Biostatistics (2001), **2**, 2, *pp.* 233–247
Printed in Great Britain

The statistical analysis of truncated data: application to the Sverdlovsk anthrax outbreak

RON BROOKMEYER*, NATALIE BLADES

Department of Biostatistics, Johns Hopkins University School of Public Health, Baltimore, MD 21205, USA

MARTIN HUGH-JONES

School of Veterinary Medicine, Louisiana State University, Baton Rouge, LA 70803, USA

DONALD A. HENDERSON

The Center for Civilian Biodefense Studies, and the School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

SUMMARY

An outbreak of anthrax occurred in the city of Sverdlovsk in Russia in the spring of 1979. The outbreak was due to the inhalation of spores that were accidentally released from a military microbiology facility. In response to the outbreak a public health intervention was mounted that included distribution of antibiotics and vaccine. The objective of this paper is to develop and apply statistical methodology to analyse the Sverdlovsk outbreak, and in particular to estimate the incubation period of inhalational anthrax and the number of deaths that may have been prevented by the public health intervention. The data available for analysis from this common source epidemic are the incubation periods of reported deaths. The statistical problem is that incubation periods are truncated because some individuals may have had their deaths prevented by the public health interventions and thus are not included in the data. However, it is not known how many persons received the intervention or how efficacious was the intervention. A likelihood function is formulated that accounts for the effects of truncation. The likelihood is decomposed into a binomial likelihood with unknown sample size and a conditional likelihood for the incubation periods. The methods are extended to allow for a phase-in of the intervention over time. Assuming a lognormal model for the incubation period distribution, the median and mean incubation periods were estimated to be 11.0 and 14.2 days respectively. These estimates are longer than have been previously reported in the literature. The death toll from the Sverdlovsk anthrax outbreak could have been about 14% larger had there not been a public health intervention; however, the confidence intervals are wide (95% CI 0–61%). The sensitivity of the results to model assumptions and the parametric model for the incubation period distribution are investigated. The results are useful for determining how long antibiotic therapy should be continued in suspected anthrax cases and also for estimating the ultimate number of deaths in a new outbreak in the absence of any public health interventions.

Keywords: Anthrax; Common source outbreak; Incubation period; Likelihood; Truncation.

*To whom correspondence should be addressed.

1. INTRODUCTION

In recent years anthrax has been the subject of increasing concern and attention because of its potential use as a biological weapon. The main concern is that an anthrax aerosol could be released that is odourless and invisible and could be disseminated widely causing significant number of cases of disease and deaths (Inglesby *et al.*, 1999). In response to these concerns, the US military initiated a vaccination programme against anthrax.

Anthrax is caused by the bacteria *Bacillus anthracis*. Transmission of anthrax infection occurs either through inhalation, ingestion or cutaneous exposure. The most common form of transmission, cutaneous, occurs through broken skin, typically when an individual butchers an anthrax-infected animal. Cutaneous anthrax is readily treated with antibiotics and is seldom fatal. Much less common is gastrointestinal anthrax, which occurs when an individual ingests undercooked contaminated meat. Inhalational anthrax occurs when an individual breathes anthrax spores into the lungs. These spores, borne in particles 1–5 μm in diameter, are carried to mediastinal lymph nodes where they may germinate into vegetative cells. Proliferation of the cells is accompanied by the release of toxins and the onset of symptoms in the host. The mortality rate from inhalational anthrax in the absence of treatment has been reported to be very high (Friedlander, 1997; Cieslak and Eitzen, 2000). The time from onset of symptoms to death is, on average, only a few days. Antibiotics are believed to be effective only against the vegetative cell shortly after germination. Inhalational anthrax was first reported in the mid-1800s among British wool sorters who were exposed to contaminated goat and alpaca hairs, and became known as wool sorter's disease (Brachman, 1980). Only 18 cases of inhalational anthrax have been reported in the United States from 1900 through 1978 and none since 1978 (Brachman, 1980; Inglesby *et al.*, 1999). Yet, inhalational anthrax poses the major threat as a biological weapon because the pathogen is widely available in nature, and stable in aerosol form, which makes it possible to disseminate the spores widely.

Knowledge of the incubation period of inhalational anthrax is important in the development of strategies for a public health response to an outbreak. First, information on the duration of the incubation period can be used to determine how long antibiotics should be given to exposed individuals. Second, information on the incubation period together with numbers of reported cases early in an outbreak could be used to determine the ultimate size of the outbreak. However, little is known about the natural history or incubation period of inhalational anthrax because of very little human experience with the disease and because only limited information is available from experimental monkeys (Friedlander *et al.*, 1993). Estimates of the incubation period of inhalational anthrax that have been cited in the scientific literature range from 1 to 7 days and have been based in part on the response of monkeys to high-dose aerosol exposure (Cieslak and Eitzen, 2000; Brachman and Friedlander, 1999; Benenson, 1995; Franz *et al.*, 1997).

In April 1979 an anthrax outbreak occurred in the city of Sverdlovsk in Russia about 900 miles east of Moscow. Originally the outbreak was attributed to the consumption of contaminated meat. However, it was subsequently concluded that the outbreak resulted from the inhalation of spores that were accidentally released from a military microbiology facility on April 2, 1979 (Meselson *et al.*, 1994). This conclusion was based on isolation of the organism and autopsy, and on pathological reports that documented inhalational anthrax as the cause of death. Interview information from relatives of cases determined that the cases lived or worked in a narrow zone extending from the military microbiology facility to the southern limit of the city and consistent with wind directions on April 2, 1979 (Guillemin, 1999). While there were less than 100 reported anthrax deaths from this outbreak, it is believed that there were several hundred additional non-fatal cases. On April 11, 1979, the epidemic was confirmed as anthrax. In the middle of April, health authorities began visiting all households with suspected cases to provide a five day supply of tetracycline for all contacts. Many others in the community also received antibiotics. There are, however, no records about who received antibiotics or for how long. During the period April

20–22, a large-scale vaccination programme was mounted during which persons of 17–59 years of age were vaccinated. About 80% of a target population of 59 000 was said to have been reached. The Russian vaccine, however, is a live, attenuated vaccine that must multiply in the host in order to produce immunity. A seven day period is said to be required for protection to develop. However, among those receiving antibiotic treatment, the organism would not multiply and little or no benefit would accrue from receiving the vaccine.

