2006 Proceedings of American Statistical Association, Statistical Computing Section [CD-ROM], Alexandria, VA:American Statistical Association:pp–pp; SAND2006-5431C, unclassified, unlimited release

Estimating Bioterror Attacks from Patient Data : A Bayesian Approach

J. Ray[1], Y. M. Marzouk[1], H. N. Najm[1], M. Kraus[2] and P. Fast[3]
Sandia National Laboratories, Livermore, CA, 94550-0969[1]
NORAD-NORTHCOM, 250 Vandenberg St, Peterson AFB, CO 80914-3812[2]
Lawrence Livermore National Laboratory, Livermore, CA 94550[3]

Abstract

Bioterrorist attacks involving an undetected release of an aerosolized pathogen have gained credibility and significance in national security, especially after the anthrax attacks of 2001. We address the problem of characterizing the release—i.e., inferring the number of index cases, the time of infection and the dosage of the pathogen—exclusively from clinical/patient data. We formulate this as a Bayesian inverse problem and develop probability density functions for the unknowns, conditioned on a short (3–4 day) time-series of diagnosed patient data. We assume (1) a single-focus epidemic, (2) identical dosages for all the infected people, and (3) a non-contagious disease, i.e., anthrax. The method is tested against simulated epidemics and the anthrax outbreak of Sverdlovsk in 1979. We also examine the impact of the data collection frequency on the quality of the estimates.

Keywords: Bayesian inference, bioterrorism, anthrax, clinical data

1 Introduction

Criminal releases of aerosolized pathogens may not always be detected via environmental sensors. Examples include small releases which do not travel far, releases where the formulation is coarse (and heavy) enough to precipitate quickly, and releases in areas which are not well-instrumented with sensors. In such cases, the first intimation of the attack will be the definitive diagnosis of the first patient, but by then the pathogen may have already established itself in the population. The ability to infer the characteristics of the release then plays an important part in formulating a medical response. The inferred characteristics can also serve as the initial condition for various epidemic models which can then be used to predict the evolution and spread of the disease in a population as well as its ramifications on society. In case of such an attack, especially with a non-contagious disease such as anthrax, this involves estimating the number of index cases, the time of the attack, and the doses received by the infected people.

Drawing these inferences can be challenging; one can only exploit the distribution of the incubation period of the disease, which in some cases is dependent on the dose received. To be relevant in an operational, consequence management sense, as opposed to forensics, inferences must be made early in the epidemic; a time-series of patient data, 3–4 days long, should thus be considered representative. Estimates are therefore expected to be rather uncertain, and quantifying this uncertainty becomes a key requirement of the inference process.

In this paper, we explore how such attacks may be characterized. This preliminary study targets single-focus attacks (i.e., a single release) with a non-contagious disease, anthrax. We also assume that the dose received by the index cases is uniform. We study how the inferences of the size, time and dosage behave with the time-resolution of the observed data by collecting data over 6- and 24-hour intervals. We adopt a Bayesian approach since it allows us to develop the inferred quantities as PDFs (probability density functions), thus quantifying uncertainty, and also allows a straightforward accommodation of additional information from disparate sources using prior distributions and sequential Bayesian learning. We operate within a self-imposed limit of 4 days of patient time-series. The results of this study will indicate whether more detailed questions, such as dose distributions and multiple attack foci, can be addressed satisfactorily by such an approach.

2 Previous Work

The exact question of estimating the size and time of an attack from a time-series of patient data has not been studied extensively. Walden & Kaplan [9] developed a Bayesian formulation and tested it on a low-dose anthrax attack corresponding, roughly, to the Sverdlovsk outbreak [5] of 1979. They also demonstrated the use of Bayesian *priors*—prior belief regarding the number N of people infected—to develop a smooth PDF for N, even for a small infected population ($N = 100$) and a 5-day time-series with data collected on a daily basis. Alternatively, a maximum likelihood method was employed by Brookmeyer & Blades [3] to infer the size of the infected population in the 2001 anthrax attacks in the US [4], preparatory to estimating the reduction of casualties by the timely administration of antibiotics. Both [3] and [9] developed similar expressions for the probability of observing a time-series of patients given a particular attack using the low-dose anthrax incubation model in [1].

