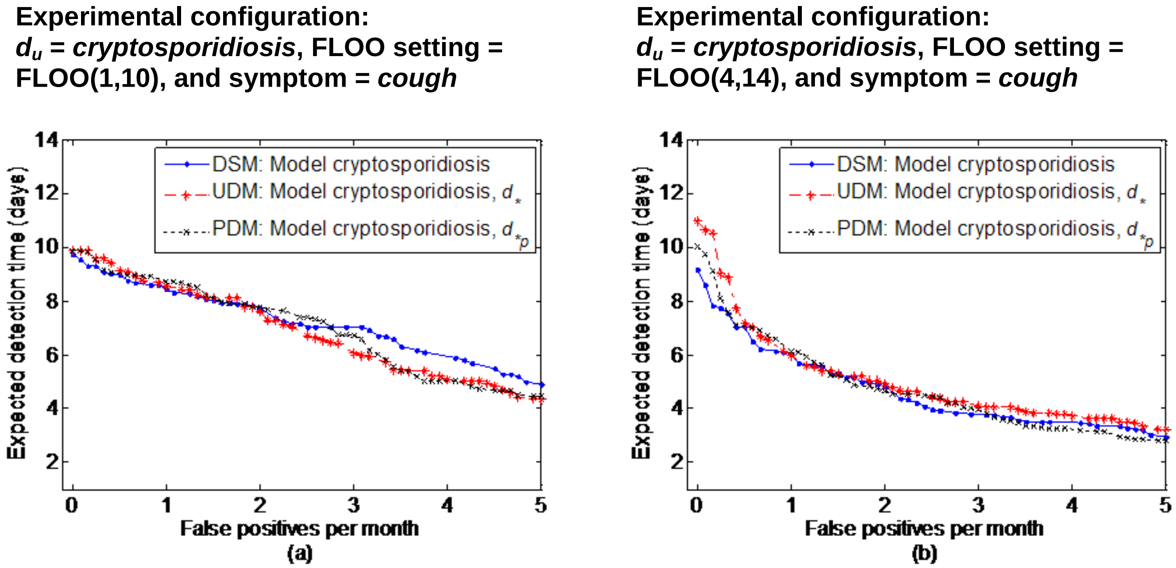
**Experimental configuration:**
***$d_u$* = *early stage anthrax*, FLOO setting = FLOO(1,10), and symptom = *abdominal pain***

**Experimental configuration:**
***$d_u$* = *early stage anthrax*, FLOO setting = FLOO(4,14), and symptom = *abdominal pain***

**Figure 5.5** AMOC curves of two experimental configurations for Exp. 1.

Figure 5.5a shows the AMOC curves for Exp. 1 in an experimental configuration of $d_u$ = *early stage anthrax*, FLOO(1,10), and *abdominal pain*, where DSM is expected to outperform UDM and PDM. However, the three models have similar expected detection performance. In particular, at zero false alerts per month, each of the three models has an expected detection time at approximately 10 days, which means the three models barely detect the ongoing outbreak in most of the 96 outbreak scenarios. Using FLOO(1,10), the maximum number of *early stage anthrax* cases overlaid onto the background cases is 1 × 10 / 2 = 5, which is less than one standard deviation ($\sigma_0$ = 8) of the background time series. It appears that the outbreak signal in this experimental configuration is too weak for any disease model to perform detection well, thus making all three disease models perform equally poorly.

I then overlaid more *early stage anthrax* cases onto the background cases using FLOO(4,14) while keeping other parameters in the experimental configuration the same as those in Figure 5.5a. Using FLOO(4,14) makes the maximum number of *early stage anthrax* cases equal to 4 × 14 / 2 = 28, which is 4 cases more than three standard deviations of the background time series. Figure 5.5b shows the AMOC curves for this new experimental configuration, where DSM outperforms UDM and PDM, as expected. In particular, at one false alert per month, UDM has an expected detection time of 6.7 days while DSM has an expected detection time of 6 days, which is 0.7 days faster.



**Experimental configuration:**
***$d_u$* = *cryptosporidiosis*, FLOO setting = FLOO(1,10), and symptom = *cough***

**Experimental configuration:**
***$d_u$* = *cryptosporidiosis*, FLOO setting = FLOO(4,14), and symptom = *cough***

**Figure 5.6** AMOC curves of two experimental configurations for Exp. 1.

Figure 5.6a also shows an exception in Exp. 1, where DSM has detection performance similar to UDM and PDM. However, it was expected that DSM would outperform UDM and PDM in this experimental configuration because an ongoing outbreak is due to *cryptosporidiosis*, which is known to us and explicitly modeled in DSM. When a stronger outbreak signal is injected in the experiment constructed in Figure 5.6b, DSM detects the outbreak 1.8 days faster than UDM at zero false alerts per month and 1.2 hours faster than UDM at one false alert per month.

It is somewhat surprising that UDM performs no worse than PDM in Exp. 2 in some outbreak scenarios, as for example the scenario shown in Figure 5.4a-b at one false alert per month. We now discuss a rationale for these results. In Exp. 2, a mixture of priors is used for implicitly modeling disease $d_u$, which is the cause of the ongoing outbreak. PDM computes a likelihood using a mixture of priors that is composed of three components: a uniform distribution and the prior distribution for each of the two diseases other than $d_u$. If each of these two diseases has a probability distribution of its symptom state that is quite different than that of disease $d_u$, then it is not surprising that the proposed mixture of priors is not as effective as the uniform prior in modeling unknown disease. For example, as shown Figure 5.1b, the three selected outbreak diseases have little overlap in their probability distributions for *cough*. Whichever among the three diseases has been selected as disease $d_u$ that causes the ongoing outbreak, the mixture of priors that contains the other two diseases is expected to not be as effective as the uniform prior in modeling disease $d_u$. Figure 5.7 shows such an example where the ongoing outbreak is due to disease *early stage anthrax,* namely $d_u$ = *early stage anthrax*. The uniform distribution shown in this graph will get a greater likelihood than the mixture distribution if the cough states of early stage anthrax cases were obtained by simulating $P(cough \mid d_u)$ with $d_u$ = *early stage anthrax*.

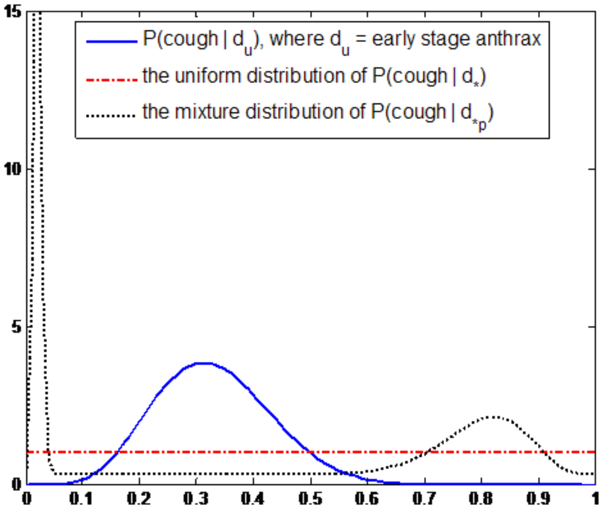


**Figure 5.7** Probability density functions of the probability of cough given outbreak disease *early stage anthrax*, the unknown disease $d_*$, and the partially-known disease $d_{*p}$. The mixture distribution is composed of a mixture of a uniform distribution, the distribution for cryptosporidiosis, and the distribution for inhalation tularemia.

### 5.5.2 AMOC curves of the MBH algorithm

Recall from the experiments associated with Figure 5.6a that each of the DSM, UDM, and PDM models barely detected the outbreak in the 96 outbreak scenarios using the UBH algorithm. In particular, the expected detection time at zero false alerts per month is approximately 10 days. Figure 5.8a-c show the experimental configurations of $d_u$ = *cryptosporidiosis* and FLOO(1,10) of Exp. 1 and Exp. 2 using the MBH algorithm. As opposed to the UBH algorithm, the MBH algorithm has an expected detection time (at zero false alerts per month) of 9.1 days for each of the DSM, UDM, and PDM models, which is 0.9 days faster than using the UBH algorithm. These results suggest that the MBH algorithm performs better than the UBH algorithm in outbreak detection. As with the case of comparing the performance of the three disease models, it was expected that DSM would outperform UDM and PDM in 5.8a, and UDM and PDM would outperform DSM in 5.8b-c. However, the performance of DSM, UDM, and PDM are at approximately the same level in the experimental configurations in 5.8a-c. I conjecture that the outbreak signal in this experimental configuration is not strong enough for the three disease models to be differentiated, in a way similar to what was seen in the UBH results above.

I then injected more outbreak cases using FLOO(4,14) while keeping other parameters in the experimental configuration the same as those in Figure 5.8a-c. Figure 5.8d-f show the AMOC curves for Exp. 1 and Exp. 2 using these stronger outbreaks. In particular, as shown in Figure 5.8d, DSM detects the outbreak 1.2 days faster than UDM and 0.3 days faster than PDM at one false alert per month, as expected. As shown in Figure 5.8e, UDM has an expected detection time that is 1.8 days shorter than DSM, and PDM has an expected detection time that is one day shorter than DSM at one false alert per month, as expected. The results in Figure 5.8f also show that UDM and PDM perform slightly better than DSM in detecting the ongoing disease outbreak due to *cryptosporidiosis*. The results in Figure 5.8d-f support the hypothesis that DSM would outperform UDM and PDM in Exp. 1, and UDM (and PDM) outperform DSM in Exp. 2.

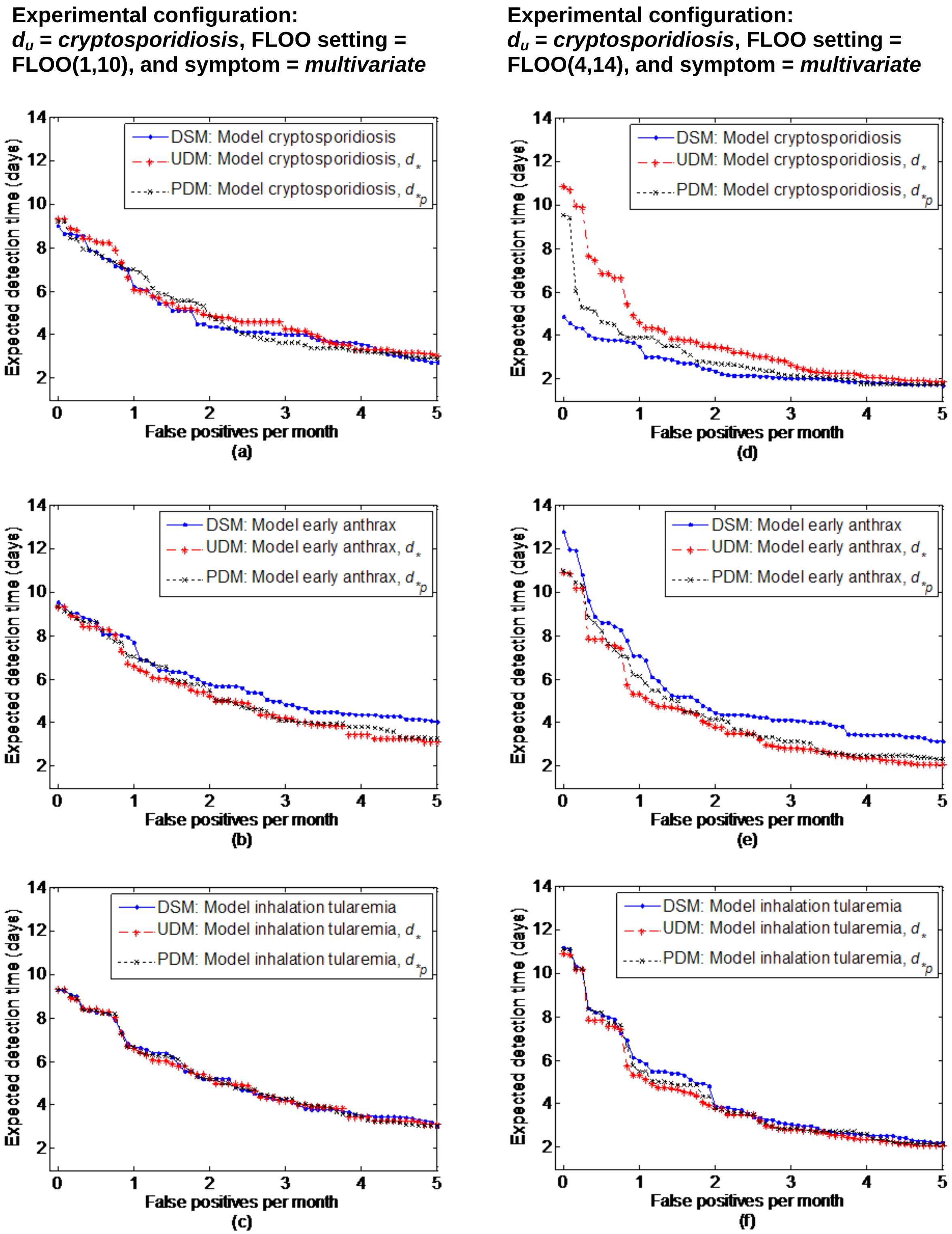


**Figure 5.8** (a) and (d) shows the AMOC curves for Exp. 1, and (b), (c), (e), (f) show the AMOC curves for Exp. 2 with $d_v$ = *early stage anthrax* in (b) and (e), and $d_v$ = *inhalation tularemia* in (c) and (f). The plots in (d)-(f) are different from those in (a)-(c) only in that FLOO($\Delta$,$T$) = FLOO(4,14), and thus, the simulated outbreaks are stronger.

## 5.6 STATISTICAL ANALYSIS

The results presented in Section 5.5 are representative of the complete set of results I obtained in the experiments, but nonetheless they do not give a complete analysis of all the results. This section describes an analysis of all the results. In particular, I performed statistical analyses to evaluate the probabilistic outputs from the UBH algorithm and the MBH algorithm. The outputs from the UBH algorithm and from the MBH algorithm were evaluated separately. Next, I describe an example procedure for performing a statistical analysis using the UBH algorithm. The results of the statistical analysis are presented for both the UBH algorithm and the MBH algorithm at the end of this section.

Recall from Section 5.3.4 that $E_{DSM}$, $E_{UDM}$, and $E_{PDM}$ are continuous linear functions in $q$. Then if $E_{PDM} < E_{UDM} < E_{DSM}$ holds under $q = 1$, then by continuity of $E_{DSM}$, $E_{UDM}$, and $E_{PDM}$, there must be a $q < 1$ such that $E_{PDM} < E_{UDM} < E_{DSM}$ holds. When $q = 1$, it is easy to see that $E_{DSM} = E_{DSM2}$, $E_{UDM} = E_{UDM2}$, and $E_{PDM} = E_{PDM2}$. Thus the problem is reduced to testing whether $E_{PDM2} < E_{UDM2} < E_{DSM2}$ holds.

To perform a statistical analysis, I ran the UBH algorithm for models DSM, UDM, and PDM on the background time series of ED cases of 2005 that are assumed to contain no disease outbreaks in order to determine their false alert rates under various detection-ratio thresholds. I first select a threshold $r$ that is used for obtaining a rate of one false alert per month[12], since one false alert per month is frequently cited as an upper bound of a tolerable rate. Then threshold $r$ is applied on the output likelihood ratios of an outbreak scenario of a specific experimental configuration to determine its detection time under one false alert per month.

[12] A statistical analysis was performed on a rate of zero false alerts per month as well using a directly analogous procedure.

