Population-wide Anomaly Detection

Weng-Keen Wong, Gregory F. Cooper
RODS Laboratory,
University of Pittsburgh
Forbes Tower, Suite 8084
200 Lothrop Street
Pittsburgh, PA 15213
412-647-7113
{wwong,gfc}@cbmi.pitt.edu

Denver H. Dash
Intel Research
3600 Juliette Lane
Santa Clara, CA 95054
408-765-0410
denver.h.dash@intel.com

John D. Levander, John N. Dowling, William R. Hogan, Michael M. Wagner
RODS Laboratory,
University of Pittsburgh
550 Cellomics Building
100 Technology Drive
412-383-8134
jdl@cbmi.pitt.edu, dowling+@pitt.edu, {wrh,mmw}@cbmi.pitt.edu

ABSTRACT
Early detection of disease outbreaks, particularly an outbreak due to an act of bioterrorism, is a critically important problem due to the potential to reduce both morbidity and mortality. One of the most lethal bioterrorism scenarios is a large-scale release of inhalational anthrax. The Population-wide Anomaly Detection and Assessment (PANDA) algorithm [1] is specifically designed to monitor health-care data for the onset of an outbreak caused by an outdoor, airborne release of inhalational anthrax. At the heart of the PANDA algorithm is a causal Bayesian network which models the effects of the outbreak on a population. The most unique aspect of the PANDA algorithm is an approach we will refer to as *population-wide anomaly detection* in which each individual in the population is represented as a subnetwork of the overall causal Bayesian network. This paper will describe the benefits of the population-wide approach used by PANDA, which include a coherent way to incorporate background knowledge as well as different types of evidence, the ability to combine multiple data sources indicative of an outbreak, and the capability to identify the evidence that contributes the most to the belief that an anthrax outbreak is occurring.

Keywords
Anomaly Detection, Syndromic Surveillance, Biosurveillance, Bayesian Networks

1. INTRODUCTION
Early detection of disease outbreaks is a critically important problem due to the potential to reduce both morbidity and mortality. Disease outbreaks can either occur naturally or they can be caused by acts of bioterrorism. One of the most lethal bioterrorism scenarios is a large-scale release of inhalational anthrax, which is estimated to kill as many as 30,000 people per day and to have a long-term economic cost of as much as $200 million per hour of the outbreak according to an analysis done by [2]. The Population-wide Anomaly Detection and Assessment (PANDA) algorithm [1] is specifically designed to monitor health-care data for the onset of an outbreak caused by an outdoor, airborne release of inhalational anthrax. At the heart of the PANDA algorithm is a causal Bayesian network[1] which models the effects of the outbreak on a population. The most unique aspect of the PANDA algorithm is an approach we will refer to as *population-wide anomaly detection* in which each individual in the population is represented as a subnetwork of the overall Bayesian network.

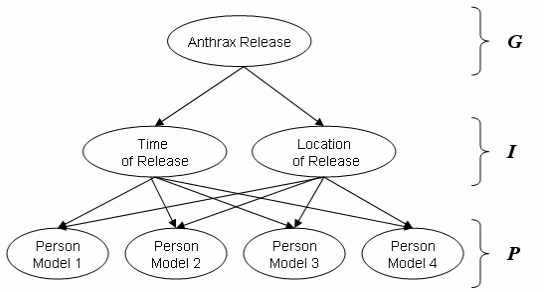

Figure 1: The causal Bayesian network structure used in PANDA.

Figure 1 provides an illustration of the causal Bayesian network structure used in PANDA to detect an outdoor release of inhalational anthrax. This model consists of three sets of nodes which we have labeled ***G***, ***I*** and ***P*** in the diagram. The nodes in the set ***G*** consist of global features that are common to all people. Included in the set ***G*** is a target node *T* which is the node that is actively monitored. In Figure 1, the target node is *Anthrax Release* and we monitor the probability that *Anthrax Release* equals true. In our example, *Anthrax Release* is the only global node. In general, the set ***G*** could include other global features such as the national terror alert level or information about local events such as major sports events or political conventions. At the next layer down, the interface nodes in set ***I*** are the nodes which contain the factors that significantly influence the status of an outbreak disease in individuals in the population. Inhalational anthrax is an infectious but non-contagious disease; the bio-agent can only infect people through the spores and not through person-to-person contact. As a result, the state of the disease in the population can be reasonably modeled with the nodes *Time of*

[1] A causal Bayesian network is a Bayesian network in which the arcs have a causal interpretation in addition to indicating a probabilistic relationship between the nodes.