A significant amount of work has focused on characterizing the incubation period of anthrax. Brookmeyer *et al* [1] developed a low-dose incubation period model applicable to the Sverdlovsk outbreak; their more recent work, based on a competing risks formulation, includes dose-dependence [2]. A more empirical approach, but based on significantly more data, was proposed recently by Wilkening [10]. He also compared four different models, including the dose-dependent model of Brookmeyer [2] (referred to as Model D); while Wilkening's Model A agreed with Model D at the high-dose limit, their low-dose behavior was different.

In this paper, in Sec. 3, we adapt the formulation in [9] to interval-aggregated data. Tests are in Sec. 4 and conclusions in Sec. 5.

3 Formulation of the problem

Consider an attack at time τ where N people are infected, with each of the N people receiving the same dose of D anthrax spores. The incubation period obeys a dose-dependent log-normal distribution; we refer to its cumulative distribution function (CDF) as $C(t, D)$. For a few days M (say 3–5 days) we can expect (1) a series $t_i, i = 0 \dots M$, of times, perhaps the endpoints of 24-hr intervals, when patients' symptoms are observed and (2) the time-series $n_i, i = 0 \dots M$, of new patients who turned symptomatic between $t_i - \Delta t$ and t_i where $t_i - t_{i-1} = \Delta t, i \neq 0$, and Δt is a constant. We define survival probability as $P_{surv}(t, D) = 1 - C(t, D)$.

We can state the problem as such: Given a time series $(t_i, n_i), i = 0 \dots M$, of patients showing symptoms over a few days M, estimate (N, τ, D) from these data. n_i patients are assumed to have developed symptoms over the time interval between t_{i-1} and t_i.

Below we reproduce from [8] the likelihood function $\mathcal{L}$—the probability of observing a $\{t_i, n_i\}$, $i = 0 \dots M$, series given a (N, τ, D) attack:

$$\begin{aligned}\mathcal{L}\left(N, \tau, D\right) &= \frac{N!}{(N-L)! \prod_{i=0}^{M} n_i!} \{P_{surv}(t_M - \tau, D)\}^R \\ & \prod_{i=0}^{M} \{C(t_i - \tau, D) - C(t_{i-1} - \tau, D)\}^{n_i}\end{aligned}$$

where $L = \sum_{i=0}^{M}$, $R = N - L$ and

$$C(t, D) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\ln(t/t_0)}{\sqrt{2}S}\right)\right], \qquad (1)$$

with $t_0 = 10.3 - 1.35 \log_{10}(D)$ and $S = 0.804 - 0.079 \log_{10}(D)$ [10].

By Bayes rule, the probability $\pi(N, \tau, D | \{t_i, n_i\}, i = 0 \dots M)$ of a (N, τ, D) attack, conditioned on the data, can be written as

$$\pi(N, \tau, D | \dots) \propto \mathcal{L}(N, \tau, D) \pi_N(N) \pi_\tau(\tau) \pi_D(D) \qquad (2)$$

where $\pi_N(N)$, $\pi_\tau(\tau)$ and $\pi_D(D)$ are the priors for N, τ and D. In the absence of additional information, we use broad uniform distributions as priors for all three parameters. The joint posterior $\pi(N, \tau, D | \dots)$ from (2) is then marginalized to obtain one-dimensional PDFs for N, τ and D. Each marginalization, which involves integrating out the effect of the two other variables, is performed using the VEGAS algorithm [7] for Monte Carlo integration, as encoded in the GNU Scientific Library (http://www.gnu.org/software/gsl/)

4 Test cases

We first consider a simulated anthrax attack where 10^4 people are infected with a dose of 10^2 spores each. The first patient exhibits symptoms 0.75 days after the attack. The time series of new patients exhibiting symptoms, collected over successive 24-hour intervals is $\{12, 187, 565, 899, 985\}$; the corresponding series collected over 6-hour intervals is $\{3, 9, 11, 36, 62, 78, 103, 125, 161, 176, 206, 212, 239, 242, 222, 243, 263, 257, 252, 253\}$. We will try to infer the correct values of (N, τ, D), which are $(10^4, -0.75, 10^2)$.