The objective of this paper is to develop and apply statistical methodology to data from the Sverdlovsk outbreak to estimate the incubation period of inhalational anthrax and to estimate the number of deaths that may have been prevented by the public health response to the outbreak. The data consist of the incubation periods among individuals who were reported to have died of anthrax. The main statistical problem is that the data on incubation periods are truncated. Truncation arises because some individuals who were exposed to the anthrax pathogen may have had their disease and death prevented because of the timely administration of antibiotics or immunization. Indeed, individuals with longer incubation periods are more likely to have received vaccine or antibiotics and thus their death could have been prevented and would not be included in the data set. Accordingly, a naive analysis that ignored the effects of truncation would lead to an underestimation of the incubation period. It is not known how many individuals would have died in the absence of the public health intervention.

The statistical analysis of truncated data in epidemiological studies has received increasing attention (Brookmeyer, 1998). For example, the first estimates of the incubation period distribution of AIDS were based on an analysis of transfusion-associated AIDS cases (Lui *et al.*, 1986; Medley *et al.*, 1987, 1988; Brookmeyer and Gail, 1988; Kalbfleisch and Lawless, 1989). This data set was truncated because it tended to include disproportionately many individuals with short incubation periods. Individuals with longer incubation periods may not yet have been diagnosed with AIDS and thus would tend to be excluded from the data set. Similar selection biases occur in the statistical analysis of the delays in disease reporting to public health registries (Brookmeyer and Liao, 1990; Wang, 1992) and in studies of pediatric AIDS (DeGruttola *et al.*, 1992). Various methods of analysis of this sort of data (e.g. transfusion-associated AIDS and disease reporting delays) are reviewed in Brookmeyer and Gail (1994). In these applications, truncation results from a limited time period for case ascertainment. In contrast, a limited time period of surveillance is not a cause of truncation in the Sverdlovsk data because disease surveillance in Sverdlovsk continued for many months after the release of the pathogen on April 2, 1979. Rather, the truncation in the Sverdlovsk data results from public health interventions that could prevent death. The analysis of the Sverdlovsk outbreak presents challenging new statistical issues because somewhat less information is available about the selection criteria by which individuals were included in the data set: for example, it is not known how many people in Sverdlovsk were exposed to the airborne anthrax, nor how many of these exposed people received any sort of public health intervention, and it is not known how effective such public health interventions were in preventing death.

The remainder of the paper is structured as follows. In Section 2 the available data are reviewed and some model assumptions are introduced. The statistical methodology and likelihood formulation are given in Section 3. The methods are applied to the Sverdlovsk outbreak in Section 4. The results are discussed in Section 5.

2. Model assumptions and data

We assume that all cases in the Sverdlovsk outbreak resulted from exposure to airborne anthrax spores during the immediate period following the initial release of the spores into the air on April 2, 1979 (i.e. a common source outbreak). This period is believed to last one day at most (Inglesby *et al.*, 1999). While it is possible that a secondary aerosol might account for some cases as a result of resuspension of spores

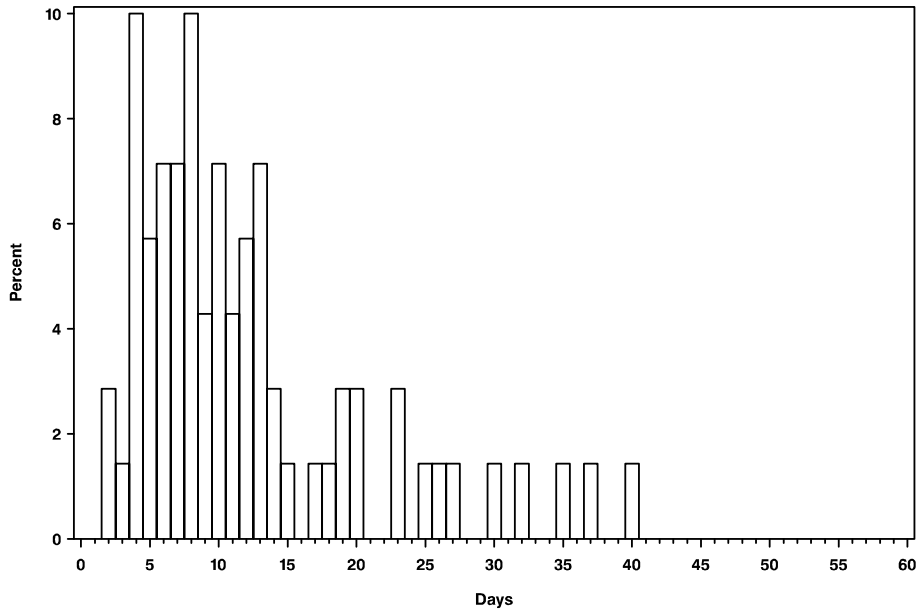

Fig. 1. Histogram of incubation periods from 70 lethal cases of the Sverdlovsk anthrax outbreak.

after they have fallen to the ground, this is very unlikely because considerable energy is required for resuspension of particles (Meselson *et al.*, 1994). Moreover, person-to-person transmission of anthrax does not occur (Pile *et al.*, 1998).

The dates of onset of symptoms of fatal cases with documented inhalational anthrax have been compiled in Meselson *et al.* (1994) and in Guillemin (1999). A number of additional fatal cases for which data are now available are also included in this analysis. Figure 1 is a histogram of the incubation periods of 70 cases who ultimately died of anthrax. The incubation period refers to the time between April 2, 1979 and the onset of symptoms. The data in Figure 1 refer only to cases that ultimately died of the disease. While there are reports of many additional cases that did not ultimately die from the disease, it is believed that there is considerable under-reporting of the number of non-lethal anthrax cases in the Sverdlovsk outbreak. In this report, we focus on the analysis of the data in Figure 1 on lethal cases of anthrax. As described in Section 1, this is a truncated sample of incubation times because individuals who were exposed to the anthrax pathogen but who recovered with or without public health intervention are excluded from the data set.

We suppose that in the absence of public health interventions that could prevent death from the disease, individuals exposed to the anthrax pathogen are a mixture of a proportion p of individuals who would die from anthrax infection and a proportion $1 - p$ of individuals who would not die from disease. Individuals who are exposed to the pathogen may not die either because they have not been exposed to a sufficient dose of the pathogen or because of naturally occurring immunity. Suppose that in the absence of public health intervention, N individuals would have died from the anthrax outbreak. The 70 deaths reported in Figure 1 may well be less than N because some deaths may have been prevented by the antibiotic and vaccination programmes. We also define the incubation period distribution of lethal anthrax, $F(t)$, to refer to the cumulative distribution function of incubation periods (time from exposure to disease onset) for these N individuals. The distribution function $F(t)$ is a proper distribution function and refers to the probability that an incubation period is less than t days among the N individuals who would eventually die without

effective medical intervention. Our goal is to estimate N and $F(t)$ from the data in Figure 1. An estimate of N can be used to determine how many deaths there could have been had there not been any intervention, and thus the number of deaths that may have been prevented. The incubation period distribution for lethal anthrax, $F(t)$, can be used to address a number of important public health questions. For example, $F(t)$ may be helpful in determining the duration of time to administer post-exposure antibiotics. Further, as discussed in Section 5, $F(t)$ together with the numbers of deaths early in an outbreak can be used to estimate the total number of deaths from an outbreak in the absence of effective medical intervention.