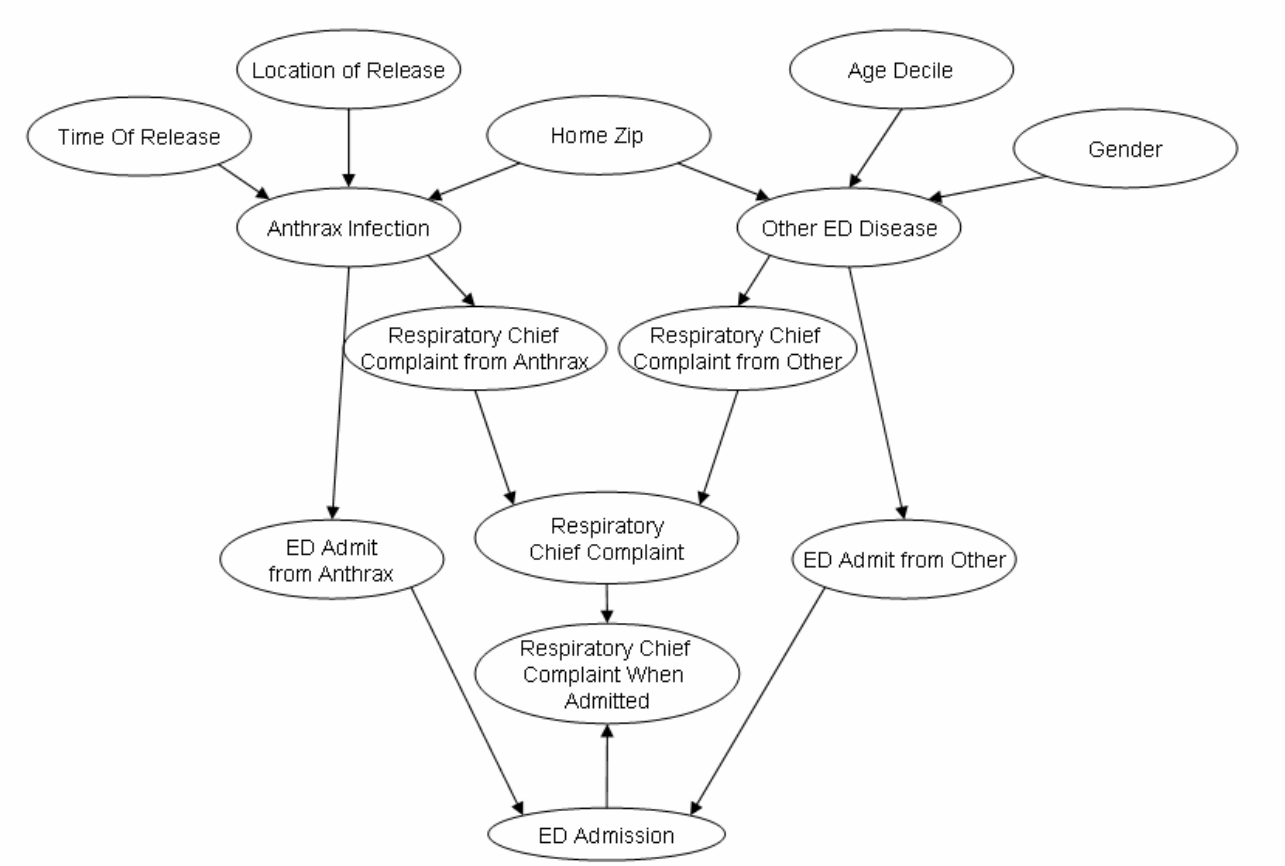

Figure 2: The person model used by PANDA to model each individual in the population.

Release and *Location of Release*. In the future, we plan to refine the model to include other interface nodes such as the amount of the release, the type of anthrax powder and meteorological information. Often the variables in $\boldsymbol{I}$ will be unmeasured. It is legitimate, however, to have measured variables in $\boldsymbol{I}$. For example, the wind direction might be a measured variable that influences the disease status of people in the population, and thus it would be located in $\boldsymbol{I}$. The last set of nodes $\boldsymbol{P}$ consists of the *person models* P_i *ie.* $\boldsymbol{P} = \{P_1, \ldots, P_n\}$ which form the core of the population-wide approach. Although we refer to these subnetworks as a *person* model, it can be generalized to entities that provide information about disease outbreaks, such as biosensors and livestock. Each person model P_i represents an individual in the population. In general, each person model can be different but for simplicity, we will use the model shown in Figure 2 for each P_i. Evidence observed on an individual basis will be entered at the person model level. In our implementation of PANDA, the information observed for each individual consists of Emergency Department (ED) records. Each ED record contains the attributes Home Zip, Age (by decile), Gender, Respiratory Chief Complaint When Admitted to the ED, and ED Admission.