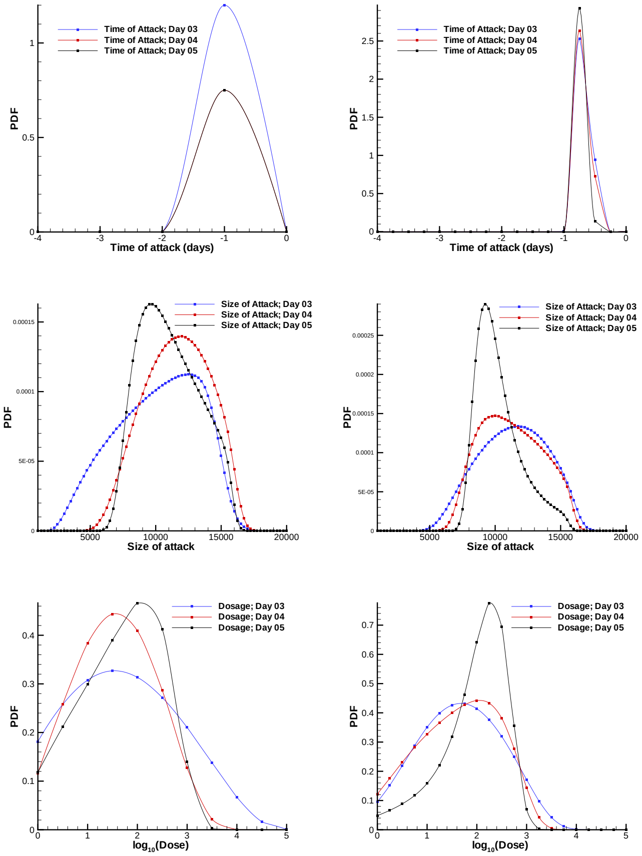

Figure 1: PDFs for N, τ and $\log_{10}(D)$ calculated from the time series of clinical data as generated by a (data) collection interval of 24 hours (left) and 6 hours (right). The higher resolution data, which captures the *structure* of the observables better, results in sharper PDFs.

In Fig. 1, on the left side, we plot the PDFs for N, τ and $\log_{10}(D)$ as calculated using the low-resolution (24-hour intervals) data; on the right are the corresponding plots from the high-resolution series. We see that the time of attack τ is easily inferred even with 3 days of data. However the low resolution time-series does not allow us to identify -0.75 days as the MAP (*maximum a posteriori*) estimate for τ due to coarseness of the resolution; instead the PDF peaks at -1.0 day. The PDFs for N calculated from the two time series are somewhat different in shape, but the MAP estimate for N by day 5 is around 10^4. The high resolution data provides a more peaked PDF, as expected. The dosage is harder to infer, but by day 5, the MAP estimate for $\log_{10}(D)$ is around 2. The higher resolution data results in sharper PDFs and tighter confidence intervals since it captures the *structure* of the observed data better. A more detailed analysis is in [8].

Next, we apply this technique to the Sverdlovsk anthrax outbreak of 1979 [5]. The cause is suspected to be an accidental release of aerosolized spores from a military facility on April 2nd, 1979 in the erstwhile Soviet Union. 70 people died and it is estimated that 80 were infected [10]. The first symptoms were exhibited on April 4th, i.e. $\tau = -2$ days. Around April 15th prophylaxis was distributed, suppressing symptoms and lengthening the incubation period. The time series of patients showing symptoms was obtained from [6]; data were collected on a daily basis. A low dose exposure is conjectured, and estimates in the literature vary between 2–300 spores [5, 10]. In Fig. 2 we develop PDFs for N and τ. The time of attack τ is again inferred easily, with the MAP estimate peaking at $\tau = -2$ within 3 days. The MAP estimate of N is less well behaved, peaking around 40 after 3–5 days and around 60 after 10 days. The dosage curve (omitted in this paper) reveals little except that the dosage was small; even after 10 days of data, no apparent peak is observed. However, given the small size of the infected population (80), the effect of prophylaxis (which is not captured in our model of the incubation period), and the noise in the data (which had to be reconstructed from grave markers), this statistical reconstruction of the event is remarkably similar to the best available analysis of the outbreak [5].