Our analysis assumes that the effect of a public health intervention, if any, is to prevent some deaths that ordinarily would have occurred but that the intervention does not alter the incubation distribution among deaths that are not prevented. This assumption certainly holds for interventions that are applied after onset of symptoms, and is discussed further in Section 5. As discussed below, our model assumes that a random sample of unknown size of the N individuals who are still at risk (alive) when the intervention programme begins, will have their deaths prevented by medical intervention.

3. Statistical analysis

3.1 *Likelihood formulation*

We estimate both $F(t)$ and N from the data in Figure 1 using methodology that acknowledges the possible truncation of the data. Suppose a common source epidemic occurs at calendar time 0 and that a public health intervention to prevent disease among exposed individuals is initiated at calendar time C. For notational convenience we centre the origin of calendar time at the start of the outbreak. Let X_1 refer to the number of cases with dates of onset of symptoms prior to calendar time C and let X_2 refer to the number of cases with dates of onset of symptoms after calendar time C. We assume that X_1 is a complete count of all cases diagnosed prior to calendar time C. In the development that follows, we develop an *instantaneous intervention model* that assumes the intervention was applied exactly at time C. Specifically, this model assumes that all individuals with (potential) incubation periods greater or equal to C had an equal probability of receiving an intervention at time C that could have prevented disease. In Section 3.2, we extend the methods to allow for a phase-in of the intervention, beginning at calendar time C.

The data consist of the incubation times of the $X_1 + X_2$ individuals in Figure 1. Let $t_1 = \{t_{1i}; i = 1, \ldots, X_1\}$ represent the incubation periods for the X_1 cases with incubation periods less than C. Let $t_2 = \{t_{2i}; i = 1, \ldots, X_2\}$ represent the incubation periods for the X_2 cases with incubation periods greater or equal to C. We assume a parametric model for $F(t)$ that involves some unknown parameter vector θ. Thus, the parameters of our model include θ and N, the number of exposed individuals who would eventually succumb to the infection in the absence of intervention. The parameter N is equal to $X_1 + X_2$ plus the number of exposed individuals whose disease was prevented by the intervention.

The observed data consist of the vectors of incubation times t_1 and t_2 along with the dimensions of these vectors, namely X_1 and X_2 respectively. The likelihood function of the observed data is $L = L(N, \theta; t_1, t_2, X_1, X_2)$. Typically, a likelihood is formulated by considering the sample size as fixed by design rather than as a random variable. Here, we do not consider X_1 and X_2 as fixed because they are random variables whose distributions involve parameters of interest (i.e. N and θ). We decompose the likelihood L into a product of two factors using conditional probabilities as follows. The first factor is the probability distribution of X_1 and X_2 and is called $L_1 = L_1(N, \theta; X_1, X_2)$. The second factor is the probability density of t_1 and t_2 conditional on X_1 and X_2, and is called $L_2 = L_2(\theta; t_1, t_2|X_1, X_2)$, which depends only on the parameter θ. Then, we can write L as

$$L(N, \theta; t_1, t_2, X_1, X_2) = L_1(N, \theta; X_1, X_2) \times L_2(\theta; t_1, t_2|X_1, X_2). \quad (1)$$

The first factor $L_1(N, \theta; X_1, X_2)$ is a binomial likelihood with N trials, where N is an unknown parameter, X_1 observed 'failures' (i.e. incubation periods $< C$) and $N - X_1$ 'successes' (i.e. incubation periods $\geqslant C$). Thus

$$L_1(N, \theta; X_1, X_2) = \frac{N!}{X_1!(N - X_1)!}[F(C)]^{X_1}[1 - F(C)]^{N-X_1} \tag{2}$$

for $N \geqslant X_1 + X_2$ and $L_1 = 0$ for $N < X_1 + X_2$.

The second factor in equation (1), $L_2(\theta; t_1, t_2|X_1, X_2)$, is the likelihood of observing incubation times t_1 in a random sample of X_1 persons with incubation times $< C$ and incubation times t_2 in a random sample of X_2 persons with incubation times $\geqslant C$. Thus

$$L_2(\theta; t_1, t_2|X_1, X_2) = \prod_{i=1}^{X_1} \frac{f(t_{1i})}{F(C)} \prod_{i=1}^{X_2} \frac{f(t_{2i})}{1 - F(C)} \tag{3}$$

where $f(\cdot)$ is the incubation period probability density corresponding to the CDF $F(\cdot)$. Equation (3) assumes that the X_2 observations with incubation times $\geqslant C$ are a random sample from the density $f(t)/\{1 - F(C)\}$. This is justified if a random sample of the individuals with incubation periods $\geqslant C$ are truncated at time C. The implicit assumption is that the probabilities of truncation do not depend on the individuals' potential (possibly unobserved) incubation periods. Substituting expressions (2) and (3) into (1), the likelihood L becomes

$$L = \frac{N!}{(N - X_1)!X_1!}[F(C)]^{X_1}[1 - F(C)]^{N-X_1} \prod_{i=1}^{X_1} \frac{f(t_{1i})}{F(C)} \prod_{i=1}^{X_2} \frac{f(t_{2i})}{1 - F(C)}. \tag{4}$$

Expression (4) for L simplifies to

$$L = \frac{N!}{(N - X_1)!X_1!}[1 - F(C)]^{N-X_1-X_2} \prod_{i=1}^{X_1} f(t_{1i}) \prod_{i=1}^{X_2} f(t_{2i}). \tag{5}$$

Expression (5) provides additional insight into the likelihood L. For a fixed value of N, expression (5) is the likelihood function for a sample of right-censored survival times with $X_1 + X_2$ observed failures at times t_1 and t_2, and with $N - X_1 - X_2$ observations right-censored at time C. The connection between expression (5) and the 'usual' likelihood for right-censored survival data for fixed N can be made transparent by noting that L is proportional to

$$L \propto [1 - F(C)]^{N-X_1-X_2} \prod_{i,j} f(t_{ij}).$$

Thus, for a fixed value of N, expression (5) can be maximized over θ, the parameters of F, using computing software for right-censored survival data to obtain estimates $\hat{\theta}(N)$. Then, a line search can be performed over values of N to determine the value $\hat{N}$ that maximizes L. Confidence intervals for N can be determined by inverting a likelihood ratio test. Specifically, a $(1 - \alpha)$ confidence interval for N consists of all N such that $2[\log L(\hat{N}, \hat{\theta}(\hat{N})) - \log L(N, \theta(N))] < \chi^2_\alpha(1)$ where $\chi^2_\alpha(1)$ is the α-level critical value of a χ^2 distribution with one degree of freedom.