The structure of the Bayesian network used by PANDA is designed by expert judgment rather than learned from data. The parameters of our model are obtained from a combination of census data, training data consisting of one year's worth of ED records, and expert assessments informed by the literature. With the structure and parameters of the model in place, we can perform inference on the Bayesian network to calculate the probability of an anthrax release. We emphasize that the population-wide anomaly detection approach is only used in designing the structure of the model and in the inference phase. We do not estimate parameters for an individual in the population. Rather, we observe evidence regarding an individual from an ED record and then propagate the effects of that evidence through the Bayesian network in order to update our belief that an anthrax attack is occurring.

At first glance, this approach appears to be intractable since the resulting model will consist of millions of nodes. For example, a typical surveillance region such as Allegheny County, Pennsylvania consists of 1.4 million people. Inference on a network of this scale seems intractable. However, in a previous paper [1], we have shown that such an approach is indeed feasible for a real-time bio-surveillance application that monitors Emergency Department (ED) data. In our initial prototype in [1], we exploit the conditional independence structure of the causal Bayesian network to produce two optimizations: incremental updating and equivalence classes. Incremental updating dramatically reduces inference time by allowing us to calculate probabilities for the entire population incrementally rather than from scratch whenever new data arrives. We also exploit the fact that individuals with exactly the same evidence are indistinguishable under the PANDA model. Individuals with the same values for the *Home Zip*, *Age Decile*, *Gender*, *Respiratory Chief Complaint When Admitted,* and *ED Admission* nodes are placed into the same equivalence class. In our surveillance of

Allegheny County, this optimization reduces the population of 1.4 million people to 24,240 equivalence classes. On a Pentium 4, 3 Gigahertz processor with 2 Gigabytes of RAM, the PANDA algorithm takes approximately 45 seconds of initialization time; after initialization, each hour of ED data can be processed in about 3 seconds.

Grouping individuals in equivalence classes may seem to contradict our claim of modeling each individual in the population. However, the use of equivalence classes is purely for computational convenience. We are indeed representing each person in the population and we are still capable of doing so without equivalence classes, albeit at a higher computational price. In future work, we intend to incorporate more information regarding the symptoms exhibited by patients in the ED. Adding this information will increase the number of features that define an equivalence class and consequently increase the number of equivalence classes beyond the number of people in the population. We plan to replace the use of equivalence classes with other optimizations such as approximate inference in order to make future extensions of the PANDA algorithm tractable.

Having addressed the most obvious downfall to population-wide anomaly detection, we will now discuss its advantages. Intuitively, it is the individuals in the population that generate the observed evidence. Thus, the most logical unit in the model is the individual, which is the finest level of granularity permitted by the data. With the modeling unit of an individual, we can exploit the conditional independence between individuals for a non-contagious disease to make inference tractable. As shown in Figure 1, if we condition on the time and location of the anthrax release, then the person models in the population are independent of each other. Another advantage gained by modeling each individual in the population is the ability to distinguish arbitrary groups from each other. This ability buys us a tremendous about of representational flexibility and power. In particular, we can coherently incorporate various forms of background knowledge and evidence into the model. Modeling at the individual level also facilitates combining information between multiple data sources, especially if the interaction between these data sources is much easier to model at an individual level than at a population level. Finally, the population-wide approach allows us to determine the contribution of each individual to the overall probability that an anthrax attack is occurring. We can determine the individuals that most influence this belief and in doing so, produce an explanation for why we believe an attack has occurred. The remainder of this paper will address these merits of a population-wide anomaly detection approach. We intend to provide an overview of this approach while leaving the details in previous papers on PANDA [1, 3].