Using this procedure, I obtained the detection time of all the three disease models (DSM, UDM, and PDM) over all proposed experimental configurations, where 96 outbreak scenarios were created using each experimental configuration. Using the UBH algorithm in Exp. 1, there are three factors (FLOO($\Delta$,*T*), $d_u$, and *symptom*) that affect the detection time (at one false alert per month) from the three disease models, thus creating a total of 3 (FLOO setting possibilities) × 3 (disease possibilities for $d_u$) × 3 (symptom possibilities: *abdominal pain*, *cough*, and *headache*) = 27 experimental configurations. Similarly, The UBH algorithm in Exp. 2 results in a total of 3 (FLOO setting possibilities) × 3 (disease possibilities for $d_u$) × 2 (disease possibilities for $d_v$) × 3 (symptom possibilities: *abdominal pain*, *cough*, and *headache*) = 54 experimental configurations because there are four factors (FLOO($\Delta$,*T*), $d_u$, $d_v$ and *symptom*) that are involved in Exp. 2.

Table 5.2 shows the detection times that were obtained for the three disease models in Exp. 1 over the 27 experimental configurations and the 96 simulated outbreak scenarios. For example, cell(1,1) constructs a single scenario of an outbreak due to a known disease *cryptosporidiosis* with an outbreak generator setting of FLOO(1,10), and it contains the detection times for this scenario for models DSM1, UDM1, PDM1, where DSM1 models *cryptosporidiosis*, UDM2 models *cryptosporidiosis* and an unknown disease $d_*$, and PDM2 models *cryptosporidiosis* and a partially-known disease $d_{*p}$.

Table 5.3 shows the detection times that were obtained for the three disease models in Exp. 2 over the 54 experimental configurations and the 96 simulated outbreak scenarios. For example, cell(1,1) contains the detection times for a single scenario for models DSM2, UDM2, PDM2, where DSM2 models *early stage anthrax*, UDM2 models *early stage anthrax* and an unknown disease $d_*$, and PDM2 models *early stage anthrax* and a partially-known disease $d_{*p}$, yet the actual outbreak disease is *cryptosporidiosis* with an outbreak generator setting of FLOO(1,10).

**Table 5.2** Detection times that were obtained for models DSM1, UDM1, and PDM1 (in Exp. 1) over 27 experimental configurations × 96 outbreak scenarios.

| Outbreak scenarios | 1 | 2 | … | 27 |
|---|---|---|---|---|
| | FLOO(1,10)<br>$d_u$ = *cryptosporidiosis*, *symptom* = *abdominal pain* | FLOO(1,10)<br>$d_u$ = *cryptosporidiosis*, *symptom* = *cough* | … | FLOO(4,14)<br>$d_u$ = *inhalation tularemia*, *symptom* = *headache* |
| 1 | Detection times (at one false alert per month) of models DSM1, UDM1, PDM1 | Detection times (at one false alert per month) of models DSM1, UDM1, PDM1 | … | Detection times (at one false alert per month) of models DSM1, UDM1, PDM1 |
| 2 | … | … | … | … |
| .<br>.<br>. | .<br>.<br>. | .<br>.<br>. | .<br>.<br>. | .<br>.<br>. |
| 96 | … | … | … | … |

**Table 5.3** Detection times that were obtained for models DSM2, UDM2, and PDM2 (in Exp. 2) over 54 experimental configurations × 96 outbreak scenarios.

| Outbreak scenarios | 1 | 2 | … | 54 |
|---|---|---|---|---|
| | FLOO(1,10)<br>$d_u$ = *cryptosporidiosis*, $d_v$ = *early stage anthrax*, *symptom* = *abdominal pain* | FLOO(1,10)<br>$d_u$ = *cryptosporidiosis*, $d_v$ = *early stage anthrax*, *symptom* = *cough* | … | FLOO(4,14)<br>$d_u$ = *inhalation tularemia*, $d_v$ = *early stage anthrax*, *symptom* = *headache* |
| 1 | Detection times (at one false alert per month) of models DSM2, UDM2, PDM2 | Detection times (at one false alert per month) of models DSM2, UDM2, PDM2 | … | Detection times (at one false alert per month) of models DSM2, UDM2, PDM2 |
| 2 | … | … | … | … |
| .<br>.<br>. | .<br>.<br>. | .<br>.<br>. | .<br>.<br>. | .<br>.<br>. |
| 96 | … | … | … | … |

To perform statistical analysis using the results that were obtained in form shown in Table 5.2[13], I evaluated the following null hypothesis $H_o$: $E_{PDM2} = E_{UDM2} = E_{DSM2}$ versus the alternative hypothesis $H_a$: at least two models have different mean detection times.

As shown in Table 5.2, for each experimental configuration (out of the total 27 configurations), there are 96 outbreak scenarios which were created by overlaying the simulated outbreak cases onto a background dataset with randomly selected outbreak onset dates. Thus, for a given configuration I assume that the 96 scenarios are independent. However, the total 96 × 27 scenarios over all 27 configurations are not independent. For example, in Table 5.2, outbreak scenario 1 for the 1st experimental configuration and outbreak scenario 1 for the 2nd experimental configuration are highly correlated in terms of the background data, the number of outbreak cases (created by using FLOO(1,10)) that were overlaid onto the background cases, and the disease (*cryptosporidiosis*) that causes the ongoing outbreak. Thus, the 27 outbreak results in any given row in Table 5.2 are correlated due to overlapping experimental configurations. The results are also correlated due to using the same background data in generating the scenarios in that row. The data shown in Table 5.2 suggests a hierarchical analysis that is clustered according to the factor *outbreak scenario*. Each row in Table 5.2 can be considered as a *group* categorized using the same outbreak scenario. An analysis that is based on individual observations without taking this clustering into account is likely to over-estimate the statistical significance of any observed effect. Thus, I did not apply traditional MANOVA analysis to test the proposed hypothesis. I instead adopted a linear mixed effects model (Davidian 2007) that takes into account (1) the

[13] A statistical analysis was performed on the results obtained in form shown in Table 5.3 using an analogous procedure as did for Table 5.2, and the results of statistical analysis are presented at the end of this section.

variation between and within groups and (2) the correlations between factors FLOO($\Delta$,$T$), $d_u$, and *symptom*.

I assigned the three factors (FLOO($\Delta$,$T$), $d_u$, and *symptom*) as fixed effects. The random effect in this application is the factor *outbreak scenario,* because the outbreak scenarios under a specific experimental configuration (such as a configuration of FLOO(1,10), $d_u$ = *cryptosporidiosis*, and *symptom* = *abdominal pain*) are assumed to be independently created using randomly generated outbreak onset dates.

After fitting the data in Table 5.2 using a linear mixed effects model, I further examined the fitted model and made certain that the underlying distributional assumptions appear valid for the data. According to (Pinheiro 2000), we should examine the fitted model both graphically and numerically.

The plot of the standardized residuals versus the fitted values from the model, shown in Figure 5.9, does not indicate a violation of the assumption of constant variance. In addition, the normal plot in Figure 5.10 indicates that the assumption of normality for the within-group errors is plausible.

I then examined a set of confidence intervals on the model parameters using the intervals function in R. Since this dissertation does not focus on the linear mixed effects model per se, I just present a summary of the analysis using it. Details regarding how to interpret the results are in (Pinheiro 2000). A summary of the confidence intervals on the parameters of the fitted model is presented as below, which shows that the parameters were estimated relatively precisely.

```
> intervals(mydata.lme)
Approximate 95% confidence intervals

 Fixed effects:
                              lower      est.      upper
(Intercept)               9.06640249  9.40235983  9.73831716
```

```
FLOOT14D2               -0.48473402 -0.40711806 -0.32950209
FLOOT14D4               -2.55022399 -2.47260802 -2.39499206
symptomcough             0.05659594  0.14621914  0.23584233
symptomheadache         -0.28046681 -0.19084362 -0.10122043
duearly_anthrax         -0.77996936 -0.70235340 -0.62473743
duinhalation_tularemia -2.31748865 -2.23987269 -2.16225672
dvearly_anthrax         -0.33697167 -0.25935571 -0.18173975
dvinhalation_tularemia -1.18322939 -1.10561343 -1.02799746
modelPDM                -0.12227260 -0.04465664  0.03295933
modelUDM                -0.13375022 -0.05613426  0.02148170
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: scenario
                          lower     est.    upper
sd((Intercept))    1.376677 1.587236 1.829999

 Within-group standard error:
        lower     est.    upper
 2.828784 2.851140 2.873672
```

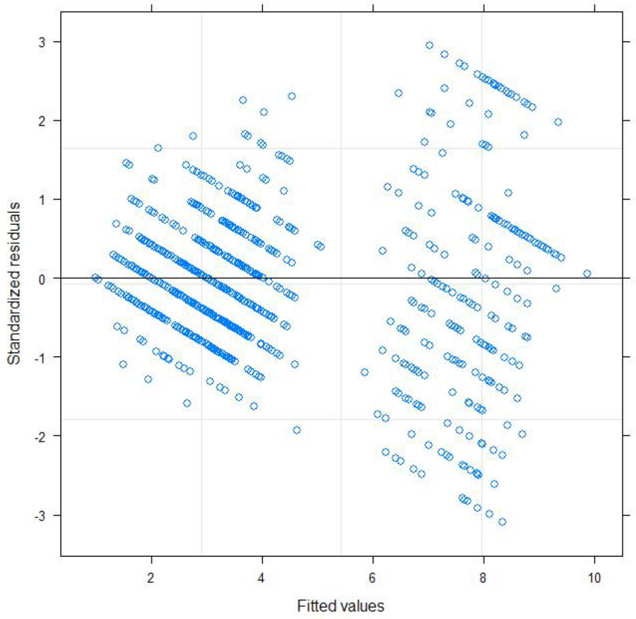


**Figure 5.9** Scatter plot of the standardized within-group residuals versus the within-group fitted values for the fitted model.

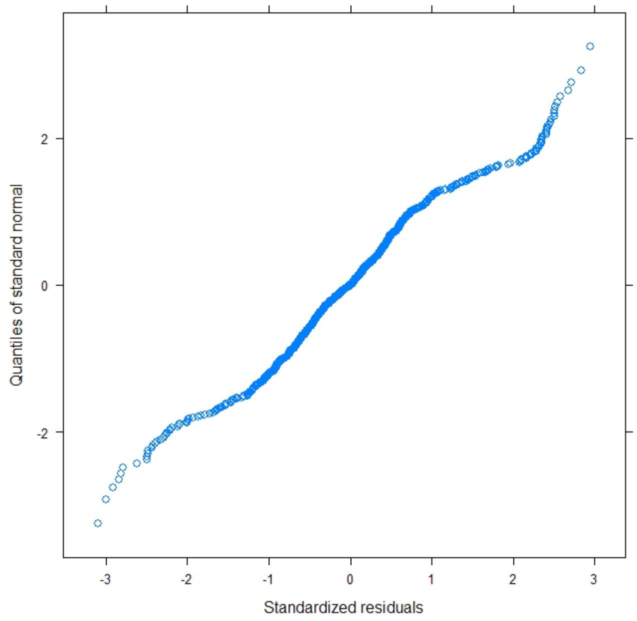


**Figure 5.10** Normal plot of within-group standardized residuals for the fitted model.

I then performed pairwise comparisons to look for inequalities between pairs of models. I applied Tukey's method (Toothaker 1993) to adjust for multiple comparisons. Table 5.4 shows the results of pairwise comparisons using the outputs from the UBH algorithm. It contains *p*-values of comparing DSM, UDM, and PDM in a pairwise fashion over all 96 × 27 experiments in Exp. 1 and over all 96 × 54 experiments in Exp. 2. The results support that DSM is at least as good as UDM and PDM in Exp. 1. A one-sided test was used for comparing DSM vs. UDM and DSM vs. PDM. Since the direction of the difference between PDM and UDM is not clear, I used a two-sided test for comparing them. Similarly, one-sided tests were used for UDM vs. DSM and PDM vs. DSM in Exp. 2, and a two-sided test was used for PDM vs. UDM in Exp. 2. All these tests used a significant level of 0.05.

Table 5.4 shows the *p*-values of pairwise comparisons of DSM, UDM, and PDM for the UBH algorithm. As shown in this table, UDM and DSM are statistically significantly different in disease detection performance in all circumstances; PDM and DSM are statistically significantly different in circumstances except in Exp. 2 at zero false alerts per month, in which the *p*-value is close to being statistically significantly different; PDM and UDM had no significant difference in the performance of disease outbreak detection using two-sided tests.

**Table 5.4** *p*-values of three pairwise comparisons at zero and one false alert per month, respectively, when using the UBH algorithm.

| | Number of false alerts per month | Pairwise comparison | | |
|---|---|---|---|---|
| | | $H_o: E_{UDM} = E_{DSM}$ | $H_o: E_{PDM} = E_{DSM}$ | $H_o: E_{PDM} = E_{UDM}$ |
| | | $H_a: E_{UDM} > E_{DSM}$ | $H_a: E_{PDM} > E_{DSM}$ | $H_a: E_{PDM} \neq E_{UDM}$ |
| Exp. 1 | 0 | 0.030 | 0.036 | 0.044 |
| | 1 | 0.041 | 0.042 | 0.052 |
| | | $H_a: E_{UDM} < E_{DSM}$ | $H_a: E_{PDM} < E_{DSM}$ | $H_a: E_{PDM} \neq E_{UDM}$ |
| Exp. 2 | 0 | 0.025 | 0.054 | 0.047 |
| | 1 | 0.019 | 0.020 | 0.034 |

**Table 5.5** *p*-values of three pairwise comparisons at zero and one false alert per month, respectively, from using the MBH algorithm.

| | Number of false alerts per month | Pairwise comparison | | |
|---|---|---|---|---|
| | | $H_o: E_{UDM} = E_{DSM}$ | $H_o: E_{PDM} = E_{DSM}$ | $H_o: E_{PDM} = E_{UDM}$ |
| | | $H_a: E_{UDM} > E_{DSM}$ | $H_a: E_{PDM} > E_{DSM}$ | $H_a: E_{PDM} \neq E_{UDM}$ |
| Exp. 1 | 0 | 0.021 | 0.033 | 0.058 |
| | 1 | 0.032 | 0.031 | 0.022 |
| | | $H_a: E_{UDM} < E_{DSM}$ | $H_a: E_{PDM} < E_{DSM}$ | $H_a: E_{PDM} \neq E_{UDM}$ |
| Exp. 2 | 0 | 0.012 | 0.016 | 0.052 |
| | 1 | 0.008 | 0.011 | 0.048 |

Table 5.5 shows the *p*-values of pairwise comparisons of DSM, UDM, and PDM for the MBH algorithm. As shown in this table, DSM performed significantly differently from UDM and PDM in all circumstances. As for the detection performance of UDM vs. PDM, PDM and UDM had significantly different detection performance in Exp. 1 at one false alert per month and had no significantly different detection performance in other circumstances.

## 5.7 DECISION ANALYSIS

As described in Section 5.6, if in Exp. 2 $E_{PDM2} < E_{UDM2} < E_{DSM2}$ holds, then there must exists a $q < 1$ such that $E_{PDM} < E_{UDM} < E_{DSM}$ holds. This section describes how to determine values of $q$ (for $q < 1$) such that $E_{PDM} < E_{UDM} < E_{DSM}$ using the UBH algorithm. The same decision analysis was also performed on the MBH algorithm using an analogous procedure, and the results of both algorithms are presented in this section[14].