An alternative approach for maximizing L (shown in the Appendix) for N large is as follows. First, we obtain an estimate of θ, called $\hat{\theta}$, by maximizing L_2 in expression (3) over θ. Then we set $\hat{N} = X_1/\hat{F}(C)$ where $\hat{F}(C) = F(C; \hat{\theta})$ is the cumulative distribution function with $\hat{\theta}$ substituted for θ and evaluated at time C. Thus, all the information in the likelihood L about the parameters θ of the incubation period distribution resides in L_2.

We can obtain confidence regions for θ by inverting a likelihood ratio test. Specifically, a $(1-\alpha)$ confidence region for θ consists of all θ such that $2[\log L_2(\hat{\theta}; t_1, t_2|X_1, X_2) - \log L_2(\theta; t_1, t_2|X_1, X_2)] \leqslant \chi^2_\alpha(d)$ where χ^2_α is the α-level critical value of a χ^2 distribution with d degrees of freedom where d is equal to the dimension of the parameter vector θ.

3.2 *A model for phased-in intervention*

The development in Section 3.1 was based on an instantaneous intervention model. This model assumed the intervention occurred instantly at time C, at which point a random sample of individuals still at risk (i.e. individuals with incubation times $> C$) were truncated because their disease was prevented. A more realistic model is that the intervention was carried out over a period of time beginning at calendar time C.

We model the probability of receiving an effective intervention that could prevent disease among the N exposed individuals as a continuous function of calendar time. These N individuals would eventually succumb to disease in the absence of an effective intervention. A simple model for the phase-in of the intervention is to assume the public health campaign begins and ends at calendar times C and L respectively and that individuals who have not yet developed disease nor received the intervention are at risk of receiving the intervention at a constant hazard rate λ between calendar times C and L. We assume that if the individual receives the intervention prior to onset of disease then disease would definitely be prevented. The survival function $S(t)$ corresponding to this hazard function for the time to intervention is

$$S(t) = \begin{cases} \exp(-\lambda(t-C)) & C \leqslant t \leqslant L \\ \exp(-\lambda(L-C)) & t > L. \end{cases} \tag{6}$$

Then $L_2 = L_2(\theta; t_1, t_2|X_1, X_2)$ becomes

$$L_2 = \prod_{i=1}^{X_1} \frac{f(t_{i1})}{F(C)} \prod_{i=1}^{X_2} \frac{f(t_{2i})S(t_{2i})}{\int_C^\infty f(u)S(u)\,\mathrm{d}u}. \tag{7}$$

We will not attempt to estimate the parameter λ from the data. Rather, external information about λ could be used. Alternatively, a sensitivity analysis to different values of λ could be performed to determine if deviations from the instantaneous intervention model of Section 3.1 could significantly affect the results. We choose different values of λ guided by external information to evaluate the sensitivity of the results. With the parameter λ fixed, as L converges to C the phased-in intervention model converges to the instantaneous intervention model. As described in Section 3.1 and the Appendix, L_2 in expression (7) can be maximized to estimate θ, and then $\hat{N} = X_1/\hat{F}(C)$.

The likelihood L_2 in expression (7) is also valid if, in addition to the phased-in intervention over continuous time, a proportion of persons were given the intervention exactly at time C and were thus truncated at time C. This is justified by noting that if a proportion of persons γ were given the intervention at exactly time C, then a factor $(1-\gamma)$ would appear both in the numerator and the denominator inside the second product sign in L_2 and would cancel. Thus, the likelihood L_2 in expression (7) is also justified under a hybrid model where there was both an instantaneous intervention given to a fraction a patients at time C and a phased-in intervention given to another fraction of persons at some point between calendar times C and L.

4. RESULTS

The methods of Section 3 were applied to the 70 observed incubation periods from the Sverdlovsk outbreak (Figure 1). The sample median and mean of the 70 incubation periods were 10 and 12.2 days

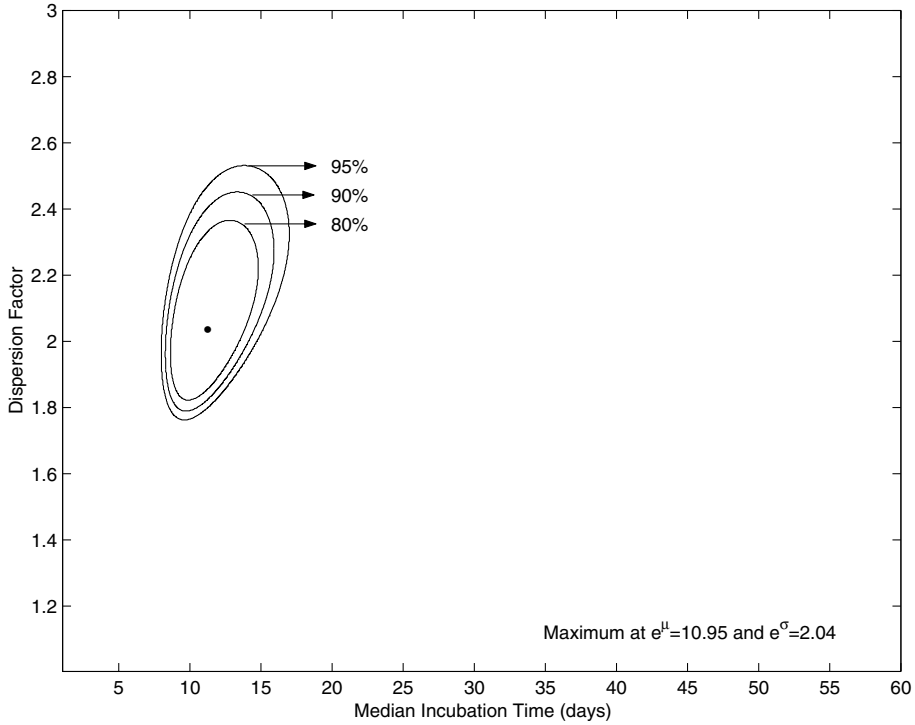

Fig. 2. 95%, 90% and 80% joint confidence regions for the median incubation period (e^{μ}) and dispersion factor (e^{σ}) under a lognormal model for the incubation distribution (instantaneous intervention model with $C = 15$).

respectively with range of 2–40. For our first set of analyses, we assumed the administration of antibiotics was initiated 15 days after the release of the pathogen, in which case $C = 15$, and fitted the instantaneous intervention model of Section 3.1. There were $X_1 = 54$ cases with incubation periods less than 15 days and $X_2 = 16$ cases with incubation periods greater or equal to 15 days. A lognormal model for the incubation period distribution was used, i.e. the log of the incubation period was normally distributed with mean μ and variance σ^2, $\log(t) \sim N(\mu, \sigma^2)$.