2. INCORPORATING BACKGROUND KNOWLEDGE

One of the main advantages of a population-wide approach is the ability to coherently represent different types of background knowledge in the model. This background knowledge is particularly useful for disease outbreak detection algorithms that monitor for a specific disease; we will refer to these detection algorithms as *specific detectors*. In contrast, a *non-specific detector* such as WSARE [4] searches for any irregularities from the normal behavior. A strategy that works well for *non-specific detectors* is to model the baseline behavior of the data and signal an alert when the deviation from this baseline exceeds some threshold. However, since this strategy raises alarms for any irregularities rather than those caused by the disease being monitored, it can result in many false positives for a specific detector. We can improve the performance of specific disease detectors by building models of the data during non-outbreak periods and building models of the effects of the specific disease during outbreak periods.

Data during non-outbreak conditions are often available and in some cases abundant. The most common approach to building a baseline model is to use standard machine learning techniques such as Bayesian network structure learning [4] to learn the structure and/or the parameters of the model. Another option is to incorporate background knowledge of this baseline behavior into our model; for instance, in PANDA we use census information to model the demographics of the population. In contrast to data during non-outbreak periods, data during outbreak periods are scarce or completely non-existent. In the case of anthrax, there are only two commonly known anthrax outbreaks – an accidental leak in Sverdlovsk, Russia [5] and the 2001 postal anthrax attacks [6-9], although the postal attacks are clearly not representative of the large-scale outdoor release of anthrax that the PANDA algorithm is intended to detect. We cannot learn a model of an anthrax outbreak from data because do not have training data available from both of these incidents.

Nevertheless, we can incorporate the assessments of domain experts who are informed by their experience and the literature. In addition to studies performed on the two known outbreaks, there is a vast body of medical literature that allows us to model what we know about the likely patterns of presentation of inhalational anthrax [10-13]. In particular, we can model the known progression of symptoms that occur after a person has inhaled anthrax spores. We can also represent the incubation period, which is the earliest period of time after infection that a person begins to physically manifest the symptoms of anthrax (the incubation period varies depending on the concentration of spores released and the amount inhaled by an individual). Finally, in the case of an airborne release of anthrax, we can model the spatial dispersion pattern of the spores as in [14, 15], enabling the detection algorithm to know that a person standing downwind in the dispersion region is more likely to be infected than someone who is standing upwind. We can coherently incorporate all of this information in the parameters of the causal Bayesian network as background knowledge. Most importantly, the majority of the background knowledge about inhalational anthrax is at an individual level and it is precisely this background knowledge that we intend to use to improve our detector.

3. INCORPORATING DIFFERENT TYPES OF EVIDENCE

Besides the power in representing different forms of prior knowledge, modeling each individual allows the model to combine spatial, temporal, demographic, and symptomatic evidence to derive a posterior probability of a disease outbreak. For instance, if many people are admitted to the ED with symptoms consistent with those of inhalational anthrax and their home locations follow roughly the spatial dispersal pattern of an

airborne anthrax release, then the posterior probability of an anthrax attack should be high. Furthermore, individual modeling permits new types of knowledge and evidence to be readily incorporated into the model. We had previously assumed that the person models are identical for the purpose of simplicity but we can easily incorporate different person models into our framework. If we know more information about one person or group of people than another, we can represent that difference. 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 readily add this new evidential variable to the person model representing those individuals.

4. DATA FUSION

Modeling each individual in the population also facilitates fusion of different data sources, because such data originate from the individuals in the population that are being explicitly modeled. In [3] we extended the PANDA model to incorporate evidence from both ED data and from over-the-counter (OTC) data. By jointly monitoring both data sources, the combined information could reinforce our belief that an anthrax outbreak is happening and improve the detection algorithm's performance. However, 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 anthrax attack on these data sources during the same time period. Although training data do not exist, we do have some background knowledge at the individual level about the plausible relationship between OTC and ED data during an anthrax outbreak. Our approach to combining multiple data streams relies on using this background knowledge and explicitly modeling the actions of individuals that result in the interaction between OTC medication purchases and ED admissions.

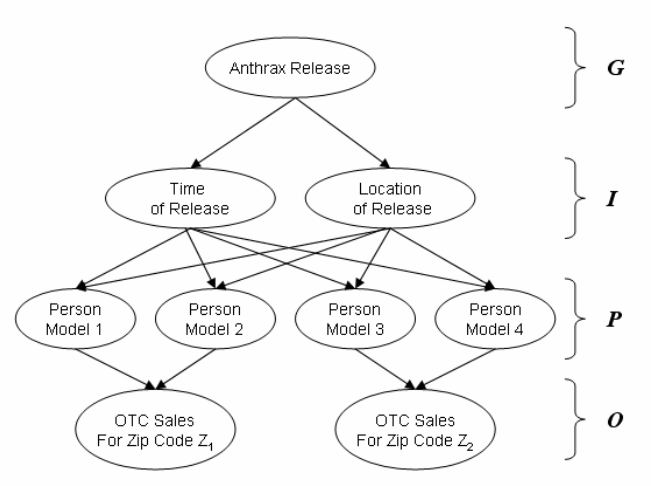

Figure 3: The causal Bayesian network used to combine ED data and OTC data.