Recall that in Exp. 1 there are 3 (FLOO setting possibilities) × 3 (disease possibilities for $d_u$) × 3 (symptom possibilities) = 27 experimental configurations and for each configuration there are 96 outbreak scenarios created, thus there are a total of 96 × 27 outbreak scenarios. I assume that each of the 96 × 27 outbreak scenarios in Exp. 1 is equally likely to occur. If so, we can calculate $E_{DSM1}$ as the average detection time over the 96 × 27 outbreak scenarios for DSM1; $E_{UDM1}$ and $E_{PDM1}$ can each be calculated analogously. Similarly, in Exp. 2, there are a total of 96 × 54 outbreak scenarios. Thus $E_{DSM2}$ can be calculated as the average detection time over the 96 × 54 outbreak scenarios for DSM2, and $E_{UDM2}$ and $E_{PDM2}$ can each be calculated analogously.

[14] Section 5.8 contains a sensitivity analysis over the prior probability of the unknown (partially-known) disease being modeled and a related decision analysis is presented in that section as well.

Using this procedure, I obtained the mean detection time (in days) of all six models over all the experiments at a false alert rate of zero and one per month, as shown in Table 5.6.

In Exp. 1, DSM has a slightly better expected detection performance than UDM and PDM at both zero and one false alert per month, which is qualitatively expected. In Exp. 2, UDM detects the ongoing outbreak 0.76 days faster than DSM, and PDM detects the ongoing outbreak 0.57 days faster than DSM at one false alert per month as expected. At zero false alerts per month in Exp. 2, UDM and PDM also have a slightly better disease outbreak detection performance than DSM.

**Table 5.6** Mean detection time (in days) at a false alert rate of zero and one per month of all six models over all the experiments when using the UBH algorithm

| | Number of false alerts per month | **DSM** | **UDM** | **PDM** |
|---|---|---|---|---|
| Exp. 1 | 0 | 9.70 | 9.99 | 9.90 |
| | 1 | 6.05 | 6.18 | 6.18 |
| Exp. 2 | 0 | 10.59 | 10.30 | 10.37 |
| | 1 | 7.14 | 6.38 | 6.57 |

Using a similar procedure as described above, I also obtained the mean detection time (in days) at a false alert rate of zero and one per month from using the MBH algorithm. Table 5.7 summarizes those results. At zero false alerts per month, DSM detects the ongoing outbreak 0.75 days faster than UDM and 0.68 days faster than PDM in Exp. 1. In Exp. 2, UDM detects the outbreak 1.21 days faster than DSM, and PDM detects the outbreak 1.08 days faster than DSM; at one false alert per month in Exp. 2, the net gain in the expected detection time of UDM relative to DSM is 1.36 days and of PDM relative to DSM is 1.16 days.

**Table 5.7** Mean detection time (in days) at a false alert rate of zero and one per month of all six models over all the experiments when using the MBH algorithm

| | Number of false alerts per month | **DSM** | **UDM** | **PDM** |
|---|---|---|---|---|
| Exp. 1 | 0 | 5.62 | 6.37 | 6.30 |
| | 1 | 3.06 | 3.25 | 3.11 |
| Exp. 2 | 0 | 7.73 | 6.52 | 6.65 |
| | 1 | 4.91 | 3.55 | 3.75 |

From the results shown in Table 5.6 and 5.7, we can see that PDM performs better than UDM in Exp. 1 of both UBH and MBH algorithms but performs worse than UDM in Exp. 2 of both algorithms. This makes sense because the partially-known disease $d_{*p}$ in PDM in Exp. 1 is modeled using a mixture of priors that includes the known outbreak disease $d_u$ that causes the ongoing disease outbreak, whereas the unknown disease $d_*$ in UDM in Exp. 1 is just modeled using a uniform prior distribution. In contrast, disease $d_u$ is not known to us in Exp. 2, and thus $d_{*p}$ in PDM in Exp. 2 is modeled using a mixture of priors that does not include disease $d_u$. I conjecture that in Exp. 2, the uniform prior employed by UDM might be better in modeling the outbreak data that were constructed by simulating from disease $d_u$ than the mixture of priors employed by PDM.

We can determine the value range of $q$ such that modeling an unknown disease $d_*$ (using UDM) or a partially-known disease $d_{*p}$ (using PDM) yields an expected decrease in detection time. I first do so for a false alert rate of one per month for the UBH algorithm, and thus, we will use the results in Table 5.6 that relates to one false alert per month. Based on deriving such an estimate of $q$, we can then determine whether to construct a DSM or an UDM model in a detection system. Figure 5.11 shows such a decision analysis. Let $q_*$ be the probability such that the equation below holds:

$$(1 - q)E_{UDM1} + qE_{UDM2} = (1 - q)E_{DSM1} + qE_{DSM2} \tag{5.3}$$

Then $q_*$ is the threshold such that any probability greater than $q_*$ renders modeling $d_*$ helpful, given the conditions and assumptions of the evaluation. Solving Equation 5.3 using the values in Tables 5.6, yields $q_* = 0.15$. If $q = 0.15$ then modeling $d_*$ is expected to be neither helpful nor harmful. Moreover, if $q > 0.15$, then including $d_*$ in the model is expected to decrease the detection time at a false alert rate of one per month.

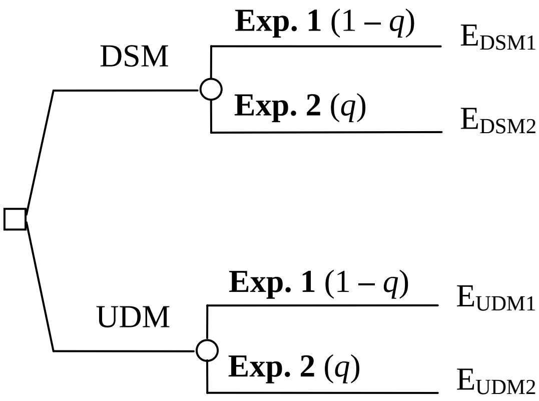


**Figure 5.11** A decision tree showing the decision analysis on selecting DSM vs. UDM

I also assessed the uncertainty for estimating $q$ by calculating its standard error. I performed random sampling of scenarios with replacement and obtained 100 scenarios at a time from the total 96 × 27 scenarios in Exp. 1 and 100 scenarios at a time from the total 96 × 54 scenarios in Exp. 2. I repeated this sampling procedure 1000 times. For each sampled 100 scenarios in Exp. 1 and Exp. 2, I calculated $E_{DSM1}$, $E_{DSM2}$, $E_{UDM1}$, and $E_{UDM2}$, and then calculated a $q$ value that satisfies Equation 5.3. I thus obtained 1000 sets of $q$ values. Let $q_*^{(s)}$ be a specific

$q$ value that was obtained out of the total 1000, where $1 \leq s \leq 1000$. I estimated the standard error (SE) of $q$ using

$$SE(q) = \sqrt{\frac{\sum_{s=1}^{1000}\left(q_*^{(s)} - q_*\right)^2}{1000-1}} ,$$

where $q_*$ = 0.15 as described above.

I also applied the same procedure, as described above, for the mean detection times of DSM1, UDM1, DSM2, and UDM2 that were obtained from using the UBH algorithm at zero false alerts per month. Similarly, we can apply this procedure for the mean detection times of DSM1, PDM1, DSM2, and PDM2 to determine a value of $q_{*p}$ with which modeling $d_{*p}$ is neither helpful nor harmful. Table 5.8 summarizes the results of $q_*$, $q_{*p}$, and their standard errors at a false alert rate of zero and one per month from using the UBH algorithm and the MBH algorithm, respectively.

**Table 5.8** A summary of the decision analysis results

| Algorithm | Number of false alerts per month | $q_*$<br>SE($q_*$) | $q_{*p}$<br>SE($q_{*p}$) |
|---|---|---|---|
| UBH | 0 | 0.49<br>0.19 | 0.48<br>0.18 |
| | 1 | 0.15<br>0.11 | 0.19<br>0.13 |
| MBH | 0 | 0.38<br>0.13 | 0.39<br>0.12 |
| | 1 | 0.12<br>0.10 | 0.039<br>0.08 |

Under the assumptions introduced, this result indicates that if the probability is greater than 0.19 of an outbreak being due to a partially-known disease $d_{*p}$, then including $d_{*p}$ in the model of the UBH algorithm is expected to decrease the detection time at a false alert rate of one per month. As discussed above, if the probability is greater than 0.15 of an outbreak being due to an unknown disease, then including $d_*$ in the model of the UBH algorithm is expected to decrease the detection time at one false alert per month. Table 5.8 also shows analogous results when the false alert rate is zero when using the UBH algorithm. In that case, the probability of an unknown disease is about 0.49 and the probability of a partially-known disease is about 0.48, which is higher than for an alert rate of one.

As for the results of the MBH algorithm, the probabilities of an unknown disease and a partially-known disease are all smaller than those of the UBH algorithm at both zero false alerts per month and one false alert per month, which supports that the hybrid detection system constructed in MBH performs better than UBH in detecting new disease outbreaks. The lowest probability 0.039 was obtained from using the MBH algorithm that includes a PDM model, and a probability 0.12 was obtained from using the MBH algorithm that includes a UDM model.

It seems plausible that there are disease-outbreak monitoring situations in which if there is an outbreak then the expected probability exceeds 0.12 of it being due to an unknown disease and 0.039 if it being due to an partially-known 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, modeling an unknown disease $d_*$ or a partially-known disease $d_{*p}$ could be beneficial.

## 5.8 SENSITIVITY ANALYSIS

Recall that Section 5.7 only reports experimental results for disease-specific model (DSM), unknown-disease model (UDM), and partially-known disease model (PDM) of the UBH and

MBH algorithm when using a uniform prior over the appearance of the outbreak diseases being modeled. Let $w$ represent the prior probability of an outbreak due to an unknown disease $d_*$ or a partially-known disease $d_{*p}$ given that an outbreak is occurring, namely $w = P(d_* \mid OB)$ or $P(d_{*p} \mid OB)$. This section describes a sensitivity analysis regarding $w$, for which a sequence of probability values 0.01, 0.02, … 0.09, 0.1, 0.2, …, 0.9, 0.91, 0.92, …, 0.99 was used for UDM and PDM, where the prior probability $P(d_* \mid OB)$ and $P(d_{*p} \mid OB)$ will be applied to UDM and PDM, respectively, to obtain the likelihood ratio results.

For each prior probability value $w$ in the sequence described above, I applied the same experimental methods described in Section 5.3 to obtain the mean detection time (over all the experiments) at zero and one false alert per month for the UBH algorithm and the MBH algorithm that includes a DSM model, an UDM model, and a PDM model, respectively. For example, when $w = 0.1$, by running the UBH algorithm, we obtain experimental results for UDM in Exp. 1 from using prior probability $P(d_* \mid OB) = 0.1$ and $P(d_u \mid OB) = 0.9$, and for PDM in Exp. 1 from using prior probability $P(d_{*p} \mid OB) = 0.1$ and $P(d_u \mid OB) = 0.9$. Note that when $w = 0$, DSM, UDM, and PDM have the same mean detection time in Exp. 1 and Exp. 2. Section 5.8.1 and 5.8.2 describe the experimental results of the sensitivity analysis using the UBH algorithm and the MBH algorithm, respectively. Each section contains experimental results that were obtained at a false alert rate of zero and one per month.

### 5.8.1 Results of the UBH algorithm

Figure 5.12 shows the mean detection time of DSM1, UDM1, and PDM1 (in Exp. 1) at one false alert per month when using a sequence of prior probability values $w$. Note that DSM does not model disease $d_*$ or $d_{*p}$, thus the mean detection time of DSM does not change relative to $w$ ($w$

= $P(d_* \mid OB)$ or $P(d_{*p} \mid OB)$), as shown in Figure 5.12 as a horizontal line at $\mathrm{E}_{\mathrm{DSM1}}$ = 6.05 days, and DSM1 has a faster mean detection time than UDM and PDM at every value of *w*. UDM1 and PDM 1 both have the fastest mean detection time at a prior probability *w* = 0.01 with $\mathrm{E}_{\mathrm{UDM1}}$ = $\mathrm{E}_{\mathrm{PDM1}}$ = 6.06 days. The overall trend of UDM1 is that the mean detection time of UDM1 increases relative to *w*.

Figure 5.13 shows the mean detection time of DSM2, UDM2, and PDM2 (in Exp. 2) at one false alert per month when using a sequence of prior probability values *w*. UDM2 and PDM2 outperform DSM2 at any probability value *w* in the sequence, as expected. The mean detection time of UDM2 decreases as *w* increases, which implies that the higher the prior probability assigned to disease $d_*$, the better the mean detection performance of UDM2 in detecting the ongoing disease outbreak that is caused by an unknown disease. PDM2 also has a similar trend as UDM2, as expected. However, PDM2 performs no better than UDM2 at any value of *w*.

Figure 5.14 shows the decision analysis results ($q_*$ and $q_{*p}$) relative to *w* at one false alert per month for the UBH algorithm. For example, when *w* = 0.5, $q_*$ is 0.15, as shown in this figure. This example indicates that given a prior probability of 0.5 for $d_*$ (conditioned on there being an outbreak), modeling $d_*$ will decrease the expected outbreak-detection time if the actual frequency of $d_*$ (conditioned on there being an outbreak) is a 0.15 or greater. As shown in this figure, the maximum value of $q_*$ is 0.18 at *w* = 0.01. When *w* = 0.6, $q_{*p}$ also has the maximum value of 0.22. Probability $q_*$ and $q_{*p}$ have the minimum value of 0.08 at *w* = 0.08 and *w* = 0.2, respectively.

Figure 5.15 shows the mean detection time of DSM1, UDM1, and PDM1 (in Exp. 1) at zero false alerts per month when using a sequence of prior probability values *w*. It is expected that DSM1 detects the known ongoing disease outbreak faster than UDM1 and PDM1. However, when the prior probability $w$ falls into the range such that $0.01 \leq w \leq 0.09$ for UDM1 and $0.03 \leq w \leq 0.1$ for PDM1, UDM1 and PDM1 have slightly faster mean detection time than DSM1 as shown in this figure, with the maximum mean detection time gain of UDM1 over DSM1 as 0.02 days and PDM1 over DSM1 as 0.02 days. The result using Tukey's method shows that there is no statistically significantly difference in the mean detection time of DSM1, UDM1, and PDM1 when $w$ is in these ranges. Thus, I conjectured that the slight net gain in mean detection time of UDM1 (or PDM1) relative to DSM1 is due to noise. When $w > 0.1$, the results in this figure show that DSM1 outperforms UDM1 and PDM1, and the mean detection time of UDM1 and PDM1 increases, as $w$ increases.

Figure 5.16 shows the mean detection time of DSM2, UDM2, and PDM2 (in Exp. 2) at zero false alerts per month when using a sequence of prior probability values $w$. UDM2 and PDM2 outperforms DSM2 at any value of $w$ in the sequence, as expected. As $w$ increases, the mean detection times of UDM2 and PDM2 decrease, as expected. PDM2 has a mean detection time that is no worse than UDM2.

Figure 5.17 shows the decision analysis results ($q_*$ and $q_{*p}$) relative to $w$ at zero false alerts per month using the UBH algorithm. Since UDM1 and PDM1 have faster mean times than DSM1 when $0.01 \leq w \leq 0.09$ for UDM1 and $0.03 \leq w \leq 0.1$ for PDM1 (as shown in Figure 5.15), and UDM2 and PDM2 outperform DSM2 (as shown in Figure 5.16), it is not surprising that the value of $q_*$ and $q_{*p}$ is smaller than zero at this range of $w$ using Equation 5.3. In this circumstance, I set $q_*$ and $q_{*p}$ to be zero in the plot of $q_*$ and $q_{*p}$ relative to $w$. Recall that $q$ is used to represent the actual frequency with which the Exp. 2 scenario occurs. Under the assumptions introduced, this result indicates that including $d_*$ (or $d_{*p}$) in the model will decrease the detection time at a false alert rate of zero alerts per month when $d_*$ (or $d_{*p}$) occurs at a frequency $q$ for $0.01 \le q \le 0.09$ for UDM1 and $0.03 \le q \le 0.1$ for PDM1. When $w \ge 0.02$, the value of $q_*$ and $q_{*p}$ increases as $w$ increases.

