Bayesian Biosurveillance Using Multiple Data Streams

Weng-Keen Wong,[1] G. F. Cooper, [1] D. H. Dash,[2] J. D. Levander, [1] J. N. Dowling, [1] W. R. Hogan, [1] M. M. Wagner[1]

[1]RODS Laboratory, University of Pittsburgh, Pittsburgh, Pennsylvania; [2]Intel Research, Santa Clara, California.

Corresponding author: Weng-Keen Wong, Center for Biomedical Informatics, University of Pittsburgh, 8084 Forbes Tower, 200 Lothrop Street, Pittsburgh, PA, 15213. Telephone: 412-647-7113; Fax: 412-647-7190; Email: wwong@cbmi.pitt.edu

ABSTRACT

Introduction: Emergency Department (ED) records and over-the-counter (OTC) sales data are two of the most commonly used data sources for syndromic surveillance. Most current detection algorithms monitor these data sources separately, and either do not combine them, or combine them in an *ad hoc* fashion. This paper introduces a causal model that coherently combines the two data sources in order to perform outbreak detection.

Objectives: This paper presents a Bayesian biosurveillance algorithm called PANDA that combines information from multiple data streams. We describe the model, along with an explication of assumptions and techniques used to make this approach scalable for real-time surveillance of a large population.

Methods: We extend the causal Bayesian network model used in *(1)* to incorporate evidence from daily OTC sales data. We model, at the level of individual people, the actions that result in the purchase of OTC products, as well as admission to an ED.

Results: The aim of this paper is to describe a detection model for monitoring both ED and OTC data. This paper provides preliminary support that despite the complexities of this model, the running time is tractable.

Conclusion: This paper introduces a new Bayesian biosurveillance algorithm that models the interaction between ED and OTC data. It also provides preliminary results that are positive regarding the run time of the algorithm.

1 INTRODUCTION

Two data sources that are routinely monitored by syndromic surveillance systems are over-the-counter (OTC) medication sales and Emergency Department (ED) chief complaint records *(2; 3)*. If a disease outbreak occurs in a region, we often expect its effects to be seen in both these data sources *(4; 5)*. Even though ED and OTC data sources contain the signal of an outbreak, detection algorithms usually monitor each type of data separately. This independent analysis limits the detection capabilities of a surveillance system.

In order to illustrate this point, consider the characteristics of OTC data. The signal of an outbreak is often expected to appear first in OTC medication sales then in ED data (4). The notion is that individuals with the initial symptoms of the disease will typically attempt to treat themselves before seeking medical care *(6-8)*. While the early signal in OTC data is appealing, this signal will be weak. In addition, OTC data is often reported as a univariate time series, in which the daily regional sales volume for a particular category of product (e.g., sales of cough medications) is recorded each day. This univariate time series consists of data aggregated over time. Case-level data about each sales transaction are not available. If such a level of detail were available, we could apply a multivariate detection algorithm and potentially exploit the additional information about each transaction to improve our detection capability.

In contrast to OTC data, ED chief complaint data does contain case-level information about each patient admitted. Details that are typically recorded include admission date and time,

age, gender, home zip code, and chief complaint. Using this data, we can improve a detection algorithm's capability by specifically looking for known spatial, temporal, demographic and symptomatic patterns of the disease in the data. However, as was previously mentioned, we expect the signal of an outbreak to often appear later in ED data than in OTC data.

One way to combine the advantages of both types of data is to develop a detection algorithm that integrates the data sources. If both data sources are being jointly monitored, then the combined information could reinforce our belief that an outbreak is happening. The key difficulty with this data fusion approach is in measuring the relationship between the data sources when an outbreak occurs. The correlation between OTC and ED data during outbreak conditions cannot be learned because no training data exists that captures the effects of a large-scale epidemic on these data sources during the same time period. Most existing data fusion approaches *(9-11)* treat the data sources as if they are independent. Although training data do not exist, we do have some background knowledge about the plausible relationship between OTC and ED data for a given disease. Our approach to combining multiple data streams relies on using this background knowledge to model the actions of individuals that result in possible OTC medication purchases and ED admissions.

