Biostatistics (2003), **4**, 2, *pp.* 279–295
Printed in Great Britain

Branching process models for surveillance of infectious diseases controlled by mass vaccination

C. P. FARRINGTON*
Department of Statistics, The Open University, Milton Keynes, MK7 6AA, UK
C.P.Farrington@open.ac.uk

M. N. KANAAN
Department of Epidemiology and Population Health, The American University of Beirut, PO Box 11-0236, Riad El Solh 11072020, Beirut, Lebanon

N. J. GAY
Communicable Diseases Surveillance Centre, 61 Colindale Avenue, London, NW9 5EQ, UK

SUMMARY

Mass vaccination programmes aim to maintain the effective reproduction number R of an infection below unity. We describe methods for monitoring the value of R using surveillance data. The models are based on branching processes in which R is identified with the offspring mean. We derive unconditional likelihoods for the offspring mean using data on outbreak size and outbreak duration. We also discuss Bayesian methods, implemented by Metropolis–Hastings sampling. We investigate by simulation the validity of the models with respect to depletion of susceptibles and under-ascertainment of cases. The methods are illustrated using surveillance data on measles in the USA.

1. INTRODUCTION

Vaccination programmes represent one of the most effective ways of controlling infectious diseases. For example, measles, though still endemic in many countries of Africa and Asia, has been eliminated from large parts of Europe and America. Local elimination is the necessary precursor of global eradication, and also represents a desirable public health objective in its own right.

Global eradication of an infection is the reduction of the number of infections to zero. Elimination, on the other hand, is the interruption of sustained endemic transmission, which may be achieved by the maintenance of a high level of vaccination coverage. Elimination is best characterized in terms of the effective reproduction number of the infection, R, namely the average number of infectious individuals produced by one infectious case during his or her infectious period (Anderson and May, 1991; Dietz, 1993; Farrington *et al.*, 2001). Elimination of an infection may then be defined as the maintenance of the reproduction number below unity. Under conditions of elimination, infections can still occur, for example due to spread from imported cases, but such spread cannot lead to large-scale epidemics. Since each infectious case is replaced by no more than one other, the infection cannot become endemic, and outbreaks peter out with probability 1. The lower the value of R, the smaller and the shorter the outbreaks will be.

*To whom correspondence should be addressed

While all values $R \leqslant 1$ in theory result in the chain of transmission being interrupted with probability 1, in practice it is advisable to maintain R well below unity.

An important aim for the surveillance of mass vaccination programmes is to monitor the value of R over time. If the reproduction number gets too close to unity, intervention is required to reduce its value. For example, a special vaccination campaign against measles was undertaken in the UK in 1994 following such warnings (Gay *et al.*, 1995). It has recently been proposed to extend the surveillance of vaccination programmes by supplementing the surveillance of cases with the surveillance outbreaks (De Serres *et al.*, 2000). In this paper we develop the statistical aspects of these new surveillance methods.

Our methods rely on the theory of branching processes. These are well suited for the purpose of epidemiological surveillance, since they require data only on cases. However, the simplicity of the branching process approach comes at a price: the methods rely on an approximation to the epidemic process. More accurate methods, such as those based on chain binomial models (Becker, 1989), require information on the denominator of susceptible individuals. Such information is only ever available in very special, much analysed datasets on single outbreaks (Bailey, 1975; Becker, 1989). It is seldom, if ever, available in a surveillance setting. We show in this paper that branching process models, applied to surveillance of mass vaccination programmes in conditions of elimination, are of direct practical use for public health.

Branching processes play a fundamental role in epidemic theory, underpinning our understanding of the threshold behaviour of epidemics and the calculation of the critical vaccination threshold, and providing a simple model for the early stages of epidemic spread. Thus, much recent work on complex epidemic models makes use of branching process approximations (Marschner, 1992; Ball and Donnelly, 1995; Clancy and O'Neill, 1998; Caraco *et al.*, 1998; Muller and Kirkilionis, 2000; Ball and Lyne, 2001). More directly, branching processes have been proposed as statistical models on which to base inferences about the reproduction number R, represented by the offspring mean (Becker, 1974a,b, 1976, 1977; Heyde, 1979; Farrington and Grant, 1999; Yanev and Tsokos, 1999). Nevertheless, in contrast to the widespread use of other types of models, particularly the deterministic models of Anderson and May (1991), statistical models based on branching processes are seldom used in practice for infectious disease control.

In this paper we describe a framework for monitoring the value of the reproduction number R in the context described above, using data on outbreak size and duration. Existing methods based on such data condition on ultimate extinction of the process (Becker, 1974a; Farrington and Grant, 1999). However, the resulting conditional likelihoods are ill-suited to making inferences about whether $R \leqslant 1$ or $R > 1$. Instead, we use an unconditional modelling approach using censored likelihoods similar to those used in survival analysis. We derive surveillance thresholds based on upper profile likelihood confidence limits on R. These likelihoods may also be used in conjunction with the Bayesian approach of Heyde (1979), in which the threshold criterion is based on the posterior probability that $R > 1$. In our application the posterior distribution is evaluated by Metropolis–Hastings sampling.

In Section 2 we introduce the branching process model, describe the likelihoods based on outbreak surveillance data, and discuss the Bayesian criterion. However, as already noted, the branching process is only an approximate model for the spread of infection. In particular, it does not take into account the depletion of susceptibles in the population. Furthermore, in a practical surveillance setting, ascertainment of cases is unlikely to be complete. So in Section 3 we use simulations to investigate the effects of depletion of susceptibles and under-ascertainment on the estimates of λ. In Section 4 we illustrate the methods by applying them to surveillance data on the spread of measles in the USA.