Another concern in data fusion is incorporating data sources of different spatial and temporal granularity. For example, ED data is available in real-time (although we process it as if it were available hourly) with each record corresponding to an individual. On the other hand, the OTC data is available at the end of each day and each record aggregates the OTC sales over a zip code. The population-wide approach models the data at the level of an individual, which is the finest granularity that makes sense and that is permissible though the data. With this level of granularity, we can always aggregate individuals to form a coarser level of granularity while taking full advantage of all the information available.

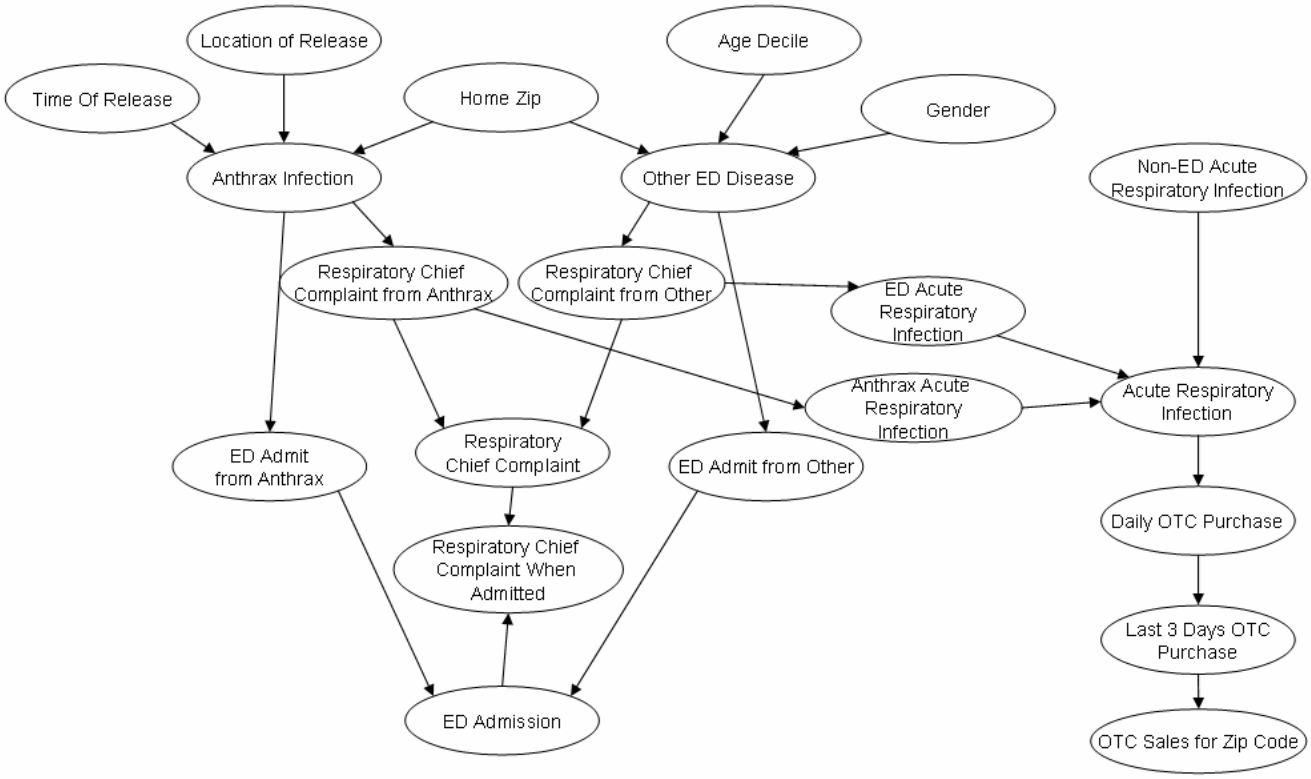

Figure 4: The person model for the PANDA algorithm that combines both ED data with OTC data.