This paper extends the Population-wide Anomaly Detection and Assessment (PANDA) algorithm *(1)*, which uses a causal Bayesian network to model an entire population of people. In the original PANDA model, the algorithm was designed to monitor only ED chief complaint data. In this paper, we enhance PANDA to simultaneously monitor data sources of different granularity – specifically aggregated regional counts for OTC sales and multivariate ED records for individual patients. Although the Bayesian network presented in *(1)* can be used to model the effects of any non-contagious disease outbreak in a geographic area, we will focus on detecting

an outdoor, aerosolized release of an anthrax-like agent within a county-wide region being monitored.

2 METHODS

The key aspect of the PANDA algorithm is the explicit modeling of each individual in the population as a subnetwork of the overall causal Bayesian network. We will refer to models of these individuals as *person models*, although these models could be generalized to entities that provide information about disease outbreaks, such as biosensors and livestock. The advantage of modeling each individual in the population is to allow the algorithm to have a great deal of representational power and representational flexibility. By having a subnetwork for each individual, we can coherently represent different types of background knowledge in the model. For instance, with an aerosolized anthrax release, we can build into the model our temporal assumption about the incubation time of anthrax as well as our spatial assumption that the release will take the shape of a downwind plume *(12; 13)*. Besides the power in representing prior knowledge, modeling the entire population allows the model to *combine* spatial, temporal, demographic, and symptomatic evidence to derive a posterior probability of a disease outbreak. As for representational flexibility, the individual modeling allows new types of knowledge and evidence to be readily incorporated into the model. As an example, if we gain access to radiology reports for a group of individuals, and we find that radiology reports are especially useful indicators of an anthrax attack, we can then easily add this new evidential variable to the model. Finally, most of the background knowledge of the characteristics of respiratory anthrax disease is at an individual rather than population level.