2. UNCONDITIONAL LIKELIHOODS

We approximate the spread of infection by means of a homogeneous Galton–Watson branching process. Thus we assume that each case infects a random Z others, known as the offspring of the case

who infected them. The distribution of Z is called the offspring distribution, and we shall assume that it belongs to the power series family

$$P(Z = r) = a_r \frac{\theta^r}{A(\theta)}$$

where θ is the canonical parameter and $A(\theta) = \sum a_r \theta^r$, $a_r \geqslant 0$. For more details see, for instance, Guttorp (1991). In this paper we will specifically be concerned with the Poisson distribution, for which $a_r = 1/r!$ and $A(\theta) = \mathrm{e}^{\theta}$, and the geometric distribution, for which $a_r = 1$ and $A(\theta) = (1 - \theta)^{-1}$. The offspring mean $\lambda = \mathbb{E}(Z)$ is

$$\lambda = \frac{\theta A'(\theta)}{A(\theta)}.$$

For the Poisson offspring distribution, $\lambda = \theta$, while for the geometric, $\lambda = \theta(1 - \theta)^{-1}$. In applications to infectious diseases, the offspring mean λ corresponds to the reproduction number, usually denoted R in the epidemiological literature.

The extinction probability $q(\lambda)$ (which we shall sometimes write q) of a branching process originating from one case, with offspring distribution from the power series family, is the smallest root of

$$A(q\theta) = qA(\theta).$$

When $\lambda \leqslant 1$, $q(\lambda) = 1$, and, when $\lambda > 1$, $q(\lambda) < 1$ (Guttorp, 1991). For example, for the geometric offspring distribution, $q(\lambda) = \min\{1, 1/\lambda\}$. If $\lambda > 1$ then, conditional on extinction, the offspring distribution is a power series distribution with canonical parameter $\theta^* = q\theta$ (Waugh, 1958).

Standard methods for inference about λ are based on the likelihood given observation of the process up to some pre-determined generation $k > 0$. Let X_k denote the total size of the outbreak up to and including the kth generation; the outbreak starts from generation zero, with $X_0 = s$. The maximum likelihood estimator of λ is

$$\tilde{\lambda} = \frac{X_k - s}{X_{k-1}}.$$

This is consistent and asymptotically unbiased in the limit where s tends to infinity (Yanev, 1975). In the alternative limit where k tends to infinity, the MLE is inconsistent; this has motivated considerable recent research (Lockhart, 1982; Sriram, 1991; Jacob and Peccoud, 1998). In this paper we shall only be concerned with the limit in which the number of outbreaks $n \longrightarrow \infty$. Since the size of an outbreak starting from s cases may be regarded as the sum of the sizes of s independent outbreaks starting with a single case, the limits $n \longrightarrow \infty$ and $s \longrightarrow \infty$ are equivalent for outbreak sizes; this is not so for outbreak durations.

In practice, however, generations of spread are not observed with any accuracy. Thus X_k and X_{k-1}, for given k, are not readily observable unless the outbreak has ended by generation $k - 1$, in which case $X_k = X_{k-1} = X$, the outbreak size. Thus it makes sense to base inferences on outbreak sizes X, which are readily observable. The problem, however, is that when $\lambda > 1$, not all outbreaks become extinct, so that the distribution of X is improper. It is common practice to condition on extinction, since the resulting conditional distribution of X is proper: see Becker (1974a) and Guttorp (1991, pages 100–102). However, inference conditional on extinction provides no information on whether or not the observed data are consistent with $\lambda > 1$, and so is of little direct use for surveillance purposes. Guttorp (1991) discusses various informal ways round this problem, while Heyde (1979) goes so far as to suggest that there is no satisfactory solution outside a Bayesian framework. Similar issues arise with observations on outbreak duration, though these have received rather less attention in the literature (Farrington and Grant, 1999).

In this paper we develop an unconditional likelihood approach using observations, possibly censored, on outbreak sizes and durations. We observe all or a random sample of outbreaks originating in some time interval $[0, \tau]$ over which it is reasonable to expect that λ is roughly constant. For each such outbreak we make observations on (S, T, X, U) where $S \geqslant 1$ is the initial number of cases (generally one), $T \in [0, \tau]$ is the time at which the outbreak started, $X \geqslant S$ is the outbreak size and $U \geqslant 0$ the outbreak duration, namely the time between the onset of symptoms in the initial introductory case and the onset of symptoms in the final case. We regard X as having domain $\{s, s+1, ...\} \cup \{\infty\}$ and U as having domain $[0, \infty]$, the values ∞ corresponding to non-extinction. Note that extending the domains of X and U in this way to include ∞ is simply a device to ensure that non-extinction is included in the events $X \geqslant x$ and $U \geqslant u$; the values $X = \infty$ and $U = \infty$ never arise in calculations and are never observed. The initial number of cases S and the originating time T depend on the pattern of importations of infections. We shall assume that S and T are uninformative about λ and hence condition on S, T.