Maximum likelihood estimates of the parameters $\theta = (\mu, \sigma)$ were obtained by maximizing L_2 given in expression (3) using a grid search over values (μ, σ). The maximum likelihood estimates of the median and mean incubation periods were 11.0 days and 14.2 days, respectively. The maximum likelihood estimate of σ was 0.713. Following Sartwell (1950), we refer to the parameter $d = \exp(\sigma)$ as the 'dispersion factor'. The estimate of d was $\exp(0.713) = 2.04$. The interpretation of the dispersion factor is that roughly 68% of incubation periods will fall in the interval, median divided by d to median multiplied by d, in this case 5.4–22.4 days. About 95% of incubation periods will fall in the interval median divided by d^2 to median multiplied by d^2, in this case 2.6–45.8 days. About 1% of incubation periods are greater than 58 days. Joint 95, 90 and 80% confidence regions for the median incubation period $\exp(\mu)$, and the dispersion factor are shown in Figure 2. A confidence interval for the median, was also found by inverting a likelihood ratio test. The 95% confidence interval for the median incubation period was 8.5–15.3 days.

The maximum likelihood estimate of N was $X_1/\hat{F}(15) = 54/0.255 = 80$. Thus, it was estimated that in the absence of a public health intervention the Sverdlovsk outbreak would have resulted in 14% more deaths or about 10 additional deaths. However, the 95% confidence interval for N is wide, ranging from

70 to 113.

The phased-in intervention model of Section 3.2 was also applied to the data. The vaccine and antibiotic programmes in Sverdlovsk were carried out over a period of days to weeks, although the exact time period is not certain. Unfortunately, there is relatively little direct information about the parameter λ in expression (6) that describes the rate at which individuals in Sverdlovsk received effective interventions. While reports indicate that perhaps 80% of eligible persons were vaccinated at least once (Meselson *et al.*, 1994), it is not known if these vaccinations were effective in preventing disease or death. Further, it is not known what proportion of persons received antibiotics or for what duration, and in some individuals the antibiotics would have nullified the effect of the live vaccine. Our approach was to try different values for λ as part of a sensitivity analysis to determine if our results would deviate substantially from the simple instantaneous intervention model of Section 3.1. We used a range of values for λ and L. We found that with L fixed, as λ increased the estimated median incubation period also increased. For example, assuming the public health campaign was carried out over a two week period (beginning at $C = 15$ and ending at $L = 29$), and $\lambda = 0.11$ (which approximately corresponds to $S(L - C) = S(14) = 0.20$), then the median incubation period was 14.1 days (95% CI 9.9–23.1) and the dispersion factor $d = 2.29$. However, it seems unlikely that 80% of individuals received effective interventions and thus we consider this set of assumptions as leading to upper bounds on the incubation period. We performed a sensitivity analysis of the incubation period distribution to different values of λ. Figure 3 shows the estimated lognormal incubation densities for λ equal to 0.11, 0.036 and 0.016 which correspond respectively to 80, 40 and 20% receiving the intervention over a two week period. The estimated medians and dispersion parameters for $\lambda = 0.036$ were 11.75 days and $d = 2.11$ respectively; and for $\lambda = 0.016$ were 11.25 days and $d = 2.07$ respectively. Thus, in this data set, the estimates of the incubation period are not particularly sensitive to the specific assumptions about the timing of the public health intervention other than that the intervention occurred in mid to late April 1979.

The above results assumed a lognormal distribution for the incubation periods. Figure 4 is a normal probability plot of the log-incubation times. It was not clear to us what a normal probability plot from a truncated normal sample would look like and whether or not it would deviate from a straight line. To answer this question, we performed a simulation study to determine whether Figure 3 is at least consistent with a truncated lognormal distribution by generating incubation times from a lognormal distribution with median 11 days and dispersion factor 2.04. We simulated 100 log-incubation periods and then randomly truncated those individuals with incubation periods greater than 15 with probability p. We performed various simulations with each of $p = 0.0, 0.4, 0.6$ and 0.8. The normal probability plots (not shown) appeared very roughly linear except for a subtle cusp that appears around day 15 when the probability of truncation is high. The normal probability plot in Figure 4 for the Sverdlovsk data also appears very roughly linear with a slight cusp about day 15. Thus, while it is not possible to draw definitive conclusions from this simulation exercise, especially in light of the known variability in such plots even with moderate sample sizes, it appears that Figure 4 is at least not inconsistent with the pattern one would expect from lognormal data with a small or moderate amount of truncation.

As part of a sensitivity analyses we also fitted a Weibull model $F(t) = 1 - \exp(-\alpha t^{\beta})$ using the instantaneous intervention model with $C = 15$. The maximum likelihood estimators of the parameters of the Weibull model were $\alpha = 0.00785$ and $\beta = 1.67$. The estimated median incubation period was 14.6 days. Figure 5 compares the incubation period densities based on the lognormal and the Weibull distributions using the instantaneous intervention model. Although the Weibull model gave a slightly larger median than the lognormal model, the tail probabilities were slightly smaller. For example, the Weibull and lognormal model predicted that about 1% of incubation periods are greater than 45 days and 58 days, respectively. The maximized log-likelihood for the two parameter lognormal model (-35.9) was greater than the two-parameter Weibull model (-37.6) suggesting that the lognormal model fits the data somewhat better.