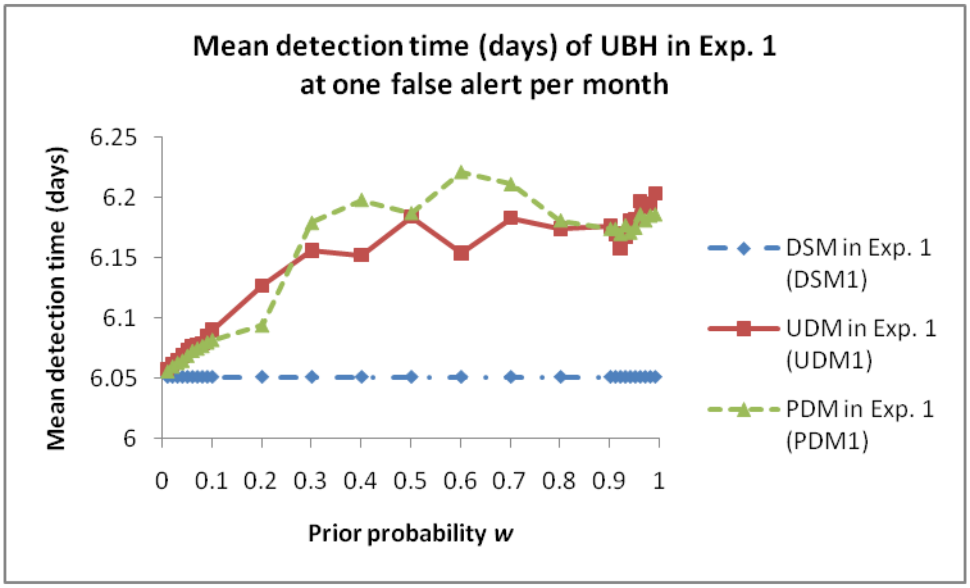


**Figure 5.12** Mean detection time (days) at one false alert per month for DSM, UDM, and PDM in Exp. 1 using the UBH algorithm

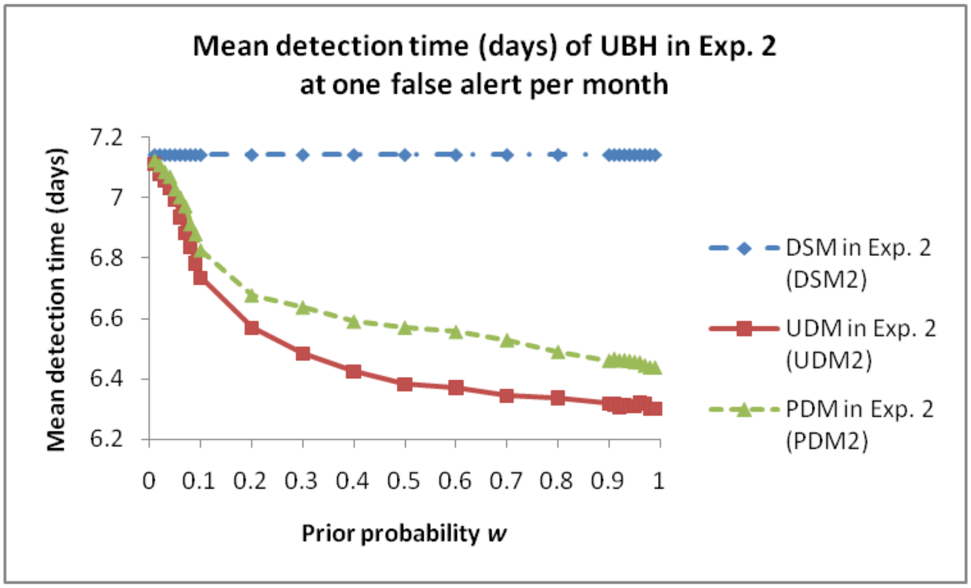


**Figure 5.13** Mean detection time (days) at one false alert per month for DSM, UDM, and PDM in Exp. 2 using the UBH algorithm

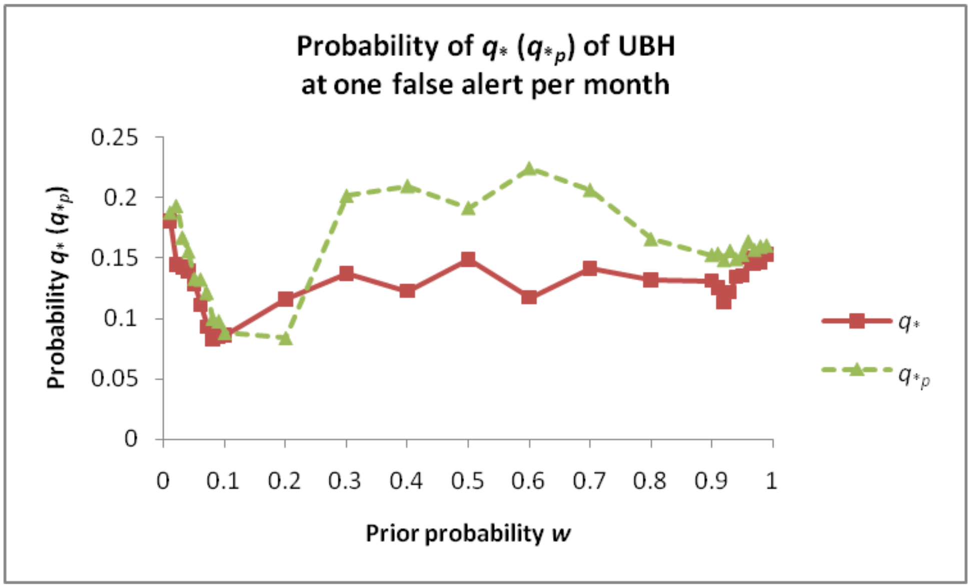


**Figure 5.14** Probability value $q_*$ and $q_{*p}$ relative to the prior probability of the appearance of the unknown (partially-known) disease at one false alert per month using the UBH algorithm

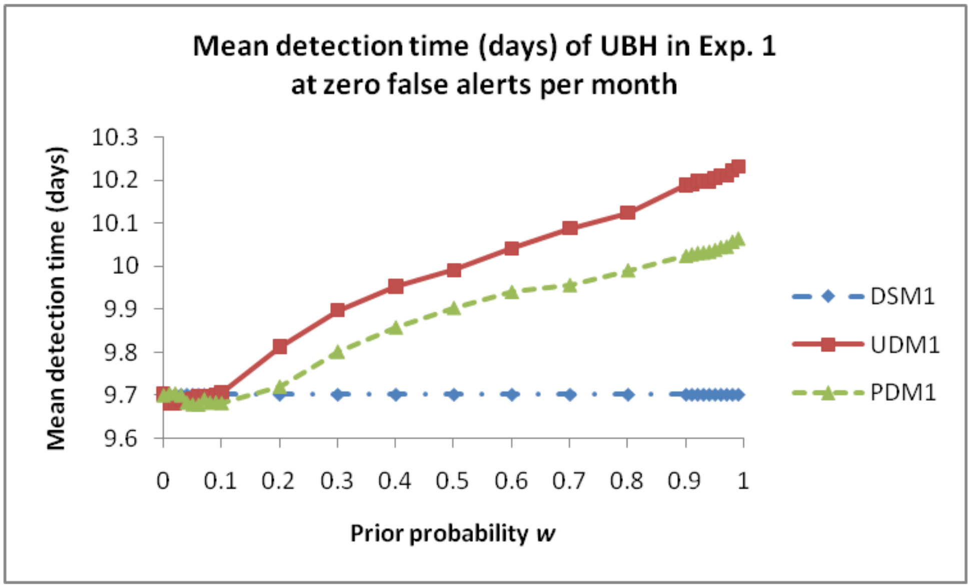


**Figure 5.15** Mean detection time (days) at zero false alerts per month for DSM, UDM, and PDM in Exp. 1 using the UBH algorithm

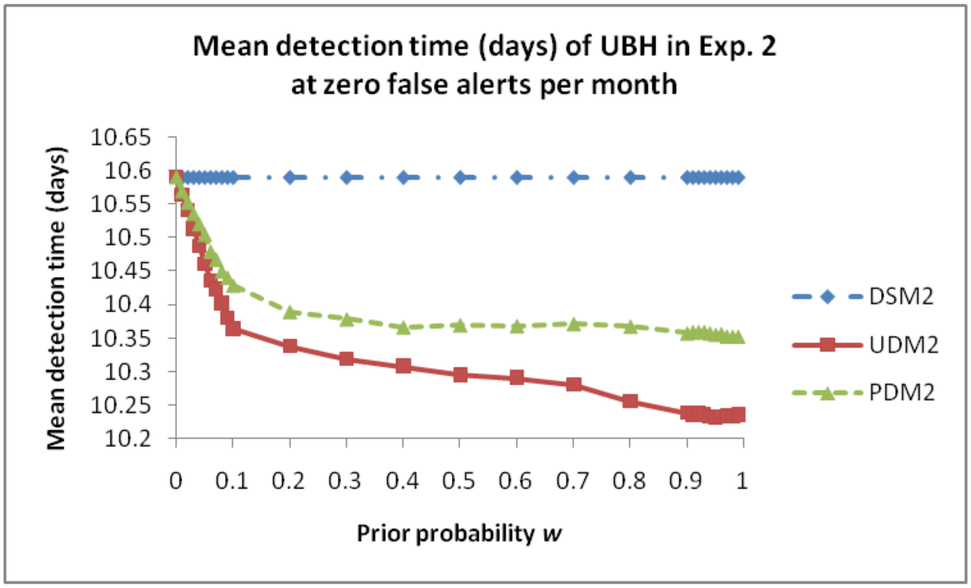


**Figure 5.16** Mean detection time (days) at zero false alerts per month for DSM, UDM, and PDM in Exp. 2 using the UBH algorithm

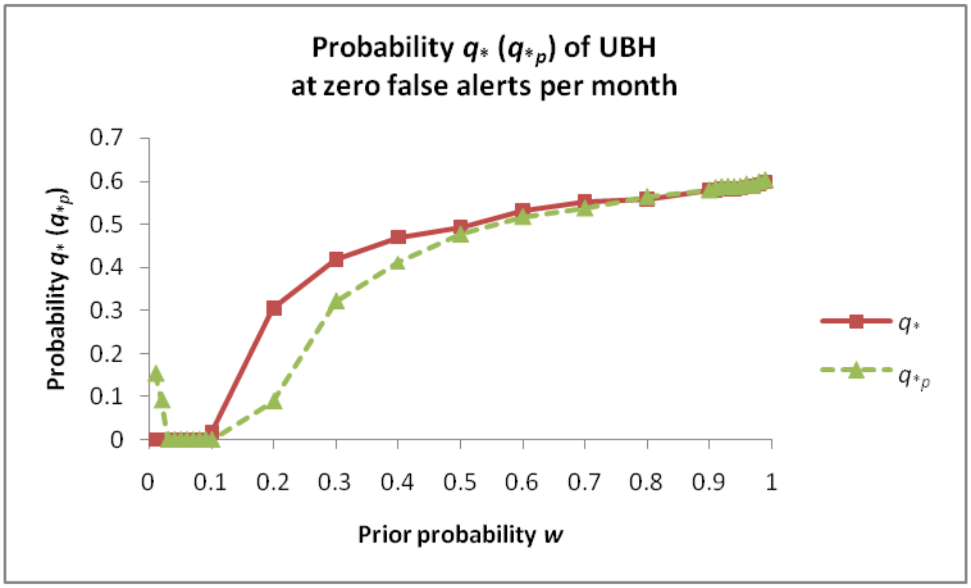


**Figure 5.17** Probability value $q_*$ and $q_{*p}$ relative to the prior probability of the appearance of the unknown (partially-known) disease at zero false alerts per month using the UBH algorithm

### 5.8.2 Results of the MBH algorithm

Figure 5.18 shows the mean detection time of DSM1, UDM1, and PDM1 (in Exp. 1) at one false alert per month when using a sequence of prior probability values *w*. The mean detection of DSM1 is approximately 3.06 days, and DSM1 has a faster mean detection time than UDM1 and PDM1 at $w \geq 0.08$ and $w \geq 0.4$, respectively. UDM1 has the fastest mean detection time at a prior probability $w = 0.01$ with $E_{UDM1} = 3.04$ days, and PDM1 has the fastest mean detection time at $w = 0.3$ with $E_{PDM1} = 3.04$ days. When $w = 0.99$, both UDM1 and PDM1 have the slowest mean detection time of 3.55 and 3.18 days, respectively, which are 2.51 and 2.88 days faster than the fastest mean detection time of UDM1 and PDM1 that were obtained using the UBH algorithm, as shown in Figure 5.12.

Figure 5.19 shows the mean detection time of DSM2, UDM2, and PDM2 (in Exp. 2) at one false alert per month when using a sequence of prior probability values *w*. UDM2 and PDM2 outperform DSM2 at any probability value *w* in the sequence, as expected. UDM2 performs slightly better than PDM2 at a probability range of $0.2 \leq w \leq 0.5$. The MBH algorithm has a better mean detection performance than the UBH algorithm (shown in Figure 5.13) regarding the absolute magnitude of the mean detection time of DSM2, UDM2, and PDM2, and the net profit gain of UDM2 over DSM2 and PDM2 over DSM2.

Figure 5.20 shows the decision analysis results ($q_*$ and $q_{*p}$) relative to *w* at one false alert per month using the MBH algorithm. When $0.01 \leq w \leq 0.07$, the value of $q_*$ that is calculated using Equation 5.3 is negative due to $E_{DSM1} > E_{UDM1}$ and $E_{DSM2} > E_{UDM2}$. I set the value of $q_*$ to be zero at this range of *w*, as shown in this figure. It indicates that when our prior belief about an unknown-disease outbreak is smaller than 0.07, modeling $d_*$ will definitely be helpful no matter what the actual frequency of an unknown-disease outbreak is. This is because that UDM surprisingly had better detection performance than DSM in Exp. 1 while it is expected DSM outperformed UDM in this experiment; when *w* = 0.08, 0.09, 0.1, and 0.3, $q_{*p}$ also has been set to zero as described above, which indicates that including disease $d_{*p}$ will improve the disease detection performance when the actual frequency of the appearance of $d_{*p}$ falls into this range.

Figure 5.21 shows the mean detection time of DSM1, UDM1, and PDM1 (in Exp. 1) at zero false alerts per month when using a sequence of prior probability values *w*. The mean detection of DSM1 is approximately 5.62 days, and DSM1 has a faster mean detection time than UDM1 and PDM1 at $w \geq 0.05$ and $w \geq 0.3$, respectively. UDM1 has the fastest mean detection time at a prior probability *w* = 0.01 with $E_{UDM1}$ = 5.56 days, and PDM1 has the fastest mean detection time at *w* = 0.1 with $E_{PDM1}$ = 5.41 days. When *w* = 0.99, both UDM1 and PDM1 have the slowest mean detection time of 6.44 and 6.48 days, respectively, which are 3.24 and 3.20 days faster than the fastest mean detection time of UDM1 and PDM1 that were obtained using the UBH algorithm, as shown in Figure 5.15.

Figure 5.22 shows the mean detection time of DSM2, UDM2, and PDM2 (in Exp. 2) at zero false alerts per month when using a sequence of prior probability values *w*. UDM2 and PDM2 outperform DSM2 at any probability value *w* in the sequence, as expected, and UDM2 performs slightly better than PDM2. The MBH algorithm has a better mean detection performance than the UBH algorithm (shown in Figure 5.16) regarding the absolute magnitude of the mean detection time of DSM2, UDM2, and PDM2, and the net profit gain of UDM2 over DSM2 and PDM2 over DSM2.