Observations are subject to censoring. For example, we might censor observations beyond some time point $\nu \geqslant \tau$. Thus each observation is either of the form $(S = s, T = t, X = x, U = u)$ for an outbreak starting at time t in $[0, \tau]$ and ending before time ν, or of the form $(S = s, T = t, X \geqslant x_\nu, U \geqslant \nu - t)$ if the outbreak has not ended before time ν, a total x_ν cases having been observed up to that time. Other censoring schemes are possible: for example we might observe all outbreaks up to the ξth case, in which case censored observations are of the form $(S = s, T = t, X \geqslant \xi, U \geqslant v_\xi - t)$ where v_ξ is now the time at which the ξth case occurs. Whatever the censoring scheme, the key point is that non-extinct processes are censored with probability 1. Note that in the application envisaged here, in which we expect λ to be well below 1, it is usually possible to arrange the surveillance system to ensure that, in practice, censoring rarely occurs. However, allowing for the possibility that it might occur, as it would if in fact λ were close to or above 1, is essential in developing a coherent inferential framework.

For the likelihood-based analyses described below, our surveillance threshold will be the upper 95% profile likelihood confidence limit on λ. If this exceeds 1, then appropriate action is taken. We consider inference about λ on the basis of observations on outbreak size X or outbreak duration U.

2.1 *Outbreak size*

Inference about λ from outbreak size data, conditional on extinction, has been discussed by Becker (1974a). Let X denote the size of an outbreak generated from $S = s$ initial cases. The distribution of X belongs to the power series family and is of the form

$$P\,(X = x; s) = b\,(x, s)\,\frac{\theta^{x-s}}{A\,(\theta)^x} \tag{2.1}$$

where $b\,(x, s)$ is a constant. For example, when the offspring distribution is Poisson the total outbreak size follows the Borel–Tanner distribution (Haight and Breuer, 1960):

$$P(X = x; s) = \frac{s x^{x-s-1} \lambda^{x-s} \mathrm{e}^{-x\lambda}}{(x-s)!}, \qquad x = s, s+1, \ldots . \tag{2.2}$$

When the offspring distribution is geometric, the total outbreak size follows the distribution

$$P(X = x; s) = \frac{s}{2x-s}\binom{2x-s}{x-s}\frac{\lambda^{x-s}}{(1+\lambda)^{2x-s}}, \qquad x = s, s+1, \ldots . \tag{2.3}$$

In this case $X - s$ has a Lagrangian generalized negative binomial distribution (Johnson *et al.*, 1992, Chapter 5, Section 12). These distributions are defined for $X < \infty$. If $\lambda > 1$ they are improper. We extend

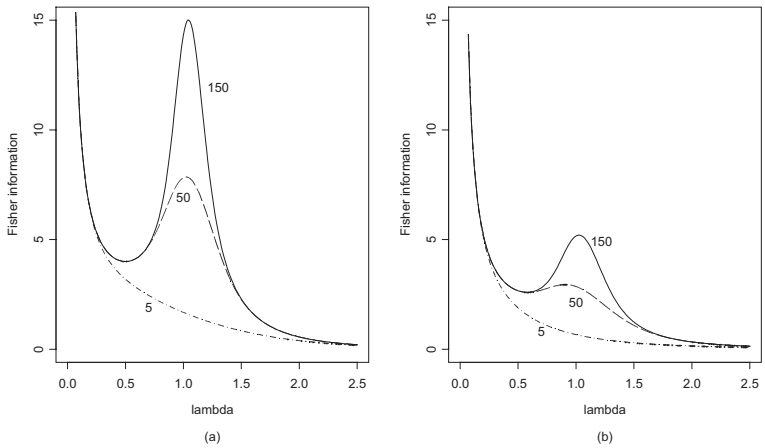

Fig. 1. Fisher information for a single observation on outbreak size, with censoring at $\xi = 5, 50, 150$; (a) Poisson offspring, (b) geometric offspring.

them to the whole domain of X, $\{s, s+1, ...\} \cup \{\infty\}$ by defining

$$P(X = \infty; s) = 1 - q(\lambda)^s.$$

Let X^* denote the outbreak size at which observation is censored. X^* may be random or determinate, depending on the censoring scheme. We assume $P(X^* < \infty) = 1$, and set $C = 1$ if $X < X^*$, $C = 0$ if $X \geqslant X^*$. We thus observe $\min\{X, X^*\}$ and C. Given observations $(s_1, x_1, c_1), \ldots, (s_n, x_n, c_n)$ on a sample of n outbreaks, the likelihood for λ is thus

$$L(\lambda; s, x, c) = \prod_{i=1}^{n} \left\{ P(X = x_i; s_i)^{c_i} \left(1 - \sum_{j=s_i}^{x_i - 1} P(X = j; s_i) \right)^{1-c_i} \right\}. \tag{2.4}$$

When all the $s_i = 1$, the likelihood has the same form as that for a parametric survival model in discrete time with time to event X, $X = \infty$ denoting non-occurrence of the event. The MLE of λ is asymptotically unbiased and consistent; the asymptotic limit considered here is the usual one in which $n \to \infty$. The MLE $\widehat{\lambda}$ can take any value in $[0, \infty)$, and profile likelihood confidence intervals for λ can straddle the value $\lambda = 1$; this threshold does not cause any particular problem.

It is of interest to examine the Fisher information $i(\lambda)$ from a single outbreak with $s = 1$. Figure 1 shows graphs of $i(\lambda)$ for three censoring schemes, in which values of X above fixed thresholds ξ are censored. The information in the neighbourhood of $\lambda = 1$ grows sharply for larger values of ξ. This may be explained heuristically as follows. When λ is just below 1, all outbreaks are finite with large expected size. When λ is just above 1, most outbreaks are finite, with large expected size conditional on remaining finite. The remainder are unbounded and hence censored. If the threshold ξ is high, then the combination of few censored outbreaks, and large size for those that are uncensored, is expected only when λ is close to 1, and hence conveys substantial information about λ. This information increases with ξ, and in fact becomes unbounded at $\lambda = 1$ as $\xi \to \infty$.