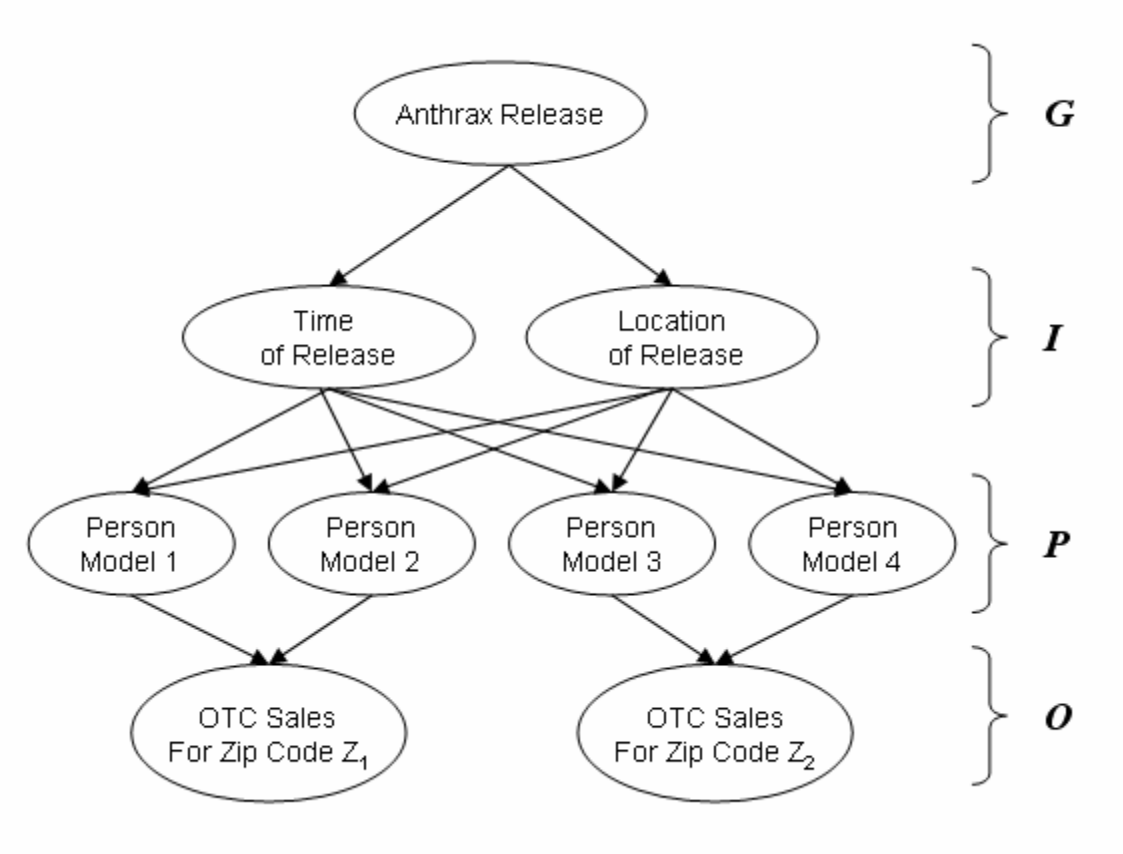

Figure 1: The anthrax model used by PANDA

2.1 The Generic Model

Figure 1 illustrates the causal Bayesian network that we use to detect an outbreak caused by an aerosolized release of an anthrax-like bio-agent. This network is an instantiation of a generic model for infectious but non-contagious diseases. We will first present the generic model, using the network in Figure 1 as an example. Then, we will present the details of the specific model in Figure 1. We can partition the nodes $\boldsymbol{X}$ of the generic model into four parts:

1. A set of *global nodes* $\boldsymbol{G}$ contains the nodes which are global features common to all people. Included in the set $\boldsymbol{G}$ is the target node T, which indicates the occurrence of a disease outbreak. By monitoring the state of T, we can derive an updated posterior probability that a disease outbreak is occurring. In Figure 1, the target node T is *Anthrax Release*, which also happens to be the only node in $\boldsymbol{G}$. The set $\boldsymbol{G}$ could include other

variables such as the national terror alert level or information about local events such as major sports events or political conventions.

2. A set of *interface nodes* ***I,*** illustrated by *Time of Release* and *Location of Release* in Figure 1. For the purposes of simplicity, we only include these two nodes to the specific model. We plan to refine our model by adding other nodes to the interface layer such as the amount of release, the type of anthrax powder, and meteorological information.
3. A set of *person models* $\boldsymbol{P} = \{P_1, P_2, \ldots, P_n\}$, one for each person. Evidence observed on an individual basis will be entered at the person model level.
4. A set of *population-wide evidence nodes* $\boldsymbol{O} = \{O_1, O_2, \ldots, O_m\}$. The set $\boldsymbol{O}$ represents evidence that is aggregated over a particular group of people, such as a group living in a particular geographic region. In the specific model, the set $\boldsymbol{O}$ consists of the aggregate OTC sales of a particular type (e.g., cough medication sales) over a particular zip code.

The generic model makes three assumptions that are intended to facilitate inference on *T*. These assumptions are:

Assumption 1: $\forall i, P_i \coprod P_{-i} \mid \boldsymbol{I}$ and any arc between a node *I* in $\boldsymbol{I}$ and a node *X* in some person model P_i is oriented from *I* to *X*. The symbol $\coprod$ means "is independent of" and the notation P_{-i} means all person models except P_i. The fact that we do not condition on the population-wide evidence in $\boldsymbol{O}$ in this assumption may seem counter-intuitive, but as we will describe in Section 3, we do not use the evidence in $\boldsymbol{O}$ when we calculate the contribution of the evidence in $\boldsymbol{P}$ to the posterior probability.

Assumption 2: $\boldsymbol{G} \coprod \boldsymbol{P} \mid \boldsymbol{I}$ and any arc between a node G in $\boldsymbol{G}$ and a node I in $\boldsymbol{I}$ is oriented from G to I.

Assumption 3: The person models P_i contain arcs that are oriented towards the population-wide evidence nodes in $\boldsymbol{O}$.

Thus, from Assumptions 1 to 3, we do not allow arcs directly between the person models. For non-contagious diseases that may cause outbreaks, Assumptions 1 to 3 are reasonable when $\boldsymbol{I}$ contains all of the factors that significantly influence the status of an outbreak disease in individuals in the population. In the case of bioterrorist-released bio-agents, for example, such information includes the time and location of release of the agent. A key characteristic of nodes in $\boldsymbol{I}$ is that they have arcs to the nodes in one or more person models, and they induce the conditional independence relationships described in Assumptions 1 and 2. On the other hand, for contagious diseases, which are not the topic of the current paper, arcs will be needed between person models, since people can infect each other. Once Assumptions 1 to 3 no longer hold, inference becomes much more computationally expensive and we cannot perform the optimizations described in later sections that allow PANDA to run efficiently.