Figure 5.23 shows the decision analysis results ($q_*$ and $q_{*p}$) relative to *w* at zero false alerts per month when using the MBH algorithm. As shown in this figure, when $0.01 \leq w \leq 0.04$, the value of $q_*$ has been set to zero; when $0.01 \leq \; w \leq 0.2$, $q_{*p}$ has been set to zero due to the reason described above. This result also indicates that including $d_*$ will improve the disease detection performance of the detection system when $d_*$ occurs at a frequency $q$ for $0.01 \leq q \leq 0.04$. A similar conclusion holds for including disease $d_{*p}$ for $0.01 \leq q \leq 0.2$.

It is interesting to notice that the overall trend of $q_*$ relative to $w$ is that when $w$ increases, $q_*$ increases as well, such as Figure 5.20 and Figure 5.23. This trend indicates that the more prior belief we think that the ongoing outbreak is due to an unknown disease, the higher the actual frequency of the unknown-disease outbreak needs to be in order for modeling $d_*$ to be helpful. This trend seems to be contradictory to our expectation. However, our prior belief about the probability of an unknown-disease outbreak might not be realistic. In particular, in the experiments constructed in Exp. 1 and Exp. 2, there is 1/3 chance that the outbreak is being caused by an unknown disease because three outbreak diseases were selected for performing leave-one-out experiments, as described in Section 5.3.

Section 5.8.1 and 5.8.2 provides support that including unknown and partially-known diseases in the model can improve the disease detection performance of the detection system. In addition, modeling unknown and partially-known diseases will yield more net profit when using the MBH algorithm than using the UBH algorithm.

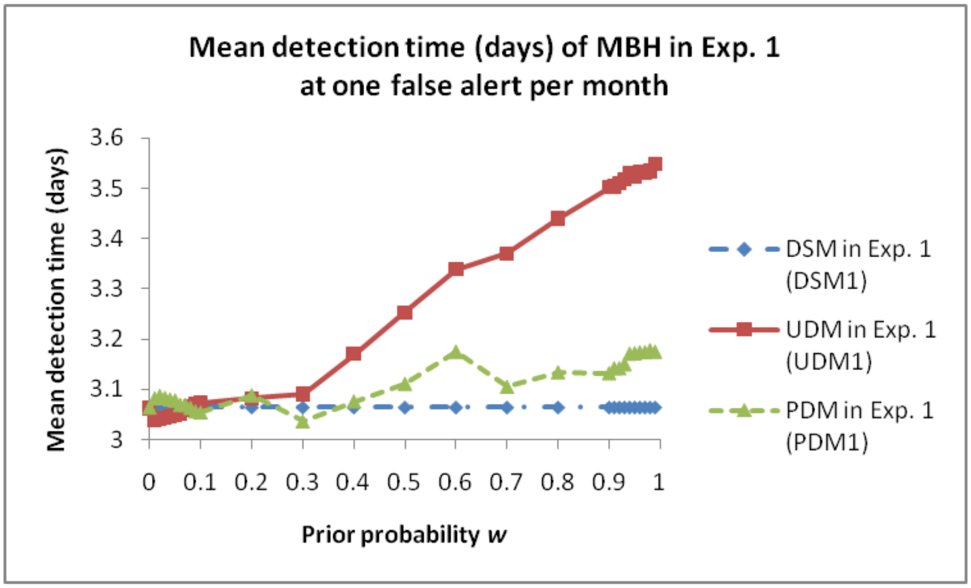


**Figure 5.18** Mean detection time (days) at one false alert per month for DSM, UDM, and PDM in Exp. 1 using the MBH algorithm

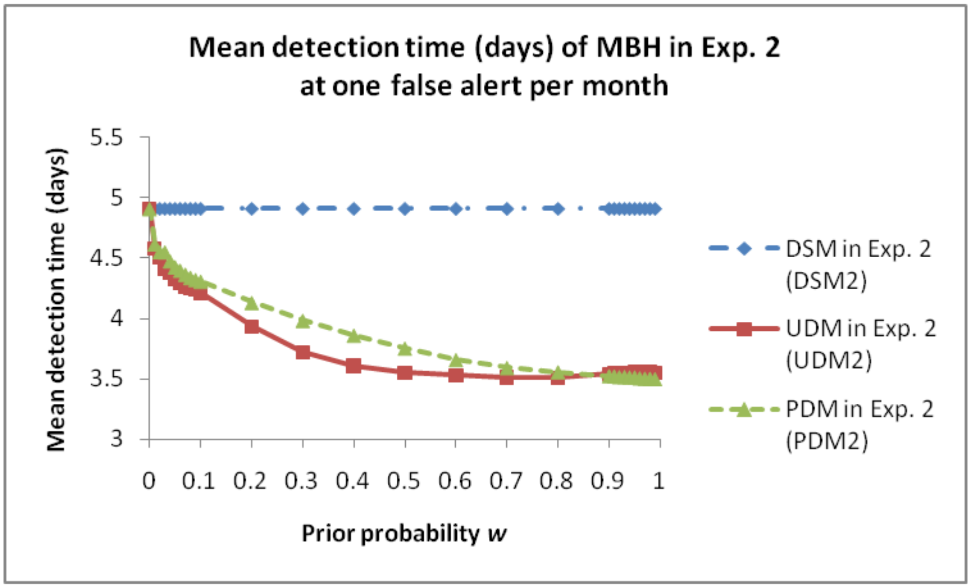


**Figure 5.19** Mean detection time (days) at one false alert per month for DSM, UDM, and PDM in Exp. 2 using the MBH algorithm

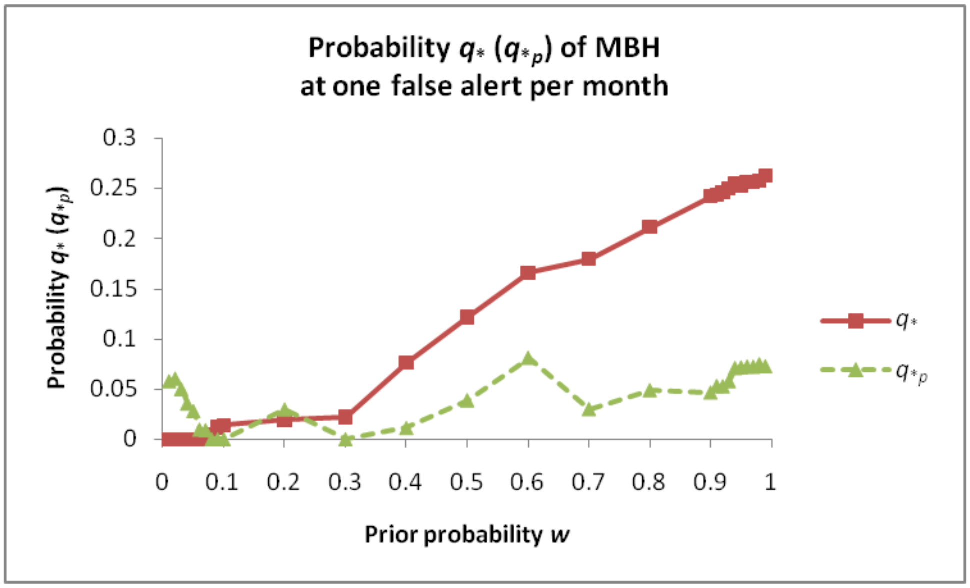


**Figure 5.20** Probability value $q_*$ and $q_{*p}$ relative to the prior probability of the appearance of the unknown (partially-known) disease at one false alert per month using the MBH algorithm

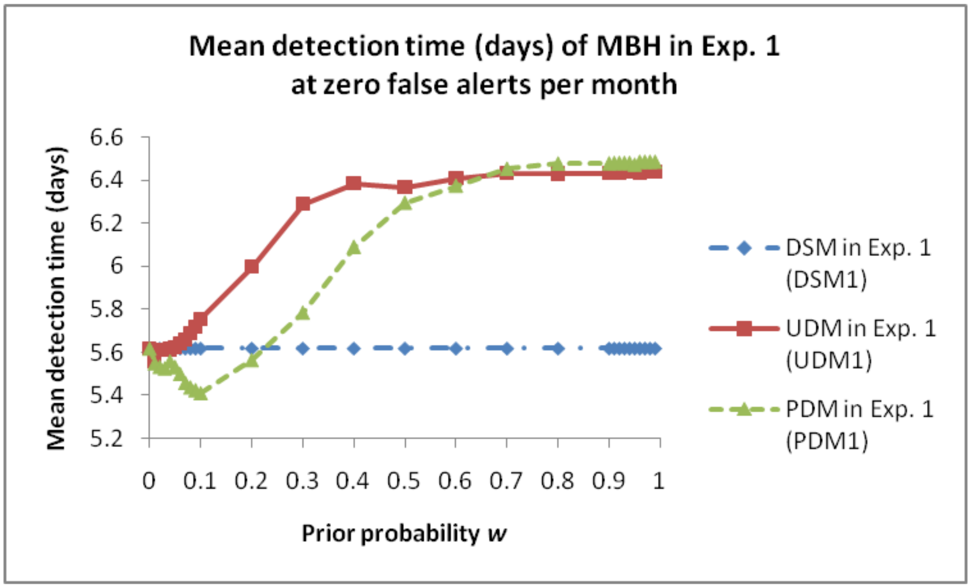


**Figure 5.21** Mean detection time (days) at zero false alerts per month for DSM, UDM, and PDM in Exp. 1 using the MBH algorithm

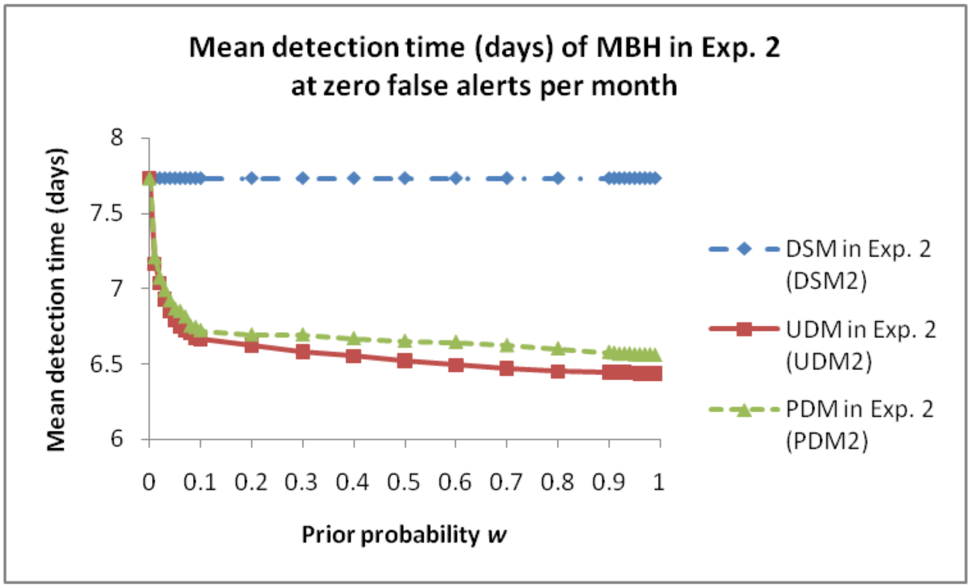


**Figure 5.22** Mean detection time (days) at zero false alerts per month for DSM, UDM, and PDM in Exp. 2 using the MBH algorithm

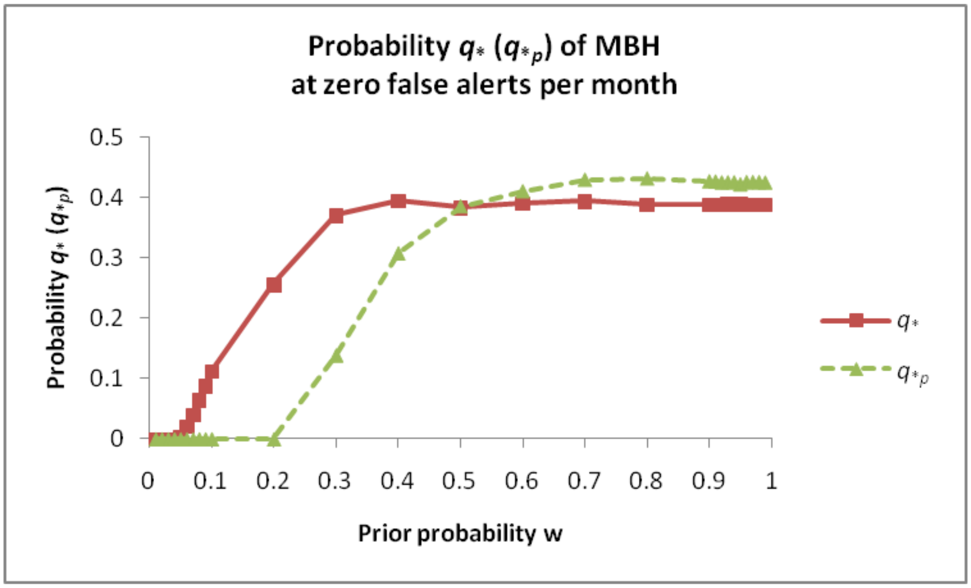


**Figure 5.23** Probability value $q_*$ and $q_{*p}$ relative to the prior probability of the appearance of the unknown (partially-known) disease at zero false alerts per month using the MBH algorithm

# 6.0 CONCLUSION AND FUTURE WORK

This dissertation investigates Bayesian modeling of unknown causes in the context of disease-outbreak detection. It introduces a Bayesian method for disease-outbreak detection that combines models of known diseases and unknown (or partially-known) diseases. In particular, I model the known non-outbreak disease $d_0$ using an informative prior estimated from past ED data, and model a known outbreak disease $d_k$ (for $k > 0$) using informative priors that were assessed from an infectious disease expert. The unknown-disease model uses a non-informative prior to model some unknown disease $d_*$. This dissertation also investigated modeling of a partially-known disease $d_{*p}$, for which a mixture of informative and non-informative priors is used.

In order to evaluate the hypothesis proposed in Chapter 1, I constructed several versions of detection systems that include a known outbreak disease $d_k$ (for $k > 0$), an unknown-disease $d_*$, and a partially-known disease $d_{*p}$, respectively.

The results presented in this dissertation provide support for the hypothesis that modeling both known and unknown outbreak diseases in a hybrid system can lead to better expected disease outbreak detection performance than modeling known outbreak diseases only.

The remainder of this chapter first summarizes the contributions of this dissertation research and then presents areas for future research.

## 6.1 CONTRIBUTIONS

### 6.1.1 Disease modeling

I investigated a continuous disease-modeling space, from disease-specific modeling to partially-known disease modeling to unknown disease modeling using the following methods:

- I used expert judgment for modeling a specific outbreak disease $d_k$, for which an informative prior distribution was derived using expert assessment on the mean and variance of the frequency of some symptom state in the people having $d_k$ in the population, where $k \geq 1$.
- A non-informative prior was proposed for modeling an unknown outbreak disease $d_*$ about which we almost completely lack knowledge. In particular, I used a uniform distribution on the interval of [0,1] to model the probability of a symptom state, as for example *cough*, given a person having disease $d_*$.
- I also proposed a mixture of informative and non-informative priors for modeling a partially-known disease $d_{*p}$. In particular, an informative prior in the mixture model represents the prior distribution of a known outbreak disease $d_k$ that we conjecture to share some similar characteristics, such as the frequency of the *cough* symptom in the population that have $d_k$, with the partially-known disease $d_{*p}$; The non-informative prior in the mixture model represents the condition that we know little about $d_{*p}$. The mixture of priors I proposed is semi-informative in the sense that it combines prior distributions that are informative and non-informative in a mixture model.