Figure 3 illustrates the extension to the model in Figure 1 while Figure 4 depicts the modifications to the person model in Figure 2. The new causal Bayesian network incorporates the OTC evidence in the set of population-wide evidence nodes ***O***. The set ***O*** represents evidence that is aggregated over a particular group of people, such as the daily OTC sales of cough medication sales over a zip code.

5. EXPLANATION

With a population-wide anomaly detection algorithm, we can not only detect anomalies but also explain why they are anomalies. Using the Bayesian network framework, we can find the subset of evidence E^* that most influences the target node T. Once this subset of evidence is found, we can trace the pathways between E^* and T that contribute the most to the belief that an attack is occurring. In the current PANDA model, E^* corresponds to evidence about individuals. We can determine the individuals whose evidence most supports the hypothesis of an anthrax attack. Once these individuals have been identified, we can determine the relationships between them, which can potentially identify the origin and subsequent spread of the anthrax release. In our current prototype, we group the individuals into equivalence classes defined by the evidence observed in the data. Thus, we can identify the equivalence class that most supports the hypothesis of an anthrax attack. We have also used this explanation method to identify the zip code that is the most likely location of the release and the day that is the most likely time of release.

6. CONCLUSION

We have approached the task of detecting a large-scale airborne release of inhalational anthrax with a population-wide anomaly detection algorithm. This method has been ideally suited for this task due to the various forms of background knowledge and evidence that need to be incorporated into the model. In addition, if an alert is raised over a possible anthrax release, we gain the capability to explain why the alarm was triggered. The results reported in [1] have been promising and indicate that modeling each individual is feasible for a real-time bio-surveillance system. We also believe that the merits of this approach can benefit anomaly detection tasks in other domains.

7. 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.

8. REFERENCES

[1] Cooper, G.F., et al. *Bayesian Biosurveillance of Disease Outbreaks*. in *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*. 2004. Banff, Canada: AUAI Press.

[2] Kaufmann, A., M. Meltzer, and G. Schmid, *The economic impact of a bioterrorist attack: Are prevention and postattack intervention programs justifiable?* Emerging Infectious Diseases, 1997. 3(2): p. 83-94.

[3] Wong, W.-K., et al. *Bayesian Biosurveillance Using Multiple Data Streams.* in *Proceedings of the Third National Syndromic Surveillance Conference*. 2004. Boston, MA.

[4] Wong, W.-K., et al. *Bayesian Network Anomaly Pattern Detection for Disease Outbreaks*. in *Proceedings of the Twentieth International Conference on Machine Learning*. 2003: AAAI Press.

[5] Meselson, M., et al., *The Sverdlovsk anthrax outbreak of 1979.* Science, 1994. 266(5188): p. 1202-1208.

[6] Barakat, L.A., et al., *Fatal inhalational anthrax in a 94-year-old Connecticut woman.* Journal of the American Medical Association, 2002: p. 287-868.

[7] Inglesby, T.V., et al., *Anthrax as a biological weapon, 2002: updated recommendations for management.* Journal of the American Medical Association, 2002. 287(17): p. 2236-2252.

[8] Jernigan, J.A., et al., *Bioterrorism-related inhalation anthrax: the first 10 cases reported in the United States.* Emerging Infectious Diseases, 2001. 7(6): p. 933-944.

[9] Mayer, T.A., et al., *Inhalational anthrax due to bioterrorism: would current Centers for Disease Control guidelines have identified the 11 patients with inhalational anthrax from October through November 2001?* Clinical Infectious Diseases, 2003. 36(10): p. 1275-1283.

[10] Brachman, P.S., *Inhalation anthrax.* Annals of the New York Academy of Science., 1980. 353: p. 83-93.

[11] Franz, D.R., et al., *Clinical recognition and management of patients exposed to biological warfare agents.* Journal of the American Medical Association, 1997. 278: p. 399-411.

[12] Penn, C.C. and S.A. Klotz, *Anthrax pneumonia.* Seminars in Respiratory Infections, 1997. 12(1): p. 28-30.

[13] Shafazand, S., et al., *Inhalational anthrax: epidemiology, diagnosis, and management.* Chest, 1999. 116: p. 1369-1376.

[14] Hogan, W.R., G.F. Cooper, and M.M. Wagner, *A Bayesian anthrax aerosol release detectors.* RODS Technical Report, 2004.

[15] Wein, L.M., D.L. Craft, and E.H. Kaplan, *Emergency response to an anthrax attack.* Proceedings of the National Academy of Sciences USA, 2003. 100: p. 4346-4351.