In the absence of censoring, the MLE $\widehat{\lambda} = 1 - (\Sigma s_i)/(\Sigma x_i)$, and Σx_i is sufficient for λ (Becker, 1974a). Thus, provided that the censoring scheme is organized so that censoring rarely occurs, there is little gain in efficiency in making use of information on outbreak duration. However, as will be shown later in the paper, if ascertainment of cases is incomplete then estimation from data on outbreak duration might be less biased.

2.2 *Outbreak duration*

The duration of outbreaks when $\lambda \leqslant 1$ has been studied in Farrington and Grant (1999). If Y denotes the number of generations of spread from a single introductory case at generation zero, then its distribution function $f_k = P(Y \leqslant k)$ satisfies the recursive relation

$$f_0 = \varphi(0), \quad f_{k+1} = \varphi(f_k) \qquad k = 0, 1, 2, \ldots$$

where $\varphi(x) = A(x\theta)/A(\theta)$ is the probability generating function of the offspring distribution (Guttorp, 1991). If Y is the number of generations of spread from s introductory cases, the distribution function is given by $f_{k,s} = \{f_k\}^s$. When the offspring distribution is Poisson with mean $\lambda \leqslant 1$,

$$f_k = \mathrm{e}^{-\lambda} E_k\left(\mathrm{e}^{\lambda \mathrm{e}^{-\lambda}}\right) \tag{2.5}$$

where $E_k(x)$ denotes the iterated exponential function $E_0(x) = 1$, $E_{k+1}(x) = x^{E_k(x)}$. When the offspring distribution is geometric with mean $\lambda < 1$,

$$f_k = \frac{1-\lambda^{k+1}}{1-\lambda^{k+2}}, \qquad k = 0, 1, 2, \ldots . \tag{2.6}$$

When $\lambda = 1$ this reduces to $f_k = (1+k)/(2+k)$.

Expressions (2.5) and (2.6) are also valid when $\lambda > 1$, though the distributions are improper. We thus define

$$f_{\infty,s} = P(Y = \infty | S = s) = 1 - q(\lambda)^s$$

to obtain proper distributions on the whole domain of Y, $\{0, 1, 2, \ldots\} \cup \{\infty\}$. If the numbers of generations of spread are known exactly or censored, then these distributions can be used in exactly the same ways as the distributions of outbreak size to obtain likelihoods which are then maximized to yield estimates of λ. Figure 2 shows the Fisher information $i(\lambda)$ for a single outbreak with $s = 1$, for three censoring schemes in which values of Y above fixed thresholds κ are censored. The shape of $i(\lambda)$ is very similar to that in Figure 1.

However, the precise number of generations is usually not known. In Farrington and Grant (1999) a simple method of analysis is described using generalized linear models, based on whether or not secondary spread has occurred. However this disregards information on the duration U of the outbreak, where U is the time between the onset of symptoms in the initial introductory case and the onset of symptoms in the final case. Such data are often available in a surveillance setting. The number of generations can then be imputed using knowledge about the mean serial interval, that is, the mean interval between the appearance of symptoms in a case and the appearance of symptoms in a secondary case infected by the first.

Alternatively, a stochastic imputation method may be used to allow for the random variation in serial intervals. We assume that serial intervals z are distributed as $h(z)$, and let $h_n(z)$ denote the distribution of the sum of n independent serial intervals. In general, the serial interval distribution may depend on the offspring mean λ, though in practice this dependence is ignorable for infections with short infectious periods. We shall assume that the first and last cases in the outbreak are necessarily those separated by the maximum number of generations of spread in the outbreak. This is reasonable provided that the variance of the serial interval distribution is small. The distribution of the outbreak duration U, given s introductory cases, is then

$$f(u; s) = \begin{cases} p_0(\lambda; s), & u = 0 \\ \sum_{n=1}^{\infty} h_n(u)\, p_n(\lambda; s), & 0 < u < \infty \\ 1 - q(\lambda)^s, & u = \infty \end{cases} \tag{2.7}$$

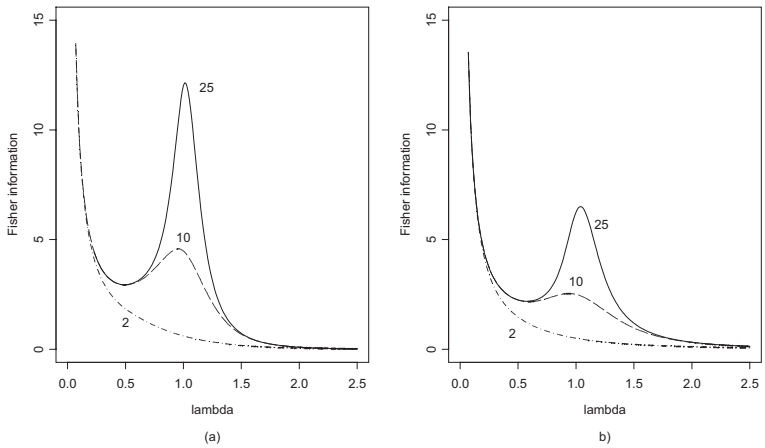

Fig. 2. Fisher information for a single observation on number of generations of spread, with censoring at $\kappa = 2$, 10 and 25; (a) Poisson offspring, (b) geometric offspring.

where $p_n(\lambda; s) = f_{n,s} - f_{n-1,s}$ is the probability mass function of the distribution of the number of generations of spread, with offspring mean λ, starting with s cases at generation 0. This method requires knowledge of the serial interval distribution $h(x)$.