### 6.1.2 A unifying framework in a hybrid disease detection system

I used a Bayesian framework that combines models of known outbreak diseases and an unknown outbreak disease to construct a hybrid outbreak detection system. This unifying framework allows us to model any number of known and unknown diseases and also provides a way to specify arbitrary prior probabilities for models of known and unknown diseases.

## 6.2 FUTURE WORK

This section describes future work and some open problems related to the dissertation research. As mentioned, the Bayesian approach I describe for modeling unknown diseases is based on specifying non-informative priors. There are numerous ways of specifying such non-informativeness, and it would be worthwhile to explore approaches beyond just using uniform distributions. One such approach is to use 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 above some specified level), but otherwise the parameter distributions are uniform (Shen 2009). I believe the investigation of non-informative and semi-informative priors holds significant promise in artificial intelligence in general and biomedical informatics applications in particular, where causes of events may often be unknown.

Recall that the MBH algorithm models the binary state of every evidential feature, as for example, cough vs. no cough, and headache vs. no headache, by assuming the evidential features are conditionally independent given the disease state of an individual in the population. Assuming independence between evidential features makes it easier to convey the basic approach in this dissertation. However, modeling independence among evidential features may ignore possible dependencies between them, and thus may affect the likelihood ratio output of the MBH algorithm. We could model dependent symptoms using Dirichlet-multinomial hierarchical model. However, it may be difficult to assess from experts the parameters of the Dirichlet distribution. Note that the inference method employed in this dissertation does not limit us from using any non-independence model because I used a simulation method to perform inference on the Bayesian network that models multivariate independent evidential features, as described in Section 4.3.3.

Besides detecting possible disease outbreaks, it would be useful for the BH algorithm to characterize the ongoing disease outbreak, as for example by computing the posterior probability *P*(*person_i_cough_state* = *present* | *outbreak_disease_in_population* = $d_*$, *evidence*). Such an analysis gives the user some insight into how unknown disease $d_*$ is presenting. To compute this posterior probability, we can perform inference on the Bayesian network model shown in Figure 4.1.

Recall that the BH algorithm was evaluated on the simulated outbreak scenarios, in which the number of outbreak cases was generated using the linear FLOO simulator, and the simulated symptom state of each patient case was generated by sampling from the Beta-Binomial model. The sampling method itself brings random effects into the outbreak scenarios to be tested. In addition, as described in Section 5.2, the probability of a symptom state in a disease was assumed to have a Beta distribution, while the data were simulated using the Beta-Binomial model, as described above. Thus, the simulated data contains another level of random effects. In the future, it would still be useful to evaluate the BH algorithm on datasets that are generated using simulators other than FLOO and that also contain additional sources of noise.

It would also be useful to evaluate the BH algorithm using real data, as for example evaluating the univariate version of the BH algorithm (UBH) using real ED chief complaints. By doing this, we could compare the detection performance of the UBH algorithm with other outbreak-detection algorithm, such as PC.

In the future, we could also apply the UDM and PDM model to the spatio-temporal version of MD-PANDA (Jiang 2008) and investigate whether including $d_*$ and $d_{*p}$ can improve the disease detection performance when using a spatio-temporal model. However, we have to consider the computational issue of the spatio-temporal version of the MBH algorithm. Recall from Section 4.3.3 (for the current version of MBH) that I applied Monte Carlo integration to approximate the likelihood given by Equation 4.20, where the time complexity of Monte Carlo integration depends on the exact inference time of the non-spatial, non-temporal Bayesian network model shown in Figure 4.3. According to (Jiang 2008), a spatio-temporal version of PC (PCTS) requires inference time of approximately 5 minutes when running PCTS on ED data from the past five days. If we still use Monte Carlo integration to approximate the likelihood of the spatio-temporal version of MBH, we will need $5M$ minutes for inference, where $M$ is the number of samples. Monte Carlo integration typically uses $M > 1000$ samples, which makes the inference time of the spatio-temporal version of MBH computationally expensive. One possible direction to future research is to develop an approximate inference method for PCTS that can more efficiently compute the likelihood (or posterior probability).

# APPENDIX A

## A COMPLETE LIST OF PARAMETERS ESTIMATED FOR DISEASE MODELS

Table A.1 shows the parameters ($\alpha$, $\beta$) estimated for the non-outbreak disease model and the disease-specific models.

**Table A.1** Parameters ($\alpha$, $\beta$) estimated for the non-outbreak disease model and the disease-specific models given the non-outbreak disease, three selected outbreak diseases, and three selected disease symptoms.

| Diseases | abdominal pain | cough | headache |
|---|---|---|---|
| non-outbreak disease | (1218.5,38781.5) | (463.77,39536.23) | (974.83,39025.17) |
| cryptosporidiosis | (22.38,5.59) | (9.91,485.77) | (1.25,11.25) |
| early stage anthrax | (1.07,10.84) | (6.84,13.90) | (6.22,16.83) |
| inhalation tularemia | (8.83,3.79) | (22.38,5.59) | (5.67,1.42) |

# APPENDIX B

## A COMPLETE LIST OF THE NUMBER OF REAL EMERGENCY DEPARTMENT CASES USED IN THE EVALUATION

Recall from Section 5.2 that I obtained real ED cases for 2004 and 2005 from a large hospital in Allegheny County, Pennsylvania. Table B.1 shows a complete list of the daily number of real ED cases for 2004 and 2005 from that hospital. This list is being provided for completeness in describing the experiments that I performed.

**Table B.1** A complete list of real ED cases for 2004 and 2005 from a large hospital in Allegheny County, Pennsylvania.

| Date | Number of ED cases | Date | Number of ED cases | Date | Number of ED cases |
|---:|---:|---:|---:|---:|---:|
| 1/1/2004 | 122 | 9/1/2004 | 126 | 5/3/2005 | 141 |
| 1/2/2004 | 141 | 9/2/2004 | 160 | 5/4/2005 | 150 |
| 1/3/2004 | 151 | 9/3/2004 | 164 | 5/5/2005 | 119 |
| 1/4/2004 | 118 | 9/4/2004 | 146 | 5/6/2005 | 133 |
| 1/5/2004 | 145 | 9/5/2004 | 165 | 5/7/2005 | 135 |
| 1/6/2004 | 114 | 9/6/2004 | 142 | 5/8/2005 | 122 |
| 1/7/2004 | 127 | 9/7/2004 | 160 | 5/9/2005 | 140 |
| 1/8/2004 | 122 | 9/8/2004 | 139 | 5/10/2005 | 124 |
| 1/9/2004 | 121 | 9/9/2004 | 141 | 5/11/2005 | 150 |
| 1/10/2004 | 122 | 9/10/2004 | 127 | 5/12/2005 | 108 |
| 1/11/2004 | 124 | 9/11/2004 | 155 | 5/13/2005 | 147 |
| 1/12/2004 | 123 | 9/12/2004 | 137 | 5/14/2005 | 133 |
| 1/13/2004 | 119 | 9/13/2004 | 171 | 5/15/2005 | 139 |
| 1/14/2004 | 112 | 9/14/2004 | 158 | 5/16/2005 | 150 |

| | |
|---|---|
| 1/15/2004 | 117 |
| 1/16/2004 | 107 |
| 1/17/2004 | 105 |
| 1/18/2004 | 137 |
| 1/19/2004 | 153 |
| 1/20/2004 | 141 |
| 1/21/2004 | 129 |
| 1/22/2004 | 105 |
| 1/23/2004 | 109 |
| 1/24/2004 | 128 |
| 1/25/2004 | 124 |
| 1/26/2004 | 109 |
| 1/27/2004 | 126 |
| 1/28/2004 | 128 |
| 1/29/2004 | 124 |
| 1/30/2004 | 143 |
| 1/31/2004 | 127 |
| 2/1/2004 | 131 |
| 2/2/2004 | 151 |
| 2/3/2004 | 132 |
| 2/4/2004 | 129 |
| 2/5/2004 | 119 |
| 2/6/2004 | 123 |
| 2/7/2004 | 107 |
| 2/8/2004 | 129 |
| 2/9/2004 | 129 |
| 2/10/2004 | 117 |
| 2/11/2004 | 122 |
| 2/12/2004 | 114 |
| 2/13/2004 | 136 |
| 2/14/2004 | 152 |
| 2/15/2004 | 138 |
| 2/16/2004 | 128 |
| 2/17/2004 | 139 |
| 2/18/2004 | 111 |
| 2/19/2004 | 132 |
| 2/20/2004 | 122 |
| 2/21/2004 | 129 |
| 2/22/2004 | 124 |
| 2/23/2004 | 155 |
| 2/24/2004 | 112 |
| 2/25/2004 | 130 |
| 2/26/2004 | 129 |
| 9/15/2004 | 138 |
| 9/16/2004 | 137 |
| 9/17/2004 | 127 |
| 9/18/2004 | 156 |
| 9/19/2004 | 131 |
| 9/20/2004 | 158 |
| 9/21/2004 | 151 |
| 9/22/2004 | 152 |
| 9/23/2004 | 142 |
| 9/24/2004 | 158 |
| 9/25/2004 | 150 |
| 9/26/2004 | 139 |
| 9/27/2004 | 141 |
| 9/28/2004 | 143 |
| 9/29/2004 | 143 |
| 9/30/2004 | 136 |
| 10/1/2004 | 142 |
| 10/2/2004 | 137 |
| 10/3/2004 | 138 |
| 10/4/2004 | 151 |
| 10/5/2004 | 141 |
| 10/6/2004 | 108 |
| 10/7/2004 | 115 |
| 10/8/2004 | 135 |
| 10/9/2004 | 157 |
| 10/10/2004 | 148 |
| 10/11/2004 | 159 |
| 10/12/2004 | 127 |
| 10/13/2004 | 111 |
| 10/14/2004 | 146 |
| 10/15/2004 | 128 |
| 10/16/2004 | 123 |
| 10/17/2004 | 140 |
| 10/18/2004 | 130 |
| 10/19/2004 | 145 |
| 10/20/2004 | 147 |
| 10/21/2004 | 144 |
| 10/22/2004 | 140 |
| 10/23/2004 | 162 |
| 10/24/2004 | 128 |
| 10/25/2004 | 142 |
| 10/26/2004 | 135 |
| 10/27/2004 | 128 |
| 5/17/2005 | 116 |
| 5/18/2005 | 142 |
| 5/19/2005 | 113 |
| 5/20/2005 | 115 |
| 5/21/2005 | 123 |
| 5/22/2005 | 138 |
| 5/23/2005 | 138 |
| 5/24/2005 | 121 |
| 5/25/2005 | 129 |
| 5/26/2005 | 107 |
| 5/27/2005 | 136 |
| 5/28/2005 | 135 |
| 5/29/2005 | 137 |
| 5/30/2005 | 153 |
| 5/31/2005 | 129 |
| 6/1/2005 | 141 |
| 6/2/2005 | 128 |
| 6/3/2005 | 103 |
| 6/4/2005 | 141 |
| 6/5/2005 | 137 |
| 6/6/2005 | 143 |
| 6/7/2005 | 143 |
| 6/8/2005 | 128 |
| 6/9/2005 | 137 |
| 6/10/2005 | 125 |
| 6/11/2005 | 125 |
| 6/12/2005 | 133 |
| 6/13/2005 | 152 |
| 6/14/2005 | 121 |
| 6/15/2005 | 134 |
| 6/16/2005 | 130 |
| 6/17/2005 | 161 |
| 6/18/2005 | 143 |
| 6/19/2005 | 130 |
| 6/20/2005 | 128 |
| 6/21/2005 | 117 |
| 6/22/2005 | 153 |
| 6/23/2005 | 134 |
| 6/24/2005 | 135 |
| 6/25/2005 | 124 |
| 6/26/2005 | 134 |
| 6/27/2005 | 144 |
| 6/28/2005 | 140 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 2/27/2004 | 138 | 10/28/2004 | 118 | 6/29/2005 | 135 |
| 2/28/2004 | 120 | 10/29/2004 | 130 | 6/30/2005 | 147 |
| 2/29/2004 | 137 | 10/30/2004 | 162 | 7/1/2005 | 147 |
| 3/1/2004 | 147 | 10/31/2004 | 135 | 7/2/2005 | 143 |
| 3/2/2004 | 138 | 11/1/2004 | 125 | 7/3/2005 | 145 |
| 3/3/2004 | 103 | 11/2/2004 | 116 | 7/4/2005 | 136 |
| 3/4/2004 | 130 | 11/3/2004 | 144 | 7/5/2005 | 137 |
| 3/5/2004 | 140 | 11/4/2004 | 125 | 7/6/2005 | 141 |
| 3/6/2004 | 130 | 11/5/2004 | 143 | 7/7/2005 | 129 |
| 3/7/2004 | 107 | 11/6/2004 | 135 | 7/8/2005 | 153 |
| 3/8/2004 | 115 | 11/7/2004 | 135 | 7/9/2005 | 138 |
| 3/9/2004 | 141 | 11/8/2004 | 140 | 7/10/2005 | 141 |
| 3/10/2004 | 122 | 11/9/2004 | 134 | 7/11/2005 | 127 |
| 3/11/2004 | 125 | 11/10/2004 | 134 | 7/12/2005 | 140 |
| 3/12/2004 | 110 | 11/11/2004 | 132 | 7/13/2005 | 145 |
| 3/13/2004 | 127 | 11/12/2004 | 134 | 7/14/2005 | 131 |
| 3/14/2004 | 124 | 11/13/2004 | 118 | 7/15/2005 | 161 |
| 3/15/2004 | 140 | 11/14/2004 | 135 | 7/16/2005 | 156 |
| 3/16/2004 | 127 | 11/15/2004 | 134 | 7/17/2005 | 145 |
| 3/17/2004 | 147 | 11/16/2004 | 129 | 7/18/2005 | 166 |
| 3/18/2004 | 138 | 11/17/2004 | 139 | 7/19/2005 | 152 |
| 3/19/2004 | 139 | 11/18/2004 | 122 | 7/20/2005 | 139 |
| 3/20/2004 | 126 | 11/19/2004 | 117 | 7/21/2005 | 142 |
| 3/21/2004 | 122 | 11/20/2004 | 138 | 7/22/2005 | 142 |
| 3/22/2004 | 130 | 11/21/2004 | 139 | 7/23/2005 | 142 |
| 3/23/2004 | 120 | 11/22/2004 | 159 | 7/24/2005 | 145 |
| 3/24/2004 | 108 | 11/23/2004 | 108 | 7/25/2005 | 147 |
| 3/25/2004 | 112 | 11/24/2004 | 128 | 7/26/2005 | 146 |
| 3/26/2004 | 126 | 11/25/2004 | 86 | 7/27/2005 | 146 |
| 3/27/2004 | 150 | 11/26/2004 | 145 | 7/28/2005 | 152 |
| 3/28/2004 | 136 | 11/27/2004 | 126 | 7/29/2005 | 132 |
| 3/29/2004 | 137 | 11/28/2004 | 126 | 7/30/2005 | 128 |
| 3/30/2004 | 137 | 11/29/2004 | 129 | 7/31/2005 | 139 |
| 3/31/2004 | 116 | 11/30/2004 | 117 | 8/1/2005 | 141 |
| 4/1/2004 | 112 | 12/1/2004 | 112 | 8/2/2005 | 147 |
| 4/2/2004 | 148 | 12/2/2004 | 118 | 8/3/2005 | 135 |
| 4/3/2004 | 123 | 12/3/2004 | 111 | 8/4/2005 | 147 |
| 4/4/2004 | 114 | 12/4/2004 | 134 | 8/5/2005 | 129 |
| 4/5/2004 | 115 | 12/5/2004 | 118 | 8/6/2005 | 147 |
| 4/6/2004 | 132 | 12/6/2004 | 124 | 8/7/2005 | 152 |
| 4/7/2004 | 128 | 12/7/2004 | 106 | 8/8/2005 | 147 |
| 4/8/2004 | 130 | 12/8/2004 | 145 | 8/9/2005 | 157 |
| 4/9/2004 | 143 | 12/9/2004 | 115 | 8/10/2005 | 147 |