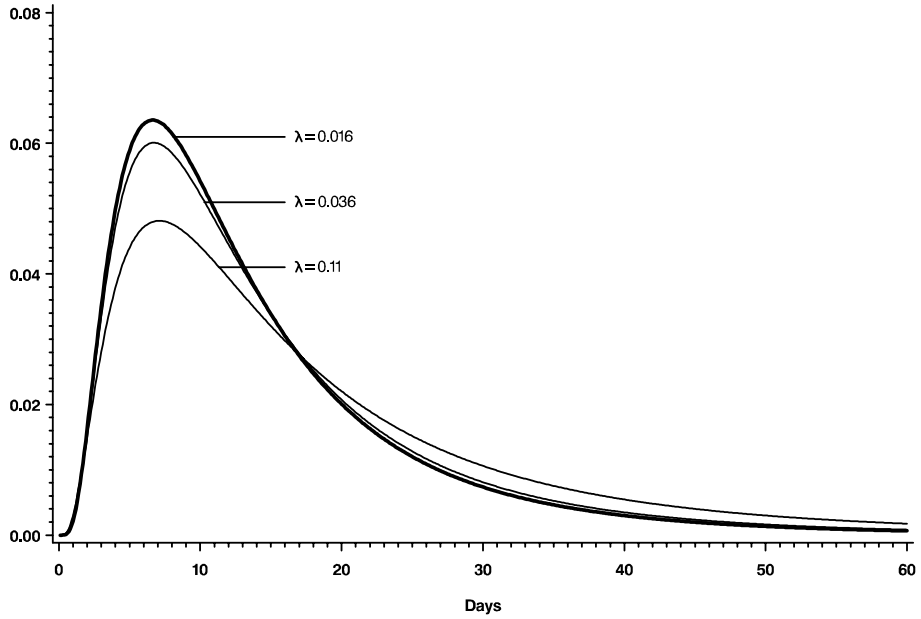

Fig. 3. Sensitivity analysis of lognormal incubation period densities using the phased-in intervention model to different values of λ ($\lambda = 0.11, 0.036$ and 0.016).

An additional analysis was performed where we analysed the times to death (as opposed to onset) of the 70 lethal cases. We estimated the cumulative distribution function of death following anthrax exposure, $F_d(t)$, using the instantaneous intervention model with $C = 15$ and assuming a lognormal model. The maximum likelihood estimates of the mean and median times to death were 18.2 and 15.8 days respectively, with dispersion factor of 1.70. Thus, as expected, death on average occurred within a few days following onset.

5. DISCUSSION

We have developed methods for estimating the incubation period distribution and the size of a population from a sample of randomly truncated incubation times, and have applied the methods to the Sverdlovsk anthrax outbreak. Our estimates of the incubation period of inhalational anthrax are somewhat longer than other, widely cited estimates of between 1 and 7 days. We found that point estimates for the median incubation period under various assumptions about the parametric model for the incubation period and duration of the intervention were about 11 days. Based on the lognormal model with instantaneous intervention at $C = 15$, we have estimated the median and mean incubation period of inhalational anthrax to be 11.0 and 14.2 days, respectively. This model implies that about 1% of incubation periods are greater than 58 days. A recent consensus recommendation for the public health response to the use of anthrax as a biological weapon against a civilian population (Inglesby *et al.*, 1999) was that antibiotic therapy should be continued for 60 days in suspected anthrax cases. This recommendation was based on the presumption that even if the treated patient survives the anthrax infection, the risk of recurrence remains for at least 60 days because of the possibility of delayed germination of spores. Our estimate of the 99th percentile of the incubation period distribution of inhalational anthrax was 58 days, lending support to the

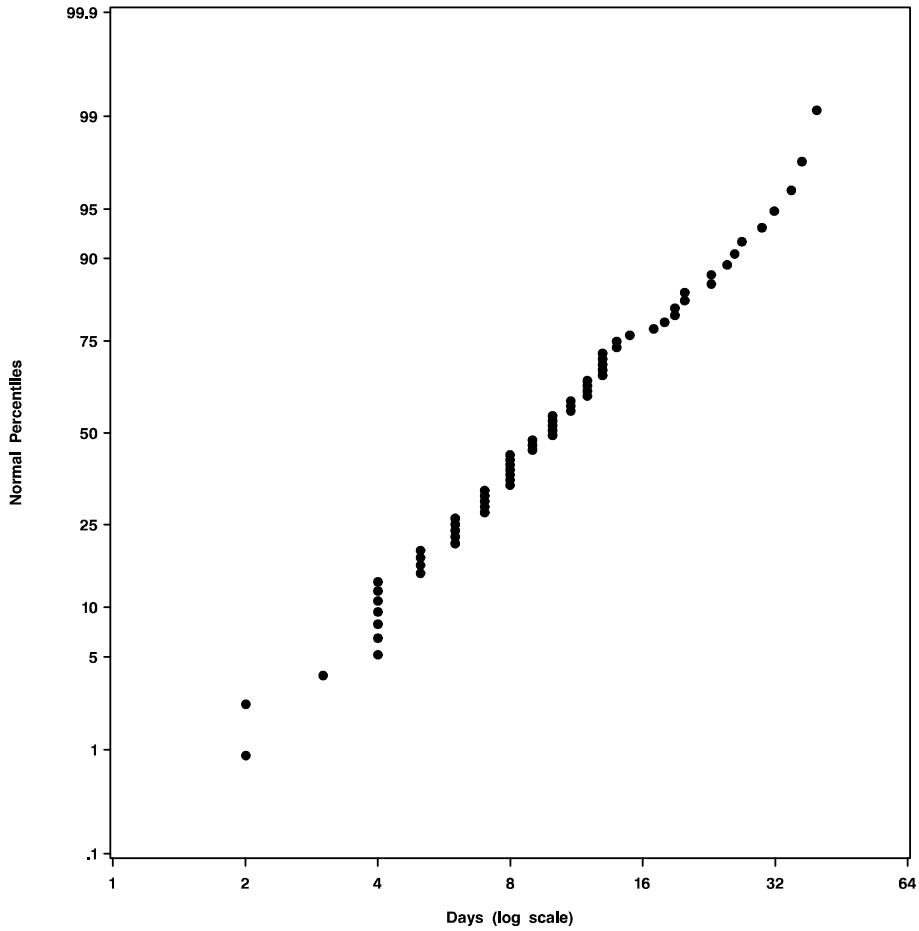

Fig. 4. Normal probability plot for log of incubation periods from Sverdlovsk anthrax outbreak.

recommendation.

Our results could also be used to assess the magnitude of a future outbreak. For example, suppose X is the cumulative number of deaths that have been observed within T days after the release of anthrax pathogen. The total number of anthrax deaths in the absence of any public health intervention is estimated from the formula $X/F_d(T)$ where F_d is the cumulative distribution function of the time to death following exposure among lethal cases. Table 1 illustrates these calculations for different values of X and T, assuming $F_d(t)$ follows a lognormal distribution with median of 15.8 days and dispersion factor 1.70. For example, if there are 50 deaths within 10 days of the release of a pathogen, the total number of deaths from the outbreak is estimated to be 255 in the absence of any public health prevention; however, if there are 50 deaths within 5 days of the release of the pathogen, the size of the outbreak is estimated to be over 3000.