Maximum likelihood estimation of λ may thus be based on observations $(s_1, u_1, c_1), \ldots, (s_n, u_n, c_n)$ on a sample of n outbreaks; as previously the c_i are censoring indicators: 1 if the observation is uncensored, 0 if it is censored. The likelihood is thus

$$L(\lambda; s, u, c) = \prod_{i=1}^{n} \left\{ f(u_i; s_i)^{c_i} \left(1 - \int_0^{u_i} f(x; s_i)\, \mathrm{d}x \right)^{1-c_i} \right\}. \tag{2.8}$$

When the $s_i = 1$, this may be recognized as the likelihood for a parametric survival model, in continuous time, with time to event U, in which there is a probability $p_0(\lambda; 1)$ of immediate failure and a probability $1 - q(\lambda)$ of never failing. Standard likelihood theory again applies, and the MLE is consistent and asymptotically unbiased in the limit $n \longrightarrow \infty$.

2.3 *Bayesian inference*

Bayesian methods for branching processes have been applied in an epidemiological context by Heyde (1979) and more recently by Yanev and Tsokos (1999), who discuss a variety of Bayesian estimators. One of the attractions of the Bayesian approach is that, given a suitable prior distribution, it allows calculation of the posterior probability that $\lambda > 1$. This provides an alternative to thresholds based on confidence intervals: for example, an intervention might be considered if the posterior probability that $\lambda > 1$ were greater than, say, 2.5%.

The Bayesian approach is readily applicable to data on outbreak size or outbreak duration of the type considered above. Given a prior distribution $\pi(\lambda)$, the posterior distribution of λ given observations, possibly censored, on outbreak size or duration, is

$$f(\lambda) = \frac{L(\lambda)\pi(\lambda)}{\int_0^\infty L(z)\pi(z)\, \mathrm{d}z} \tag{2.9}$$

where $L(\lambda)$ denotes either of the likelihoods from equation (2.4) or (2.8).

Heyde (1979) and Yanev and Tsokos (1999) make use of conjugate prior distributions. For outbreak size distributions, these are most easily expressed in terms of the canonical parameter θ rather than λ; the conjugate distributions are the gamma for the Borel–Tanner, and the beta for the Lagrangian negative binomial. However, there are no simple conjugate priors for outbreak duration distributions. In any case, for our surveillance application, it would make sense to specify a single prior for all models, that is in some sense neutral with respect to whether $\lambda < 1$ or $\lambda > 1$. We shall therefore eschew conjugacy. Since λ is in effect a multiplier, it seems natural to choose a prior for which the distribution of $\log(\lambda)$ is symmetric. We shall thus select priors from the lognormal family $LN(z; \mu, \sigma) \sim \exp\left\{N\left(\mu, \sigma^2\right)\right\}$. When $\mu = 0$, the median is 1; we shall refer to the corresponding prior $\pi(z) = LN(z; 0, \sigma)$ as neutral since $\lambda < 1$ and $\lambda > 1$ have the same prior probability 0.5. We emphasize that the term 'neutral' relates specifically to $P(\lambda > 1)$; if the aim was to estimate λ we might prefer to use a different prior distribution, for example one with mode at 1.

The computational burden involved in calculating the posterior distribution (2.9), especially in the case of outbreak durations, is considerable. We therefore use Metropolis–Hastings sampling to evaluate the posterior distribution. Letting $g(w|z)$ denote the proposal density, we accept proposals with probability

$$\alpha(w, z) = \min\left\{\frac{L(z)\pi(z)\, g(w|z)}{L(w)\pi(w)\, g(z|w)}, 1\right\}.$$

A natural choice of proposal density is the lognormal $g(w|z) = LN(w; \log(z), \gamma)$ with dispersion parameter γ. With a neutral lognormal prior, the acceptance probability then becomes

$$\alpha(w, z) = \min\left\{\frac{L(z)}{L(w)} \exp\left(-\frac{(\log z)^2 - (\log w)^2}{2\sigma^2}\right), 1\right\}.$$

As already stated, the Bayesian approach may be implemented for both data on outbreak size and outbreak duration, the appropriate likelihood (2.4) or (2.8) being substituted for $L(\lambda)$ in the above equations.

3. SIMULATIONS

The branching process model provides only an approximation to the initial stages of the spread of an infection in a large population, since no account is taken of the depletion of susceptibles as the outbreak progresses. Generally, depletion of susceptibles will act to limit outbreak sizes and durations, and will therefore tend to bias the estimated value of λ towards 0. In the application we consider here we expect that $\lambda < 1$, the aim being to signal when λ becomes too close to 1. We are thus in a situation in which extensive spread of the infection is unlikely, and we would therefore expect the branching process approximation to be valid, at least for outbreaks in large communities.

Outbreaks are usually identified by space–time clustering of infections within a defined community such as a town or perhaps an institution such as a university or large school, rather than by identification of chains of transmission. This method of defining outbreaks is reasonable under conditions of elimination, in which relatively few infections occur. In these circumstances, clustering of cases provides good evidence that they are part of the same outbreak. However, not all the infections occurring in an outbreak may be ascertained. Such under-ascertainment will also tend to bias towards 0 the estimates of λ based on outbreak sizes and outbreak durations.

In this section we investigate the impact of these biases by simulation.