| | |
|---|---|
| 4/10/2004 | 108 |
| 4/11/2004 | 98 |
| 4/12/2004 | 152 |
| 4/13/2004 | 132 |
| 4/14/2004 | 132 |
| 4/15/2004 | 135 |
| 4/16/2004 | 162 |
| 4/17/2004 | 134 |
| 4/18/2004 | 155 |
| 4/19/2004 | 135 |
| 4/20/2004 | 127 |
| 4/21/2004 | 154 |
| 4/22/2004 | 136 |
| 4/23/2004 | 124 |
| 4/24/2004 | 138 |
| 4/25/2004 | 115 |
| 4/26/2004 | 115 |
| 4/27/2004 | 130 |
| 4/28/2004 | 121 |
| 4/29/2004 | 140 |
| 4/30/2004 | 145 |
| 5/1/2004 | 124 |
| 5/2/2004 | 117 |
| 5/3/2004 | 140 |
| 5/4/2004 | 133 |
| 5/5/2004 | 123 |
| 5/6/2004 | 126 |
| 5/7/2004 | 159 |
| 5/8/2004 | 105 |
| 5/9/2004 | 115 |
| 5/10/2004 | 116 |
| 5/11/2004 | 161 |
| 5/12/2004 | 125 |
| 5/13/2004 | 117 |
| 5/14/2004 | 146 |
| 5/15/2004 | 135 |
| 5/16/2004 | 141 |
| 5/17/2004 | 133 |
| 5/18/2004 | 116 |
| 5/19/2004 | 137 |
| 5/20/2004 | 129 |
| 5/21/2004 | 136 |
| 5/22/2004 | 136 |
| 12/10/2004 | 138 |
| 12/11/2004 | 137 |
| 12/12/2004 | 120 |
| 12/13/2004 | 127 |
| 12/14/2004 | 107 |
| 12/15/2004 | 118 |
| 12/16/2004 | 116 |
| 12/17/2004 | 125 |
| 12/18/2004 | 111 |
| 12/19/2004 | 111 |
| 12/20/2004 | 129 |
| 12/21/2004 | 137 |
| 12/22/2004 | 107 |
| 12/23/2004 | 94 |
| 12/24/2004 | 98 |
| 12/25/2004 | 98 |
| 12/26/2004 | 122 |
| 12/27/2004 | 152 |
| 12/28/2004 | 137 |
| 12/29/2004 | 152 |
| 12/30/2004 | 124 |
| 12/31/2004 | 115 |
| 1/1/2005 | 138 |
| 1/2/2005 | 143 |
| 1/3/2005 | 142 |
| 1/4/2005 | 144 |
| 1/5/2005 | 131 |
| 1/6/2005 | 157 |
| 1/7/2005 | 152 |
| 1/8/2005 | 148 |
| 1/9/2005 | 145 |
| 1/10/2005 | 153 |
| 1/11/2005 | 123 |
| 1/12/2005 | 157 |
| 1/13/2005 | 116 |
| 1/14/2005 | 105 |
| 1/15/2005 | 130 |
| 1/16/2005 | 109 |
| 1/17/2005 | 137 |
| 1/18/2005 | 126 |
| 1/19/2005 | 114 |
| 1/20/2005 | 140 |
| 1/21/2005 | 120 |
| 8/11/2005 | 133 |
| 8/12/2005 | 141 |
| 8/13/2005 | 153 |
| 8/14/2005 | 127 |
| 8/15/2005 | 152 |
| 8/16/2005 | 140 |
| 8/17/2005 | 143 |
| 8/18/2005 | 133 |
| 8/19/2005 | 163 |
| 8/20/2005 | 129 |
| 8/21/2005 | 146 |
| 8/22/2005 | 153 |
| 8/23/2005 | 137 |
| 8/24/2005 | 119 |
| 8/25/2005 | 143 |
| 8/26/2005 | 158 |
| 8/27/2005 | 117 |
| 8/28/2005 | 151 |
| 8/29/2005 | 139 |
| 8/30/2005 | 130 |
| 8/31/2005 | 149 |
| 9/1/2005 | 146 |
| 9/2/2005 | 153 |
| 9/3/2005 | 138 |
| 9/4/2005 | 165 |
| 9/5/2005 | 146 |
| 9/6/2005 | 160 |
| 9/7/2005 | 156 |
| 9/8/2005 | 143 |
| 9/9/2005 | 163 |
| 9/10/2005 | 141 |
| 9/11/2005 | 138 |
| 9/12/2005 | 152 |
| 9/13/2005 | 147 |
| 9/14/2005 | 135 |
| 9/15/2005 | 134 |
| 9/16/2005 | 150 |
| 9/17/2005 | 129 |
| 9/18/2005 | 126 |
| 9/19/2005 | 161 |
| 9/20/2005 | 134 |
| 9/21/2005 | 138 |
| 9/22/2005 | 126 |

| | |
|---|---|
| 5/23/2004 | 128 |
| 5/24/2004 | 173 |
| 5/25/2004 | 150 |
| 5/26/2004 | 124 |
| 5/27/2004 | 122 |
| 5/28/2004 | 132 |
| 5/29/2004 | 141 |
| 5/30/2004 | 144 |
| 5/31/2004 | 147 |
| 6/1/2004 | 151 |
| 6/2/2004 | 135 |
| 6/3/2004 | 121 |
| 6/4/2004 | 133 |
| 6/5/2004 | 126 |
| 6/6/2004 | 132 |
| 6/7/2004 | 146 |
| 6/8/2004 | 119 |
| 6/9/2004 | 120 |
| 6/10/2004 | 116 |
| 6/11/2004 | 135 |
| 6/12/2004 | 116 |
| 6/13/2004 | 122 |
| 6/14/2004 | 170 |
| 6/15/2004 | 124 |
| 6/16/2004 | 138 |
| 6/17/2004 | 143 |
| 6/18/2004 | 153 |
| 6/19/2004 | 134 |
| 6/20/2004 | 119 |
| 6/21/2004 | 133 |
| 6/22/2004 | 150 |
| 6/23/2004 | 124 |
| 6/24/2004 | 124 |
| 6/25/2004 | 138 |
| 6/26/2004 | 124 |
| 6/27/2004 | 157 |
| 6/28/2004 | 149 |
| 6/29/2004 | 130 |
| 6/30/2004 | 131 |
| 7/1/2004 | 124 |
| 7/2/2004 | 133 |
| 7/3/2004 | 140 |
| 7/4/2004 | 128 |
| 1/22/2005 | 101 |
| 1/23/2005 | 108 |
| 1/24/2005 | 156 |
| 1/25/2005 | 158 |
| 1/26/2005 | 161 |
| 1/27/2005 | 146 |
| 1/28/2005 | 136 |
| 1/29/2005 | 133 |
| 1/30/2005 | 126 |
| 1/31/2005 | 151 |
| 2/1/2005 | 146 |
| 2/2/2005 | 139 |
| 2/3/2005 | 137 |
| 2/4/2005 | 143 |
| 2/5/2005 | 132 |
| 2/6/2005 | 138 |
| 2/7/2005 | 167 |
| 2/8/2005 | 139 |
| 2/9/2005 | 143 |
| 2/10/2005 | 127 |
| 2/11/2005 | 175 |
| 2/12/2005 | 144 |
| 2/13/2005 | 142 |
| 2/14/2005 | 149 |
| 2/15/2005 | 146 |
| 2/16/2005 | 155 |
| 2/17/2005 | 141 |
| 2/18/2005 | 141 |
| 2/19/2005 | 161 |
| 2/20/2005 | 117 |
| 2/21/2005 | 145 |
| 2/22/2005 | 142 |
| 2/23/2005 | 152 |
| 2/24/2005 | 113 |
| 2/25/2005 | 147 |
| 2/26/2005 | 126 |
| 2/27/2005 | 138 |
| 2/28/2005 | 132 |
| 3/1/2005 | 119 |
| 3/2/2005 | 135 |
| 3/3/2005 | 119 |
| 3/4/2005 | 131 |
| 3/5/2005 | 116 |
| 9/23/2005 | 140 |
| 9/24/2005 | 121 |
| 9/25/2005 | 137 |
| 9/26/2005 | 145 |
| 9/27/2005 | 139 |
| 9/28/2005 | 137 |
| 9/29/2005 | 136 |
| 9/30/2005 | 139 |
| 10/1/2005 | 144 |
| 10/2/2005 | 152 |
| 10/3/2005 | 148 |
| 10/4/2005 | 133 |
| 10/5/2005 | 150 |
| 10/6/2005 | 155 |
| 10/7/2005 | 131 |
| 10/8/2005 | 122 |
| 10/9/2005 | 152 |
| 10/10/2005 | 149 |
| 10/11/2005 | 133 |
| 10/12/2005 | 151 |
| 10/13/2005 | 143 |
| 10/14/2005 | 136 |
| 10/15/2005 | 139 |
| 10/16/2005 | 125 |
| 10/17/2005 | 151 |
| 10/18/2005 | 139 |
| 10/19/2005 | 169 |
| 10/20/2005 | 124 |
| 10/21/2005 | 158 |
| 10/22/2005 | 143 |
| 10/23/2005 | 140 |
| 10/24/2005 | 130 |
| 10/25/2005 | 141 |
| 10/26/2005 | 125 |
| 10/27/2005 | 118 |
| 10/28/2005 | 135 |
| 10/29/2005 | 132 |
| 10/30/2005 | 135 |
| 10/31/2005 | 151 |
| 11/1/2005 | 163 |
| 11/2/2005 | 122 |
| 11/3/2005 | 125 |
| 11/4/2005 | 136 |

| | | | | | |
|---|---|---|---|---|---|
| 7/5/2004 | 155 | 3/6/2005 | 138 | 11/5/2005 | 148 |
| 7/6/2004 | 150 | 3/7/2005 | 139 | 11/6/2005 | 139 |
| 7/7/2004 | 134 | 3/8/2005 | 131 | 11/7/2005 | 142 |
| 7/8/2004 | 151 | 3/9/2005 | 133 | 11/8/2005 | 126 |
| 7/9/2004 | 141 | 3/10/2005 | 149 | 11/9/2005 | 142 |
| 7/10/2004 | 157 | 3/11/2005 | 146 | 11/10/2005 | 122 |
| 7/11/2004 | 135 | 3/12/2005 | 157 | 11/11/2005 | 123 |
| 7/12/2004 | 160 | 3/13/2005 | 165 | 11/12/2005 | 122 |
| 7/13/2004 | 138 | 3/14/2005 | 145 | 11/13/2005 | 127 |
| 7/14/2004 | 137 | 3/15/2005 | 156 | 11/14/2005 | 116 |
| 7/15/2004 | 114 | 3/16/2005 | 121 | 11/15/2005 | 139 |
| 7/16/2004 | 130 | 3/17/2005 | 125 | 11/16/2005 | 134 |
| 7/17/2004 | 132 | 3/18/2005 | 151 | 11/17/2005 | 126 |
| 7/18/2004 | 136 | 3/19/2005 | 128 | 11/18/2005 | 139 |
| 7/19/2004 | 131 | 3/20/2005 | 132 | 11/19/2005 | 156 |
| 7/20/2004 | 118 | 3/21/2005 | 145 | 11/20/2005 | 122 |
| 7/21/2004 | 139 | 3/22/2005 | 141 | 11/21/2005 | 133 |
| 7/22/2004 | 119 | 3/23/2005 | 140 | 11/22/2005 | 129 |
| 7/23/2004 | 120 | 3/24/2005 | 147 | 11/23/2005 | 108 |
| 7/24/2004 | 138 | 3/25/2005 | 158 | 11/24/2005 | 102 |
| 7/25/2004 | 119 | 3/26/2005 | 126 | 11/25/2005 | 133 |
| 7/26/2004 | 132 | 3/27/2005 | 112 | 11/26/2005 | 116 |
| 7/27/2004 | 138 | 3/28/2005 | 164 | 11/27/2005 | 124 |
| 7/28/2004 | 135 | 3/29/2005 | 148 | 11/28/2005 | 146 |
| 7/29/2004 | 115 | 3/30/2005 | 139 | 11/29/2005 | 130 |
| 7/30/2004 | 148 | 3/31/2005 | 147 | 11/30/2005 | 135 |
| 7/31/2004 | 117 | 4/1/2005 | 118 | 12/1/2005 | 125 |
| 8/1/2004 | 135 | 4/2/2005 | 137 | 12/2/2005 | 132 |
| 8/2/2004 | 137 | 4/3/2005 | 134 | 12/3/2005 | 125 |
| 8/3/2004 | 130 | 4/4/2005 | 138 | 12/4/2005 | 114 |
| 8/4/2004 | 117 | 4/5/2005 | 155 | 12/5/2005 | 135 |
| 8/5/2004 | 141 | 4/6/2005 | 145 | 12/6/2005 | 138 |
| 8/6/2004 | 136 | 4/7/2005 | 151 | 12/7/2005 | 108 |
| 8/7/2004 | 116 | 4/8/2005 | 177 | 12/8/2005 | 133 |
| 8/8/2004 | 126 | 4/9/2005 | 151 | 12/9/2005 | 103 |
| 8/9/2004 | 131 | 4/10/2005 | 141 | 12/10/2005 | 142 |
| 8/10/2004 | 127 | 4/11/2005 | 142 | 12/11/2005 | 108 |
| 8/11/2004 | 138 | 4/12/2005 | 147 | 12/12/2005 | 124 |
| 8/12/2004 | 117 | 4/13/2005 | 134 | 12/13/2005 | 111 |
| 8/13/2004 | 125 | 4/14/2005 | 132 | 12/14/2005 | 106 |
| 8/14/2004 | 130 | 4/15/2005 | 152 | 12/15/2005 | 102 |
| 8/15/2004 | 110 | 4/16/2005 | 152 | 12/16/2005 | 129 |
| 8/16/2004 | 152 | 4/17/2005 | 153 | 12/17/2005 | 114 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 8/17/2004 | 133 | 4/18/2005 | 161 | 12/18/2005 | 111 |
| 8/18/2004 | 144 | 4/19/2005 | 130 | 12/19/2005 | 106 |
| 8/19/2004 | 125 | 4/20/2005 | 157 | 12/20/2005 | 107 |
| 8/20/2004 | 142 | 4/21/2005 | 138 | 12/21/2005 | 110 |
| 8/21/2004 | 127 | 4/22/2005 | 144 | 12/22/2005 | 128 |
| 8/22/2004 | 132 | 4/23/2005 | 124 | 12/23/2005 | 102 |
| 8/23/2004 | 131 | 4/24/2005 | 104 | 12/24/2005 | 89 |
| 8/24/2004 | 170 | 4/25/2005 | 139 | 12/25/2005 | 99 |
| 8/25/2004 | 118 | 4/26/2005 | 152 | 12/26/2005 | 157 |
| 8/26/2004 | 131 | 4/27/2005 | 124 | 12/27/2005 | 135 |
| 8/27/2004 | 150 | 4/28/2005 | 125 | 12/28/2005 | 132 |
| 8/28/2004 | 146 | 4/29/2005 | 131 | 12/29/2005 | 148 |
| 8/29/2004 | 128 | 4/30/2005 | 136 | 12/30/2005 | 133 |
| 8/30/2004 | 158 | 5/1/2005 | 126 | 12/31/2005 | 93 |
| 8/31/2004 | 128 | 5/2/2005 | 138 |  |  |

# BIBLIOGRAPHY

Andrieu, C., N. D. Freitas, et al. (2003). An introduction to MCMC for machine learning. Machine Learning **50**: 5-43.

Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics **49**: 803-821.

Banks, J. (1989). Principles of quality control. New York, John Wiley.

Bar-Shalom, Y. and T. E. Fortmann (1988). Tracking and data association. San Diego, CA, Academic Press.

Baron, M. I. (2002). Bayes and asymptotically pointwise optimal stopping rules for the detection of influenza epidemics. Case Studies in Bayesian Statistics **6**: 153-63.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). Journal of the Royal Statistical Society. Series B **41**: 113-147.

Bertsekas, D. P. and J. N. Tsitsiklis (2002). Section 4.3: More on conditional expectation and variance. Introduction to Probability, Athena Scientific.

Bishop, C. M. (2006). Pattern recognition and machine learning. New York, Springer.

Borchers, D. L., S. T. Buckland, et al. (2002). Estimating animal abundance. New York, Springer.

Box, G. E. P. and G. M. Jenkins (1976). Time series analysis: forecasting and control. San Francisco, CA, Holden-Day.

Box, G. E. P. and G. C. Tiao (1973). Bayesian inference in statistical analysis. MA, Addison-Wesley.

Buckeridge, D. L. (2007). Outbreak detection through automated surveillance: A review of the determinants of detection. Journal of Biomedical Informatics **40**: 370-379.

Buckeridge, D. L., H. S. Burkom, et al. (2005). Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics **38**: 99-113.

Buntine, W. (1994). Operations for learning with graphical models. Journal of Artificial Intelligence Research **2**: 159-225.

Burkom, H. S. (2003). Biosurveillance applying scan statistics with multiple, disparate data sources. Journal of Urban Health **80**(2, suppl 1): i57-i65.

Campbell, J. G., C. Fraley, et al. (1997). Linear flaw detection in woven textiles using model-based clustering. Pattern Recognition Letters **18**: 1539-1548.

Campbell, J. G., C. Fraley, et al. (1999). Model-based methods for real-time textile fault detection. International Journal of Imaging Systems and Technology **10**: 339-346.

Carbonetto, P., J. Kisynski, et al. (2005). Nonparametric Bayesian logic. In Proc. 21th Conference on Uncertainty in AI: 85-93.

Carlstein, E. (1988). Nonparametric change-point estimation. The Annals of Statistics **16**(1): 188-197.

Casella, G. and L. R. Berger (2002). Statistical Inference (Second Edition). Australia; Pacific Grove, CA Thomson Learning.

Castillo, E. and B. M. Colosimo (2007). An introduction to Bayesian inference in process monitoring, control and optimization. Bayesian process monitoring, control and optimization. B. M. Colosimo and E. Castillo. Boca Raton, Chapman and Hall**:** 3-46.

CDC. "CDC | Bioterrorism Agents/Diseases | Emergency Preparedness & Response <http://www.bt.cdc.gov/agent/agentlist-category.asp>."

Chakraborty, D. (2002). Statistical power in observer-performance studies. Comparison of the receiver-operating characteristic and free-response methods in tasks involving localization. Academic Radiology **9**: 147-156.

Chandola, V., A. Banerjee, et al. (2009). Anomaly detection: a survey. To appear in ACM Computing Surveys.

Chapman, W. W., J. N. Dowling, et al. (2004). Fever detection from free-text clinical records for biosurveillance. Journal of Biomedical Informatics **37**(2): 120-127.

Chapman, W. W., J. N. Dowling, et al. (2005). Classification of emergency department chief complaints into seven syndromes: A retrospective analysis of 527,228 patients. Annals of Emergency Medicine **46**(5): 445-455.

Chu, D. (2007). Clinical feature extraction from emergency department reports for biosurveillance. Master's Thesis. Department of Biomedical Informatics. University of Pittsburgh. Pittsburgh.

Codd, E. F. (1970). A relational model of data for large shared data banks. Communications of the ACM **13**(6): 377-387.

Cooper, G. F. (1995). A Bayesian method for learning belief networks that contain hidden variables. Journal of Intelligent Information Systems **4**: 71-88.

Cooper, G. F., D. H. Dash, et al. (2004). Bayesian biosurveillance of disease outbreaks. Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence: 94-103.

Cooper, G. F., J. N. Dowling, et al. (2006). A Bayesian algorithm for detecting CDC category A outbreak diseases from emergency department chief complaints. Advances in Disease Surveillance **2**: 45.

Cooper, G. F. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning **9**: 309-347.

Crosier, R. B. (1988). Multivariate generalizations of cumulative sum quality-control schemes. Technometrics **30**: 291-303.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. Journal of American Statistical Association **93**: 294-302.

Davidian, M. (2007). Linear mixed effects models for multivariate normal data. class notes for Applied Longitudinal Data Analysis, North Carolina State University**:** 363-422.

Dempster, A. P., N. M. Laird, et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) **39**(1): 1-38.

Edgeworth, F. Y. (1887). On discordant observations. Philosophical Magazine **23**(5): 364-375.

Fawcett, T. and F. Provost (1997). Adaptive fraud detection. Data Mining and Knowledge Discovery **1**(3): 291-316.

Fawcett, T. and F. Provost (1999). Activity monitoring: Noticing interesting changes in behavior. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining 53-62.

Frisen, M. and P. Wessman (1999). Evaluations of likelihood ratio methods for surveillance: differences and robustness. Communications in Statistics. Simulations and Computations **28**: 597-622.

Gajewski, B. J. and M. S. Mayo (2006). Bayesian sample size calculations in phase II clinical trials using a mixture of informative priors. Statistics in Medicine **25**: 2554-2566.

Gelman, A., J. B. Carlin, et al. (1995). Bayesian data analysis. London, Chapman & Hall.

Getoor, L., N. Friedman, et al. (2002). Learning probabilistic relational models of link structure. Journal of Machine Learning Research **3**: 679-707.

Gilks, W. R., A. Thomas, et al. (1994). A language and program for complex Bayesian modeling. The Statistician **43**(1): 169-177.

Goldenberg, A., G. Shmueli, et al. (2002). Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proceedings of National Academy of Sciences **99**(8): 5237-40.

Hamilton, J. D. (1994). Time series analysis. Princeton, NJ, Princeton University Press.

Hardy, G. F. (1889). Insurance record. (reprinted in Transactions of Actuaries, 1920) **8**.

Hartigan, J. A. (1965). The asymptotically unbiased prior distribution. Annals of Mathematical Statistics **36**: 1137-1152.

Harvey, A. (1981). The Kalman filter and its applications in econometrics and time series analysis. Methods of Operations Research **44**: 3-18.

Heckerman, D., C. Meek, et al. (2004). Probabilistic entity-relationship models, PRMs, and Plate models. SRL2004: Statistical Relational Learning and its Connections to Other Fields - ICML 2004 Workshop. Banff, Alberta, Canada.

Heckerman, D., C. Meek, et al. (2004). Probabilistic models for relational data. Redmond, WA, Technical Report MSR-TR-2004-30, Microsoft Research.

Hogan, W. R., G. F. Cooper, et al. (2004). A Bayesian anthrax aerosol release detector. RODS Technical Report.

Hogan, W. R., G. F. Cooper, et al. (2007). The Bayesian aerosol release detector: An algorithm for detecting and characterizing outbreaks caused by an atmospheric release of Bacillus anthracis. Statistics in Medicine **26**: 5225-5252.

Hosmer, D. W. and S. Lemeshow (1989). Applied logistic regression. New York, John Wiley & Sons, Inc.

Hutwagner, L. C., W. Thompson, et al. (2003). The bioterrorism preparedness and response early aberration reporting system (EARS). Journal of Urban Health **80**(2, Supplement 1): i89-i96.

Jaynes, E. T. (1957). Information theory and statistical mechanics I, II. Physical Review **106**: 620-630.

Jaynes, E. T. (1968). Prior probabilities. IEEE Transactions on Systems Science and Cybernetics **SSC-4**: 227-241.

Jeffreys, H. (1961). Theory of probability (3rd ed.). London, Oxford University Press.

Jiang, X. (2008). A Bayesian network model for spatio-temporal event surveillance. Doctoral Dissertation. Department of Biomedical Informatics. University of Pittsburgh. Pittsburgh.

Kass, R. E. and L. Wasserman (1996). The selection of prior distributions by formal rules. Journal of the American Statistical Association **91**(435): 1343-1370.

Koller, D. and A. Pfeffer (1997). Object-oriented Bayesian networks. Thirteenth Conference on Uncertainty in Artificial Intelligence, Providence, RI, Morgan Kaufmann, San Mateo, CA.

Koller, D. and A. Pfeffer (1998). Probabilistic frame-based system. In Proc. 15th AAAI National Conference on Artificial Intelligence: 580-587.

Kulldorff, M. (1997). A spatial scan statistic. Communications in Statistics. Theory and Methods **26**: 1481-1496.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. Journal of the Royal Statistical Society. Series A (Statistics in Society) **164**: 61-72.

Lawson, A. B. (2009). Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Chapman & Hall.

Lawson, A. B. and K. P. Kleinman (2005). Spatial and syndromic surveillance for public health, John Wiley & Sons.

Lazarus, R., K. P. Kleinman, et al. (2002). Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. Emerging Infectious Diseases **8**(8): 753-760.

Lehmann, E. L. and G. Casella (1998). Theory of point estimation. New York, Springer-Verlag.

LeStrat, Y. and F. Carrat (1999). Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in Medicine **18**: 3463-78.

Lowry, C. A., W. H. Woodall, et al. (1992). A multivariate exponentially weighted moving average control chart. Technometrics **34**(1): 46-53.

Mandl, K. D., O. J.M., et al. (2004). Implementing syndromic surveillance: a practical guide informed by the early experience. Journal of the American Medical Informatics Association **11**(2): 141-50.

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

Milch, B. (2006). Probabilistic Models with Unknown Objects. Doctoral Dissertation. Computer Science Division, University of California, Berkeley.

Milch, B., B. Marthi, et al. (2007). BLOG: probabilistic models with unknown objects. Introduction to Statistical Relational Learning. L. Getoor and B. Taskar. Cambridge, MA, MIT Press.

Mitchell, T. M. (1997). 6.9 Naive Bayes Classifier. Machine learning, McGraw Hill**:** 177.

Montgomery, D. C. (1991). Introduction to statistical quality control. New York, Wiley.

Moore, A. W., G. F. Cooper, et al. (2003). Summary of biosurveillance-relevant statistical and data mining technologies. Pittsburgh, RODS Laboratory. University of Pittsburgh.

Mostashari, F. and J. Hartman (2003). Syndromic surveillance: a local perspective. Journal of Urban Health: Bulletin of the New York Academy of Medicine **80**(2): i1-7.

Mukherjee, S., E. D. Feigelson, et al. (1998). Three types of Gamma ray bursts. The Astrophysical Journal **508**: 314-327.

Murdoch, D. J. and X.-L. Meng (2001). Towards perfect sampling for Bayesian mixture priors. "citeseer.ist.psu.edu/article/murdoch01towards.html".

Neapolitan, R. E. (2003). Learning Bayesian networks, Prentice Hall.

Neill, D. B., A. W. Moore, et al. (2005). Detection of emerging space-time clusters. 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Nobre, F. F., A. B. S. Monteiro, et al. (2001). Dynamic linear models and SARIMA: a comparison of their forecasting performance in epidemiology. Statistics in Medicine **20**: 3051-3069.

Palomo, J., D. R. Insua, et al. (2005). On combing expertise in dynamic linear models, Statistical and Applied Mathematical Sciences Institute.

Pasula, H., B. Marthi, et al. (2003). Identity uncertainty and citation matching. Neural Information Processing Systems Conference. Vancouver, B.C.

Pearl, J. (2000). Causality: models, reasoning, and Inference, Cambridge University Press.

Pinheiro, J. C. and D. M. Bates (2000). Mixed-Effects Models in S and S-Plus. New York, Springer-Verlag New York, Inc.

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

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of the IEEE **77**: 257-285.

Rath, T., M. Carreras, et al. (2003). Automated detection of influenza epidemics with hidden Markov models. Proceedings of the Fifth International Symposium on Intelligent Data Analysis **2810**: 521-532.

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

Serfling, R. E. (1963). Methods for current statistical analysis of excess pneumonia-influenza deaths. Public Health Reports **78**: 494-506.

Shen, Y. and G. F. Cooper (2007). A Bayesian biosurveillance method that models unknown outbreak diseases. Proceedings of Intelligence and Security Informatics: BioSurveillance: 209-215.

Shen, Y. and G. F. Cooper (2009). A new prior for Bayesian anomaly detection, DBMI-09-369, Department of Biomedical Informatics, University of Pittsburgh.

Shiryaev, A. N. (1978). Optimal stopping rules. New York, Springer.

Siegrist, D. and J. A. Pavlin (2004). Bio-ALIRT biosurveillance detection algorithm evaluation. MMWR Morbility and Mortality Weekly Report **Sep 24**(53): 152-158.

Sittler, R. W. (1964). An optimal data association problem in surveillance theory. IEEE Transactions on Military Electronics **MIL-8**: 125-139.

Sonesson, C. and D. Bock (2003). A review and discussion of prospective statistical surveillance in public health. Journal of the Royal Statistical Society: Series A (Statistics in Society) **166**(Part 1): 5-21.

Stroup, F. D. and S. B. Thacker (1995). A Bayesian approach to the detection of aberrations in public health surveillance data. Epidemiology(4): 435-443.

Toothaker, L. E. (1993). Multiple comparison procedures, Sage Publications Inc.

Tuyl, F., R. Gerlach, et al. (2009). Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. Bayesian Analysis **4**(1): 151-158.

Ullman, J. and J. Widom (1997). First course in database systems. Upper Saddle River, NJ, Prentice Hall.

Wagner, M. M., A. W. Moore, et al., Eds. (2006). Handbook of biosurveillance, Elsevier.

Warrender, C., S. Forrest, et al. (1999). Detecting intrusions using system calls: alternative data models. IEEE Symposium on Security and Privacy: IEEE Computer Society: 133-45.

Wasserman, L. (2004). All of Statistics: A Concise Course in Statistical Inference. New York, NY, Springer.

Weisstein, E. W. (2009). ""Hypergeometric Distribution." From MathWorld -- A Wolfram Web Resource." from http://mathworld.wolfram.com/HypergeometricDistribution.html.

Welch, B. L. and H. W. Peers (1963). On formulae for confidence points based on integrals of weighted likelihoods. Journal of the Royal Statistical Society. Series B **25**: 318-329.

West, M. and J. Harrison (1989). Bayesian forecasting and dynamic models. New York, Springer-Verlag.

Whitworth, W. A. (1897). Exercise in choice and chance. (reprinted by Hafner, New York, 1965).

Wong, W.-K. (2004). Data mining for early disease outbreak detection. Doctoral Dissertation. Carnegie Mellon University, Pittsburgh.

Zabell, S. L. (1982). W.E. Johnson's 'sufficientness' postulate. The Annals of Statistics **10**(4): 1090-1099.

Zellner, A., Ed. (1977). Maximal data information prior distribution. New Developments in the Applications of Bayesian Methods. Amsterdam, North-Holland.

Zhang, J., F. C. Tsui, et al. (2003). Detection of outbreaks from time series data using wavelet transform. AMIA Annual Symposium Proceedings: 748-52.