2.2 The Anthrax Model

In this prototype model, we make the simplifying assumption that individuals living in a zip code only purchase OTC medication within their own home zip code. Consequently, the OTC purchases in each zip code are independent of each other. This assumption will of course be violated in the event of a large-scale bioterrorist attack. We plan to address this issue in future

work but we will operate using this assumption for this initial prototype. One straightforward way to avoid this assumption is to model the population-wide evidence ***O*** as the OTC sales over the entire region of Allegheny county; however, doing so would lose spatial information that might be helpful in detecting an outbreak. The *OTC Sales for Zip Code* nodes are integer-valued nodes representing the aggregate number of units of OTC medication sold over the specified zip code. These nodes are considered to be observed nodes because they are instantiated with values from the OTC data.

The person model is shown in more detail in Figure 2. The structure of this model was created by expert judgment. Some of the nodes in the person model are temporal nodes, which we model over a three-day period. We use three days in this initial prototype because it is the shortest period of time that we believe is meaningful for modeling a disease outbreak. We intend to expand this time period to two weeks for modeling an anthrax outbreak. Finally, we will refer to the nodes *Home Zip*, *Age Decile*, *Gender*, *Respiratory Chief Complaint When Admitted*, and *ED Admission* as evidence nodes because their values are observed in the ED data.

The parameters of some of the nodes in Figure 2 were estimated from a training set consisting of one year's worth of ED patient data from the year 2000 or from one year's worth of OTC data from 2004. The parameters of other variables were obtained from U.S. Census data about the region. The rest had their respective probabilities (whether prior or conditional) derived as a logical function of their parents or assessed subjectively. The subjective assessments were informed by the literature and by general knowledge about infectious diseases. Due to space restrictions, we will only describe below the nodes that are different from the original PANDA model in (1)

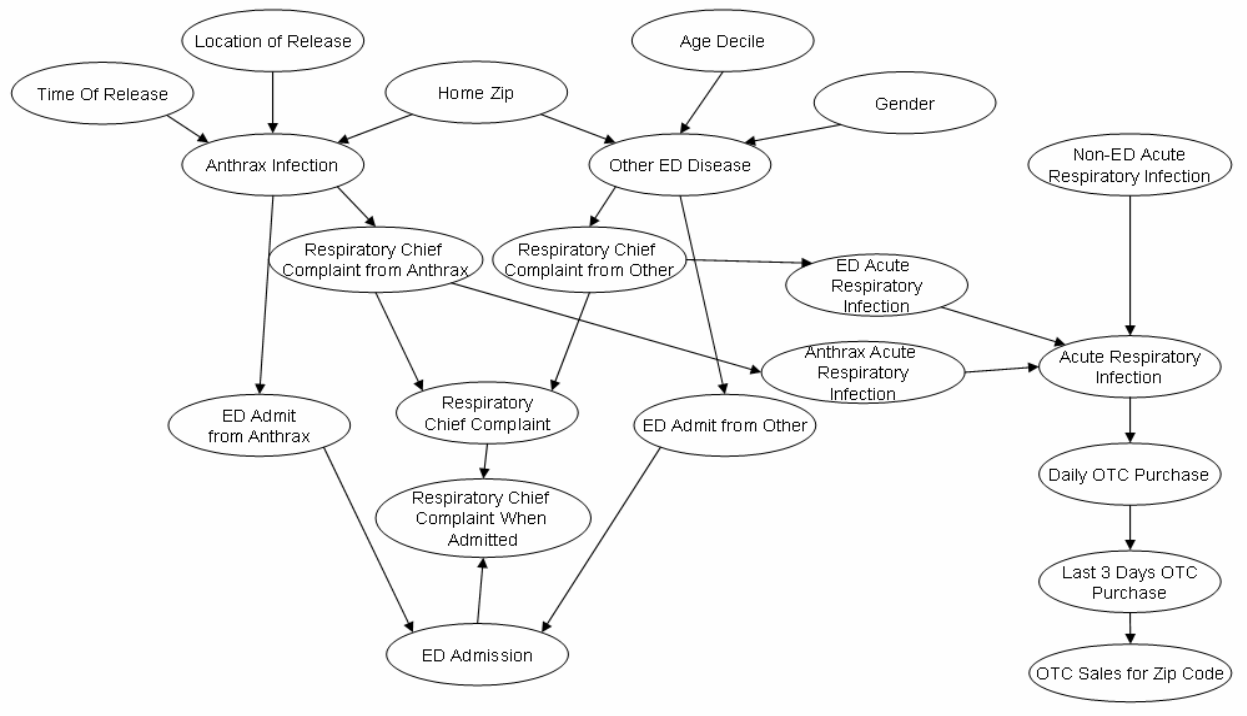

Figure 2: The person model used by PANDA. Note that the *Time of Release* and *Location of Release* nodes are part of the interface node layer. The *OTC Sales for Zip Code* node is part of the population-wide evidence layer. Although the OTC *Sales for Zip Code* node is only shown with input from a single person model, it actually has arcs from all person models in a given zip code, as illustrated in the simple example in Figure 1.