3.1 *Depletion of susceptibles*

The simulations were set up using the Reed–Frost model (Frost, 1976; Kotz and Johnson, 1986). Each outbreak was simulated by introducing a single infective in a susceptible population of size m; thus the total effective population size is $m + 1$. The escape probability was taken to be $\rho = 1 - \lambda/m$ rather than the perhaps more natural $e^{-\lambda/m}$ to ensure that the reproduction number in this population was λ. According to the Reed–Frost model, the number of cases arising in generation $k + 1$ is $X_{k+1} \sim \text{Bin}\left(m + 1 - \sum_0^k X_i, 1 - \rho^{X_k}\right)$, with $X_0 = 1$. These chain binomials were iterated until $X_k = 0$ for some k. Thus the total outbreak size for each outbreak was $\sum X_k$ and the number of generations of spread was $\min\{k : X_{k+1} = 0\}$. To simulate outbreak durations, we sampled individual serial intervals from a gamma distribution with mean 14 days and shape 2, and hence standard deviation $14/\sqrt{2} = 9.9$ days. We chose this distribution because it corresponds to the sum of two independent exponential variables with mean 7 days, representing the latent and infectious periods. The large standard deviation gives a stringent test of performance. We then estimated λ by stochastic imputation using equations (2.7) and (2.8) and using the same gamma serial interval distribution, which we assumed known.

In the following simulation experiments, 100 outbreaks were generated for fixed values of m and λ. It was assumed that no observations were censored. For each set of 100 outbreaks, λ was estimated by maximum likelihood based on total outbreak size, on number of generations of spread and on outbreak duration. The analysis based on numbers of generations of spread corresponds to the serial interval being a known constant, and provides a contrast with the assumption of gamma serial intervals. When $m \longrightarrow \infty$, the offspring distribution tends to a Poisson distribution with mean λ. Accordingly, we used the likelihoods based on Poisson offspring. This whole procedure was then repeated 100 times to obtain average estimates.

Figure 3 shows the average estimates plotted against the true λ, for values of $\lambda = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 1$, and initial susceptible population sizes $m = 10, 100, 1000, 10\,000, 100\,000$. Unsurprisingly, when the pool of susceptibles is only 10, then the estimates are seriously biased. However, for $m \geqslant 100$, the bias is moderate. There is little difference in the bias according to whether outbreak sizes, numbers of generations or duration are used.

To get an idea of the kind of population in which the model would be valid, suppose that vaccine coverage and efficacy are both about 95%. Then about 10% remain susceptible, so the total population size would need to be in excess of about 1000.

3.2 *Under-ascertainment of cases*

We repeated the simulations assuming individual case ascertainment probabilities $p = 0.5$ and 0.75. Thus, after simulating the outbreaks, we picked cases with probability p. Non-ascertainment of the index case may change the observed value of s, the number of index cases in the outbreak. We assumed, however, that the cases ascertained were still regarded as part of the same outbreak. For example, the outbreak represented in Figure 4 starts from a single case A, so that $s = 1$, and involves a total 12 cases in 5 generations of spread and duration A to B. However, the observed outbreak starts from cases C and D, so that $s = 2$, and involves a total 7 cases. The duration for the observed outbreak is the time interval D to E. This method of introducing incomplete asertainment is realistic when, as described above, outbreaks are defined in terms of spatio-temporal clustering within defined communities, rather than by explicitly identifying the chains of transmission.

The plots of the mean estimated values versus the true values of λ, for susceptible populations of initial size $m = 10\,000$, using the outbreak size and duration with ascertainment probabilities $p = 0.75$ and 0.5 are shown in Figures 5(a) and (b), respectively. For comparison, the results with complete ascertainment of cases are also shown.

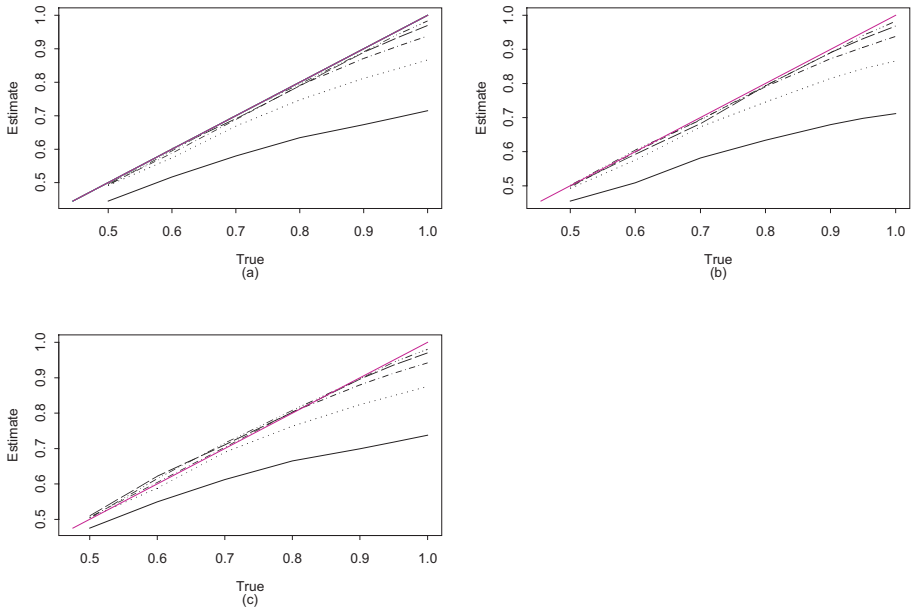

Fig. 3. Plots of average simulation estimates against true λ with population sizes $n = 10$ ——, 100 ....., 1000 ._._._., 10000 – – – –, 100000 ._..._..._.. The line $\widehat{\lambda} = \lambda$ is also shown. (a) Outbreak size, (b) number of generations, (c) outbreak duration.