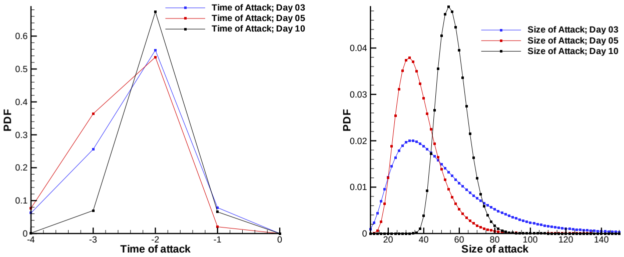

Figure 2: PDFs for N and τ as calculated from the time series of clinical data from the Sverdlovsk outbreak

The PDFs developed above clearly demonstrate the effect of data in reducing uncertainty. In Fig. 1, the PDFs from the 6-hour time series indicate tighter ranges for the inferred quantity, while in Fig. 2, the sudden shift in the inference of N towards the more accurate MAP estimate of 60 is accompanied by a reduction in uncertainty as shown by the steeper PDF.

5 Conclusions

We have developed a promising approach to reconstructing a bioterrorist attack purely from clinical data. Based on Bayesian inference, we see that very little data (3–4 days of observations) suffice to infer the size, time, and dosage received in such attacks. Improvements in the inference do not necessarily require more data from longer observation periods; instead, better resolved data—collected over 6 hour intervals, for instance—can tighten confidence intervals significantly. Nimble reporting protocols may prove sufficient in achieving this aim. Further, the inferences drawn here are "good enough" for operational purposes, e.g., to plan a response. In addition, the Bayesian construction allows for a straightforward incorporation of prior information in case one has additional knowledge affecting N, τ and/or D, such as via atmospheric dispersion modeling.

Acknowledgments

This work was funded as a Laboratory Directed Research and Development project. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

References

[1] R. Brookmeyer, N. Blades, M. Hugh-Jones, and D. A. Henderson. The statistical analysis of truncated data: application to the Sverdlovsk anthrax outbreak. *Biostatistics*, 2:233–247, 2001.

[2] R. Brookmeyer, E. Johnson, and S. Barry. Modelling the incubation period of anthrax. *Statistics in Medicine*, 24:531–542, 2005.

[3] Ron Brookmeyer and Natalie Blades. Statistical models and bioterrorism : Application to the U.S. anthrax attacks. *Journal of the American Statistical Association*, 98(464):781–788, 2003.

[4] John A. Jernigan et al. Bioterrorism-related innhalational anthrax: The first 10 cases reported in the United States. *Emerging Infectious Diseases*, 7(6):933–944, 2001.

[5] Matthew Meselson et al. The Sverdlovsk anthrax outbreak of 1979. *Science*, 266:1202–1208, 1994.

[6] Thomas V. Inglesby et al. Anthrax as a biological weapon - Medical and public health management. *J. Am. Med. Assoc.*, 281(18):1735–1745, 1999.

[7] G. Peter Lepage. A new algorithm for adaptive multidimensional integration. *Journal of Computational Physics*, 27:192–203, 1978.

[8] J. Ray, Y. M. Marzouk, H. N. Najm, M. Kraus, and P. Fast. A Bayesian method for characterizing distributed micro-releases: I. The single-source case for non-contagious diseases. SAND Report SAND2006-1491, Sandia National Laboratories, Livermore, CA 94551-0969, March 2006. Unclassified unlimited release.

[9] J. Walden and E. H. Kaplan. Estimating time and size of bioterror attack. *Emerging Infectious Diseases*, 10(7):1202–1205, 2004.

[10] D. Wilkening. Sverdlovsk revisted : Modeling human inhalational anthrax. *Proceedings of the National Academy of Science*, 103(20):7589–7594, May 2006.