- *Anthrax Acute Respiratory Infection*: Indicates whether an individual has an acute respiratory infection (ARI) due to anthrax. This node models the presence of ARI over a three-day period, much like the *Anthrax Infection* node.

- *ED Acute Respiratory Infection*: Indicates whether an individual, who presents to the ED, has an acute respiratory infection (ARI) due to an ED disease other than anthrax. This is a temporal node, and it has the same states as those of *Anthrax Infection.*

- *Non-ED Acute Respiratory Infection*: Indicates whether an individual, who does not present to the ED, has an acute respiratory infection (ARI) due to a disease other than anthrax. This is also a temporal node, and it has the same type of states as those of *Anthrax Infection*.

- *Acute Respiratory Infection*: Indicates if the patient has ARI. This node is a "logical or" of *Anthrax Acute Respiratory Infection*, *ED Acute Respiratory Infection* and *Non-ED Acute Respiratory Infection.*

- *Daily OTC Purchase*: Captures the probability of purchasing a respiratory-related OTC medication (e.g., a cough medication) today, yesterday, the day before yesterday, or never. We derived the conditional probability of this variable based on the literature and data from OTC sales of respiratory-related medication in the county-wide region being modeled.
- *Last 3 Days OTC Purchase*: A Boolean node describing if a respiratory-related OTC medication was purchased in the last three days by this person being modeled.

2.3 Inference

The goal of PANDA is to monitor the state of the target node *T*, which captures the probability of a disease outbreak occurring. Specifically, PANDA calculates the posterior probability of *T* as new ED and OTC data arrive. Let ***o*** be the set of population-wide evidence, namely the OTC sales volume for each zip code in the county-wide region. Similarly, let ***e*** be the collective set of evidence from individual people consisting of case information from those that were recently seen in EDs in the region. Note that from the ED data, we can use demographic information from the most recent U.S. Census to infer information about individuals who have not been recently admitted to the ED. The sets ***e*** and ***o*** can be described as:

$\boldsymbol{e} = \{X = e: X \in P_i, P_i \in \boldsymbol{P}\}$ and $\boldsymbol{o} = \{X = o: X \in O_j, O_j \in \boldsymbol{O}\}$

The goal of the algorithm is to calculate the probability of a disease outbreak given the OTC and the ED data. Mathematically, this objective is expressed as:

$$P(T \mid \boldsymbol{o}, \boldsymbol{e}) = k \cdot P(\boldsymbol{o}, \boldsymbol{e} \mid T) \cdot P(T), \qquad (1)$$

where the proportionality constant is

$$k = 1 \Big/ \sum_T P(\boldsymbol{o}, \boldsymbol{e} \mid T) \cdot P(T).$$

We can calculate the term $P(T)$ using Bayesian network (BN) inference on just the portion of the model that includes $\boldsymbol{G}$. Performing BN inference over just the nodes in $\boldsymbol{G}$ is much preferable to inference over all the nodes in X, because the number of nodes in X is approximately 10^7 in the current model. Since the set $\boldsymbol{I}$ renders the nodes in $\boldsymbol{P}$ (including $\boldsymbol{e}$) independent from the nodes in $\boldsymbol{G}$, we can derive the term $P(\boldsymbol{o}, \boldsymbol{e} \mid T)$ as follows:

$$P(\boldsymbol{o}, \boldsymbol{e} \mid T) = \sum_{\boldsymbol{i}} P(\boldsymbol{o}, \boldsymbol{e} \mid \boldsymbol{I} = \boldsymbol{i}) \cdot P(\boldsymbol{I} = \boldsymbol{i} \mid T) \qquad (2)$$

The above summation can be very demanding computationally, because $\boldsymbol{e}$ usually contains many nodes. We can factor the term $P(\boldsymbol{o}, \boldsymbol{e} \mid \boldsymbol{I} = \boldsymbol{i})$ as follows:

$$P(\boldsymbol{o}, \boldsymbol{e} \mid \boldsymbol{I}) = P(\boldsymbol{o} \mid \boldsymbol{e}, \boldsymbol{I}) P(\boldsymbol{e} \mid \boldsymbol{I})$$

When combined with Equation 2, we can consider the term $P(\boldsymbol{o} \mid \boldsymbol{e}, \boldsymbol{I})$ to be the conditional contribution of the OTC evidence to the posterior probability $P(T \mid \boldsymbol{o}, \boldsymbol{e})$, while the term $P(\boldsymbol{e} \mid \boldsymbol{I})$ can be considered as the conditional contribution of the ED evidence.