The results we have presented are based principally on the lognormal model for the incubation period of inhalational anthrax. The lognormal model for the incubation period distribution of infectious diseases has a long history (Sartwell, 1950). Sartwell applied the lognormal distribution to 18 data sets representing 13 different infectious diseases including measles, poliomyelitis, and salmonelliosis. Although the median

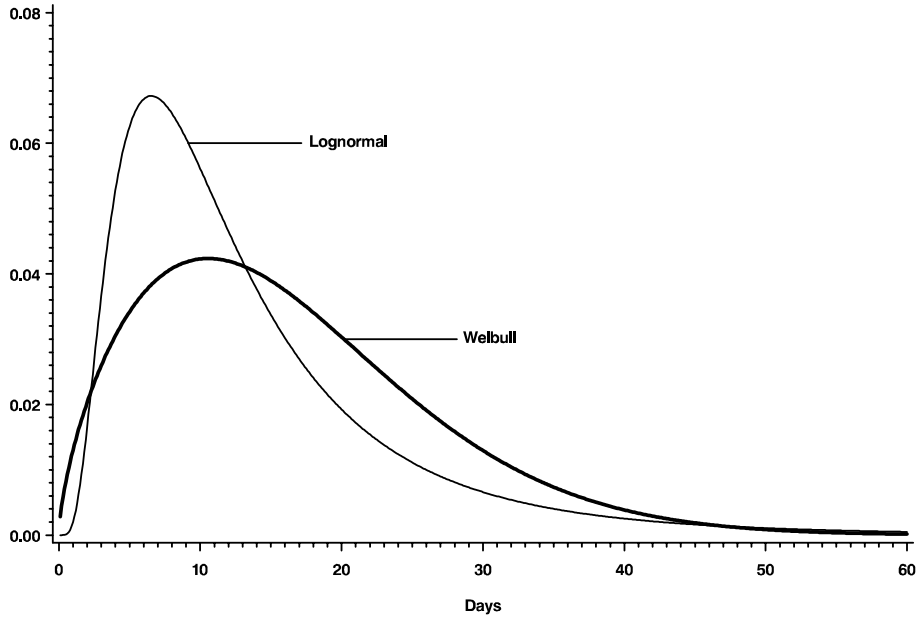

Fig. 5. Incubation period densities under lognormal and Weibull models using the instantaneous intervention model ($C = 15$).

Table 1. *Estimated total number of deaths* $(X/F_d(T))$ *from an anthrax outbreak in the absence of any public health intervention if a cumulative number of X deaths are observed within T days following release of the pathogen. Calculations are based on a lognormal distribution for* F_d *with median* $\exp(\mu) = 15.8$ *days and dispersion factor* $\exp(\sigma) = 1.70$

Number of cases (X)	Days (T)			
	5	10	15	20
5	322	26	11	7
10	644	51	22	15
20	1289	102	43	30
30	1933	153	65	45
40	2577	204	87	60
50	3221	255	108	74
60	3866	306	130	89
70	4510	357	151	104
80	5154	408	173	119
90	5798	459	195	134
100	6443	510	216	149

incubation periods for these very different diseases ranged from 56 hours to 100 days, Sartwell found that the estimated dispersion factors were not very different, ranging between 1.1 and 2.1. Our estimate of the dispersion factor of inhalational anthrax from the Sverdlovsk outbreak of 2.04 is close to the upper end of this range. The higher dispersion in incubation periods of inhalational anthrax could be consistent with high variability in doses of spores delivered to exposed cases.

Our estimates are sensitive to a number of model assumptions. For example, our assumption in Section 3.1 that the probability of truncation (i.e. prevention of onset of symptomatic disease) does not depend on the (potential) incubation period could be an oversimplification: interventions may be less effective when the onset of disease in imminent. We also assumed that the intervention may prevent some deaths (that is, possibly decrease p) but does not alter the incubation distribution $F(t)$ among deaths that are not prevented. This assumption is met for interventions that are given after the onset of symptoms. One could question the assumption if an intervention given before onset, such as a vaccine, only delays symptoms but does not prevent death, in which case we could overestimate $F(t)$. However, there is no clear evidence to support such a scenario. Further, one could argue that for public health planning purposes it would be a more serious error to underestimate than to overestimate the incubation period distribution. If we underestimate the incubation period distribution we could both underestimate the size of an outbreak and underestimate the optimal duration to treat suspected cases. An alternative analysis to circumvent the assumption that $F(t)$ is not altered by the intervention could be based on only the X_1 persons with incubation times less than C; such an analysis would use only the first factor in L_2 in expression (3) corresponding to the likelihood of the X_1 cases. However, as pointed out previously in AIDS applications, such an approach is nearly non-identifiable and provides extremely imprecise estimates that depend very strongly on parametric assumptions (Kalbfleisch and Lawless, 1989; Brookmeyer and Gail, 1994).

Our estimates of quantities such as the 99th percentile of the incubation period distribution clearly depend on model extrapolation of the parametric model for the incubation period distribution. We cannot rule out the possibility that the data in Figure 1 were generated from a more complex parametric model for the incubation period distribution with, in fact, no truncation. An area of possible future work is to take a semiparametric approach to estimation of the incubation period distribution. For example, one could attempt to use the likelihood L_2 in expression (3) to estimate the incubation period distribution under a wider class of flexible and smooth distribution functions.

APPENDIX

In this appendix we show that L given in expression (1) can be maximized by the following proposed estimators: define $\hat{\theta}$ as the value that maximizes L_2, and $\hat{N} = X_1/\hat{F}(C)$ where $\hat{F}(C) = F(C; \hat{\theta})$.

Suppose we were to replace $F(C)$ by a free parameter α in the factor L_1 in expression (1) for L. Then, writing $G = \log L$, we have

$$G(N, \theta, \alpha) = \log N! - \log(N - X_1)! - \log X_1! + X_1 \log \alpha + (N - X_1) \log(1 - \alpha) + \log L_2. \quad \text{(A.1)}$$

Let U be the maximum of G in (A.1) over the parameters α, θ and N. Now U must be greater than or equal to the maximum of L over N and θ because we have introduced an additional free parameter α in (A.1). If we can show that $L(\hat{N}, \hat{\theta}) = U$, that is $L(\hat{N}, \hat{\theta})$ attains the upper bound U, then $\hat{N}$ and $\hat{\theta}$ must also maximize L. To show this, we note that the value of θ that maximizes (A.1) is $\hat{\theta}$ because θ occurs only in the term L_2 in (A.1). Next, for large N, we treat (A.1) as a continuous function of N and differentiate with respect to N and α. We approximate the digamma function $d \log N! \, \mathrm{d}N$ by $\log N$ for large N (Abramowitz and Stegun, 1970). Setting the derivatives of (A.1) with respect to N and α equal to

zero, we obtain

$$\frac{\partial G}{\partial N} = \log N - \log(N - X_1) - \log(1 - \alpha) = 0$$
$$\frac{\partial G}{\partial \alpha} = \frac{X_1}{\alpha} - \frac{N - X_1}{1 - \alpha} = 0.$$

The above two equations are satisfied by $\hat{N} = X_1/\hat{F}(C)$ and $\hat{\alpha} = \hat{F}(C)$. Thus, $\hat{N}$, and $\hat{\alpha} = \hat{F}(C)$ maximizes G, that is $U = G(\hat{N}, \hat{\theta}, \hat{F}(C))$. By simple inspection we note that $L(\hat{N}, \hat{\theta})$ is also equal to $U = G(\hat{N}, \hat{\theta}, \hat{F}(C))$ and thus $\hat{N}$ and $\hat{\theta}$ must maximize L.