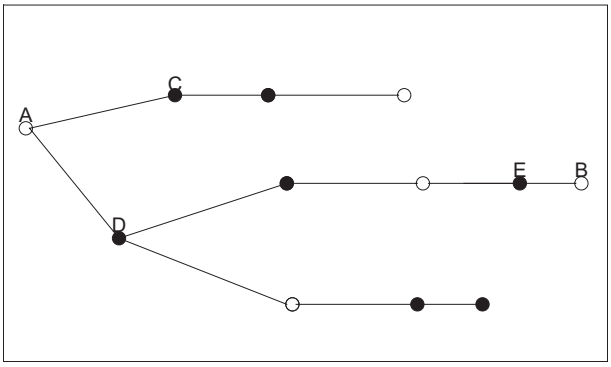

Fig. 4. Schematic representation of an outbreak. ●: case ascertained; ○: case not ascertained.

Figure 5 shows that the bias due to under-ascertainment reduces as λ increases within the range 0 to 1. Also, estimation based on outbreak size is more severely biased than ascertainment based on outbreak duration, the difference being particularly marked for larger values of λ. This is not unexpected, since under-ascertainment always affects observed outbreak size, but may sometimes not affect observed outbreak duration. Similar results were obtained using other population sizes m. For some infections, such as mumps or whooping cough, not all infected individuals exhibit clinical symptoms and hence only a proportion are likely to be ascertained. For such infections, estimation based on outbreak duration may yield results that are less biased than those based on outbreak size.

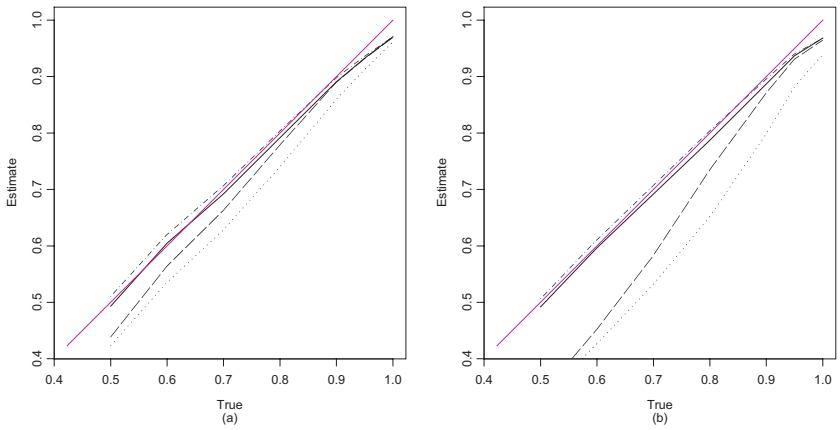

Fig. 5. Average simulation estimates against true values for λ. Outbreak size: —— complete ascertainment, ... under ascertainment. Duration: _ _ _ _ _ complete ascertainment, - - - under ascertainment. Ascertainment probability: (a) 0.75 and (b) 0.5. The line $\widehat{\lambda} = \lambda$ is also shown.

4. EXAMPLE: MEASLES IN THE USA

We illustrate the methods with a dataset on measles in the USA. The dataset includes outbreaks occurring between 1997 and 1999, observed until the end of 1999. No outbreaks were censored. Further details of the data collection are given in Gay *et al.* (2002, to appear). In all the examples based on Bayesian methods we report results using a neutral prior distribution with $\sigma = 1$. We repeated the calculations with non-neutral prior distributions with $\mu = \pm 0.5$; in all cases the results were insensitive to the choice of prior.

So far we have used the term 'outbreak' to include importations of infected individuals not resulting in any further spread. In this study, outbreaks were defined to include at least one generation of spread and hence at least two cases including the introductory case. Importations not involving secondary spread were excluded, because most are singletons and hence are likely to be under-reported. The likelihoods must therefore be conditioned on spread having occurred. The dataset comprises 41 outbreaks, each originating from a single case.

4.1 *Outbreak size data*

The total number of cases in the 41 outbreaks was 207. The frequency distribution of outbreak sizes (minimum 2 cases, maximum 33) is shown in Figure 6. All likelihoods are conditioned on the event that each of the outbreaks involves at least two cases, including the introductory case. The probability of this event is $\left(1 - \mathrm{e}^{-\lambda}\right)^{41}$ for the Poisson offspring distribution and $\lambda^{41}\,(1+\lambda)^{-41}$ for the geometric offspring distribution.

For the measles data the log likelihood kernels are therefore $166 \log(\lambda) - 207\lambda - 41 \log\left(1 - \mathrm{e}^{-\lambda}\right)$ assuming Poisson offspring, and $125 \log(\lambda) - 332 \log(1+\lambda)$ assuming geometric offspring. Figure 7(a) shows the log likelihood profiles. The estimates and profile 95% confidence intervals are $\widehat{\lambda} = 0.66$ $(0.55, 0.78)$ assuming Poisson offspring, and $\widehat{\lambda} = 0.60$ $(0.48, 0.75)$ assuming geometric offspring. The results using the Poisson and geometric models differ slightly, owing in part to the effect of conditioning. The fit was very good ($p > 0.4$) for both models, as assessed by Chi-squared tests.