2.3.1 Incorporating ED Evidence

As described in (1), calculation of the term $P(\boldsymbol{e} \mid \boldsymbol{I})$ is done efficiently using *equivalence classes* and *incremental updating*. We save both space and reduce inference time by using equivalence classes to group individuals who are indistinguishable based on their evidence. Individuals in the same equivalence class have the same values for the *Home Zip, Age Decile, Gender, Respiratory Chief Complaint When Admitted,* and *ED Admission* nodes. Incremental updating dramatically reduces inference time by allowing us to avoid calculating $P(\boldsymbol{e} \mid \boldsymbol{I})$ for the entire population every time new ED data arrives.

2.3.2 Incorporating OTC Evidence

In order to incorporate OTC evidence into the posterior probability, we need to compute the probability $P(\boldsymbol{o} \mid \boldsymbol{e}, \boldsymbol{I})$. If for the purpose of this initial prototype we make the simplifying assumption that individuals living in a zip code only purchase OTC medications within their home zip code, then the OTC purchases for each zip code are independent of each other, conditioned on the nodes in $\boldsymbol{I}$. We also assume that the OTC purchases within a given equivalence class can be modeled as with a binomial distribution, and the distribution of OTC purchases within a given zip code can be modeled as the sum of independent binomial distributions of the equivalence classes within that zip code. Let $\boldsymbol{Z}$ be the set of all zip codes in the region under surveillance and let O_{Z_k} be the variable representing the OTC cough-medication sales volume for zip code Z_k. Furthermore, let the observed OTC cough-medication sales volume during the past three days for zip code Z_k be o_{Z_k}. The zip-code independence assumption allows us to factor the probability $P(\boldsymbol{o} \mid \boldsymbol{e}, \boldsymbol{I})$ as follows:

$$P(\boldsymbol{o} \mid \boldsymbol{e}, \boldsymbol{I}) = \prod_{Z_k \in \mathbf{Z}} P(O_{Z_k} = o_{Z_k} \mid \boldsymbol{e}, \boldsymbol{I}) \qquad (3)$$

In order to model the probability $P(O_{Z_k} = o_{Z_k} \mid \boldsymbol{e}, \boldsymbol{I})$, which corresponds to the probability of the OTC cough-medication data for zip code Z_k, we need to determine the contribution of the equivalence classes that belong to this zip code. Let Ω_{Z_k} be the set of equivalence classes that have *Home Zip* equal to Z_k. For clarity, we will assume that there is only one person model, namely the one shown in Figure 2, common to all the people in the population; in general, there can be as many person models as is useful to represent different types of people. An equivalence class Q_j is defined by a tuple $\boldsymbol{e}_j$, which is a (possibly incomplete) set of values for the evidence nodes in the person model. An example of such a

tuple is { *Home Zip=15213, Age Decile=2, Gender=Male, Respiratory Chief Complaint When Admitted=True, ED Admission=Today}.*

We model the OTC sales volume for each equivalence class Q_j as a binomial distribution with parameters n_j and p_j. The parameter n_j is simply the number of people currently within the equivalence class. The second parameter p_j represents the probability of an individual in the equivalence class making an OTC cough-medication purchase within the previous three days. This probability is calculated by conditioning on the evidence $\boldsymbol{e}_j$ that defines the equivalence class, and computing the probability that *Last 3 Days OTC Purchase* in Figure 2 equals *True* by performing Bayesian network inference on the person model.