REFERENCES

ABRAMOWITZ, M. AND STEGUN, I. (1970). *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables (Applied Mathematics Series)*. Washington, DC: National Bureau of Standards.

BENENSON, A. S. (1995). *Control of Communicable Diseases Manuals*, 16th edn. Washington, DC: American Public Health Association.

BRACHMAN, P. S. (1980). Inhalation anthrax. *Annals of the New York Academy* **353**, 83–93.

BRACHMAN, P. S. AND FRIEDLANDER, A. M. (1999). Anthrax. In Plotkin, S. A. and Orenstein, W. A. (eds), *Vaccines*, 3rd edn. Philadelphia: Saunders, p. 630.

BROOKMEYER, R. (1996). AIDS, epidemics and statistics. *Biometrics* **52**, 781–796.

BROOKMEYER, R. (1998). Biased sampling of cohorts in epidemiology. In Armitage, P. and Colton, T. (eds), *Encyclopedia of Biostatistics*, New York: Wiley, pp. 338–350.

BROOKMEYER, R. AND GAIL, M. H. (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association* **83**, 301–308.

BROOKMEYER, R. AND GAIL, M. H. (1994). *AIDS Epidemiology: A Quantitative Approach*. London: Oxford University Press.

BROOKMEYER, R. AND LIAO, J. (1990). The analysis of delays in disease reporting: methods and results for the acquired immunodeficiency syndrome. *American Journal of Epidemiology* **132**, 355–365.

CENTERS FOR DISEASE CONTROL (2000). Surveillance for adverse events associated with anthrax vaccination, US Department of Defense 1998-2000. *Morbidity and Mortality Weekly Report* **49**, 341–345.

CIESLAK, T. AND EITZEN, E. (2000). Clinical and epidemiologic principles of anthrax. *Emerging Infectious Diseases* **5**, 552, www.cdc.gov/ncidod/EID/vol5no4/cieslak.htm.

DEGRUTTOLA, V., TU, X. AND PAGANO, M. (1992). Pediatric AIDS in New York city: estimating the distribution of infection, latency and reporting delay and projecting future incidence. *Journal of the American Statistical Association* **87**, 633–640.

FRANZ, D. R., JAHRLING, P. B., FRIEDLANDER, A., MCCAIN, D. J., HOOVER, D. L., BRYNE, W. R., PAVLIN, J., CHRISTOPHER, G. W. AND EITZEN, E. M. (1997). Clinical recognition and management of patients exposed to biological warfare agents. *Journal of the American Medical Association* **278**, 399–411.

FRIEDLANDER, A. (1997). Anthrax. In Zajtchuk, R. and Bellamy, R. E. (eds), *Textbook of Military Medicine: Medical Aspects of Chemical and Biological Warfare*, Washington, DC: Office of the Surgeon General, US Department of the Army, pp. 467–478.

FRIEDLANDER, A., WELKOS, S. L., PITT, M. L., EZZEL, J. W., WORSHAM, Y., ROSE, K. J., IVINS, B. E., LOWE, J. R., HOWE, G. B. AND MIKESELL, P. (1993). Postexposure prophylaxis against experimental inhalation anthrax. *Journal of Infectious Diseases* **167**, 1239–1242.

GUILLEMIN, J. (1999). *Anthrax: The Investigation of a Lethal Outbreak*. Berkeley, CA: University of California Press.

INGLESBY, T. V., HENDERSON, D. A., BARTLETT, J. G., ASCHER, M., EITZEN, E., FRIEDLANDER, A., HAVER, J., MCDADE, J., OSTERHOLM, M., O'TOOLE, T., PARKER, G., PEVI, T., RUSSELL, P. AND TONAT, K. (for the Working Group on Civilian Biodefense) (1999). Anthrax as a biological weapon: medical and public health management. *Journal of the American Medical Association* **281**, 1735–1745.

KALBFLEISCH, J. AND LAWLESS, J. F. (1989). Inference based on retrospective ascertainment: an analysis of data on transfusion related AIDS. *Journal of the American Statistical Association* **84**, 36–72.

LUI, K., LAWRENCE, D. N., MORGAN, W. M., PETERMAN, T. A., HAVERKOS, H. W. AND BREGMAN, D. J. (1986). A model-based approach for estimating the mean incubation period of transfusion-associated acquired immunodeficiency syndrome. *Proceedings of the National Academy of Science* **83**, 3051–3055.

MEDLEY, G. F., ANDERSON, R. M., COX, D. R. AND BILLARD, L. (1987). Incubation period of AIDS in patients infection via blood transfusion. *Nature* **328**, 719–721.

MEDLEY, G. F., BILLARD, L., COX, D. R. AND ANDERSON, R. M. (1988). The distribution of the incubation period for the acquired immunodeficiency syndrome. *Proceedings of the Royal Society of London* B **233**, 367–377.

MESELSON, M., GUILLEMIN, J., HUGH-JONES, M., LANGMUIR, A., POPOVA, I., SHELOKOV, A. AND YAMPOLSKAYA, O. (1994). The Sverdlovsk anthrax outbreak of 1979. *Science* **266**, 1202–1208.

PILE, J. C., MALONE, J. D., EITZEN, E. M. AND FRIEDLANDER, A. (1998). Anthrax as a potential biological warfare agent. *Archives of Internal Medicine* **158**, 429–434.

SARTWELL, P. E. (1950). The distribution of incubation periods of infectious diseases. *American Journal of Hygiene* **51**, 310–318.

WANG, M-C (1992). The analysis of retrospectively ascertained data in the presence of reporting delays. *Journal of the American Statistical Association* **87**, 397–406.

[*Received August 29, 2000; revised December 4, 2000; accepted for publication December 11, 2000*]