We combine the distributions for each equivalence class in Ω_{Z_k} by using a normal approximation to the binomial distribution *(14)* to represent the OTC sales distribution for each equivalence class. The normal approximation is needed because there is no efficient way directly to derive the distribution over the sum of binomial variates. In contrast, it is straightfoward to derive the distribution over the sum of normal variates. With this approximation we can convert a binomial distribution with parameters n_j and p_j into a normal distribution with mean $n_j p_j$ and variance $n_j p_j (1-p_j)$. The distribution for the entire zip code is therefore represented as a normal distribution $N_{Z_k}(\mu_{Z_k}, \sigma^2_{Z_k})$ with mean μ_{Z_k} and variance $\sigma^2_{Z_k}$ that are as follows:

$$\mu_{Z_k} = \sum_{Q_j \in \Omega_{Z_k}} n_j p_j \quad \text{and} \quad \sigma^2_{Z_k} = \sum_{Q_j \in \Omega_{Z_k}} n_j p_j (1-p_j)$$

Finally, to derive the probability of observing the OTC sales for each zip code, we compute

$$P(O_{Z_k} = o_{Z_k} \mid \boldsymbol{e}, \boldsymbol{I}) = \int_{O_{Z_k}-0.5}^{O_{Z_k}+0.5} N_{Z_k}(x; \mu_{Z_k}, \sigma^2_{Z_k}) dx$$

3 RESULTS AND DISCUSSION

	Average initialization time (sec)	Average processing time for three-day window of data (sec)
ED Model	45.31	3
ED+OTC Model	209.73	3.91

Table 1: Mean running times for the ED and ED+OTC models, averaged over 100 simulation files.

We compare the average running times for the model from (1), which we will call the ED model, and the model described in this paper, which we will call the ED+OTC model, with both models operating on a Pentium-4 three Gigahertz machine with two Gigabytes of RAM. As shown in Table 1, the initialization time for the ED+OTC model is nearly four times that of the ED model. However, when actually processing the three-day window of data, the models take approximately the same amount of time. For example, suppose we run PANDA in real-time starting at t = 72 hours. The algorithm first performs an initialization phase then processes the data from t = 0 to t = 71 hours. When data accumulates for the next hour, PANDA moves its window of cases forwarded by an hour and analyzes the data from t = 1 to t = 72 hours. For each subsequent 72 hour window, the running time of PANDA only takes about four seconds. Even with the OTC extension, the new PANDA model is capable of processing all of the current data well before the next hour's worth of data arrives. These timing results provide support that the method is practical for real-time biosurveillance. We will be evaluating the false positive rate and detection time of our approach using data created by injecting simulated anthrax cases into existing ED and OTC data streams.

4 RELATED WORK

For an overview of algorithms used in syndromic surveillance, we refer the reader to (15). In this section, we will focus on the most relevant, representative work of which we are aware. Burkom (10) suggests two approaches for using the Spatial Scan Statistic (16) to combine multiple data sources in performing syndromic surveillance. The first method treats the multiple sources as covariates. The Spatial Scan Statistic is calculated using the sum of the observed counts from the data sources and the sum of the expected counts. One of the main problems of this approach is that a data source with a large count may mask data sources with smaller counts. As an alternative, Burkom proposes calculating the log likelihood ratio for each data source and summing these ratios to form the scan statistic, similar to the approach taken by Kulldorff (11). Burkom (9; 17) also suggests combining multiple univariate statistical process control methods using a technique such as Edgington's consensus method (18). However, Edgington's consensus method assumes independence among the data sources. In order to capture the correlation between data streams, Burkom (17) uses multivariate methods such as Hotelling's T^2 (19), MCUSUM (20) and MEWMA (21) on multiple univariate signals. In these multivariate methods, the covariance matrix for the data streams is typically estimated from a baseline period**.** If the covariance matrix changes significantly during an outbreak, then this estimate will not capture the actual relationship between the data streams during an outbreak.

5 CONCLUSIONS

This paper introduces a data fusion approach to biosurveillance that is based on modeling the effects of an outbreak disease on the individuals in the population. We extend the causal Bayesian network in (1) to incorporate evidence from both ED and OTC data by modeling the actions of individuals in terms of purchasing OTC products and visiting the ED. In addition, we

demonstrate that this data fusion model can process a three-day window of ED and OTC data in approximately four seconds, making it a feasible algorithm for real-time surveillance. In future work, we plan to extend the model to cover a two-week period and relax the zip code independence assumption. We then intend to perform a thorough and high fidelity evaluation of the detection algorithm that involves injecting simulated anthrax cases into actual ED and OTC data streams.

6 ACKNOWLEDGEMENTS

This research was supported by grants IIS-0325581 from the National Science Foundation, F30602-01-2-0550 from the Department of Homeland Security, and ME-01-737 from the Pennsylvania Department of Health.

7 REFERENCES

(1) Cooper GF, Dash DH, Levander JD, Wong WK, Hogan WR, Wagner MM. Bayesian Biosurveillance of Disease Outbreaks. In Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence. Banff, Canada: AUAI Press; 2004. 94-103 p. http://www.cbmi.pitt.edu/panda/publications.html

(2) Espino J, Wagner M, Szczepaniak C, Tsui F-C, Su H, Olszewski R et al. Removing a Barrier to Computer-Based Outbreak and Disease Surveillance: The RODS Open Source Project. Morbidity and Mortality Weekly Report 2004; 53 (Supplement 1):32-9.

(3) Wagner MM, Tsui F-C, Espino J, Hogan W, Hutman J, Hersh J et al. National Retail Data Monitor for Public Health Surveillance. Morbidity and Mortality Weekly Report 2004; 53 (Supplement 1):40-2.

(4) Hogan WR, Tsui F-C, Ivanov O, Gesteland PH, Grannis S, Overhage JM et al. Detection of Pediatric Respiratory and Diarrheal Outbreaks from Sales of Over-the-counter Electrolyte Products. Journal of the American Medical Informatics Association 2003; 10:555-62.

(5) Wagner M, Pavlin J, Brillman JC, Stetson D, Magruder S, Campbell M. Synthesis of Research on the Value of Unconventional Data for Early Detection of Disease Outbreaks DARPA, 2004 October.

(6) Labrie J. Self-care in the New Millennium: American Attitudes Towards Maintaining Personal Health.Consumer Healthcare Products Association, 2001.

(7) McIsaac WJ, Levine N, Goel V. Visits by Adults to Family Physicians for the Common Cold. J Fam Pract 1998; 47:366-9.

(8) Zeng X, Wagner MM. Modeling the Effects of Epidemics on Routinely Collected Data. Journal of the American Medical Informatics Association 2002; 9 (6 Supplement 1):S17-S22.

(9) Burkom H, Elbert Y, Feldman A, Lin J. Role of Data Aggregation in Biosurveillance Detection Strategies with Applications from ESSENCE. Morbidity and Mortality Weekly Report 2004; 53 (Supplement):67-73.

(10) Burkom HS. Biosurveillance Applying Scan Statistics with Multiple, Disparate Data Sources. Journal of Urban Health 2003; 80 (2 Supplement 1):i57-i65.

(11) Kulldorff M, Yih K, Kleinman K, Platt R, Mostashari F, Duczmal L. The Space-Time Permutation Scan Statistic for Multiple Data Streams. In Proceedings of the National Syndromic Surveillance Conference [CD-ROM]. Boston, MA; 2004.

(12) Hogan WR, Cooper GF, Wagner MM. A Bayesian Anthrax Aerosol Release detector. RODS Technical Report 2004.

(13) Wein LM, Craft DL, Kaplan EH. Emergency Response to an Anthrax Attack. Proceedings of the National Academcy of Sciences USA 2003; 100:4346-51.

(14) DeGroot MH. Probability and Statistics. 2nd ed. Reading, MA: Addison-Wesley; 1989.

(15) Wong W-K. Data Mining for Early Disease Outbreak Detection [Doctoral Dissertation]. Pittsburgh: Carnegie Mellon University; 2004.

(16) Kulldorff M. A Spatial Scan Statistic. Communications in Statistics: Theory and Methods 1997; 26 (6):1481-96.

(17) Burkom H, Coberly J, Murphy S, Elbert Y, Hurt-Mullen K. Public Health Monitoring Tools for Multiple Data Streams. In Proceedings of the National Syndromic Surveillance Conference [CD-ROM]. Boston, MA; 2004.

(18) Edgington ES. A Normal Curve Method for Combining Probability Values from Independent Experiments. Journal of Psychology 1972; 82:85-9.

(19) Hotelling HH. Multivariate Quality Control. In: C Eisenhart; MW Hastay; WA Wallis, editors, translator and editor Techniques of Statistical Analysis. New York: McGraw-Hill; 1947; p. 111-84.

(20) Crosier RB. Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. Technometrics 1988; 30 (3):291-303.

(21) Lowry CA, Woodall WH. A Multivariate Exponentially Weighted Moving Average Control Chart. Technometrics 1992; 34 (1):46-53.