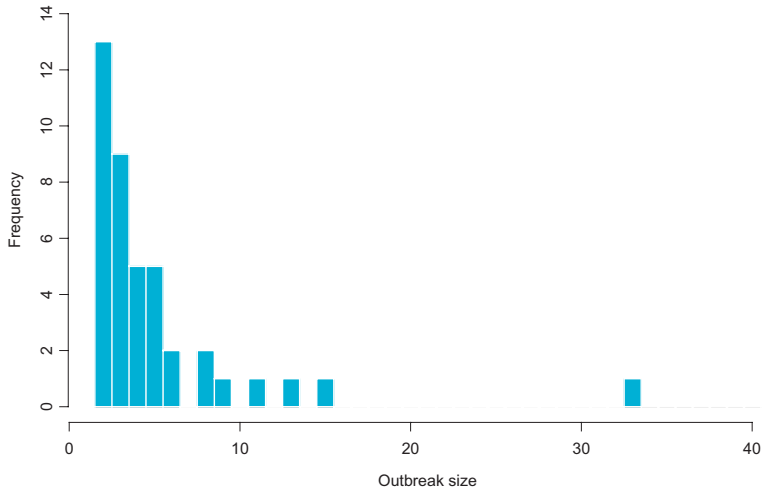

Fig. 6. Frequency distribution of 41 measles outbreaks with secondary spread.

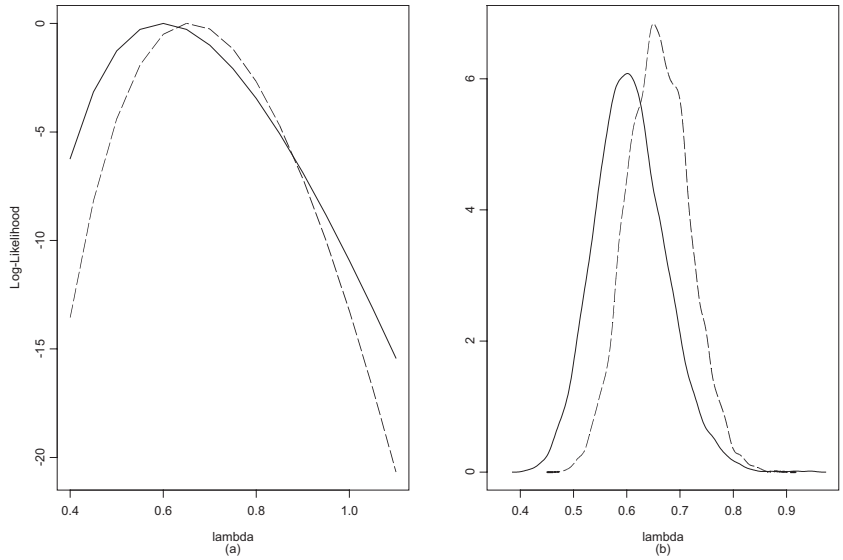

Fig. 7. Estimation from outbreak size from outbreaks of 2 or more cases: (a) log likelihoods for λ; (b) posterior distribution for λ. —— geometric offspring, - - - Poisson offspring.

Figure 7(b) shows the posterior distribution of λ using outbreak size. Clearly, the posterior probability that $\lambda > 1$ is negligible; the corresponding prior probability was 0.5. Both likelihood and Bayesian criteria lead us to conclude that elimination of measles was maintained.

4.2 *Outbreak duration data*

The 41 outbreak durations, calculated as the number of days between the first and last onsets, are shown in Table 1.

Table 1. *Durations (days) of 41 measles outbreaks with secondary spread*

Duration	9	10	11	12	13	14	15	17	19	20
Frequency	1	2	4	3	6	4	4	1	1	3
Duration	21	27	28	33	35	37	46	65	66	100
Frequency	1	1	2	1	1	2	1	1	1	1

All likelihoods are conditioned on the number of generations in each outbreak being at least one, and hence on $U > 0$. Thus the densities $f(u; s)$ in (2.8) are replaced by

$$f(u; s) = \begin{cases} \displaystyle\sum_{n=1}^{\infty} h_n(u) \frac{p_n(\lambda; s)}{1 - p_0(\lambda; s)} & 0 < u < \infty \\ \dfrac{1 - q(\lambda)^s}{1 - p_0(\lambda; s)} & u = \infty \end{cases}$$

where the terms $1 - p_0(\lambda; s)$ are $1 - \mathrm{e}^{-s\lambda}$ and $1 - (1 + \lambda)^{-s}$, respectively, for Poisson and geometric offspring distributions.

In order to model outbreak durations we first need an estimate of the serial interval distribution for measles. One approach is to use the durations of outbreaks involving two cases; there were 13 such outbreaks, with a mean serial interval of 12.6 days. However, 13 outbreaks are too few to obtain a reliable estimate of the serial interval distribution. We therefore use Hope Simpson's data on measles in 264 households of two (Bailey, 1975). This much-used data set has been analysed by Bailey (1975), Becker (1989) and more recently by O'Neill *et al.* (2000), who estimate the infectious period to be very short (2–3 days). Two measles cases were observed in 219 out of the 264 households. We shall exclude the 32 households in which the two cases were separated by 5 days or less; these pairs of cases are most probably co-primary. Fitting a gamma distribution to the remaining 187 observations on the serial interval T gives the distribution shown in Figure 8. We also fitted a gamma distribution to $T - \delta$ for some value δ to be estimated; this produced a nearly identical fit (see Figure 8). The fitted gamma distribution has mean 11.03 days and shape 20.68. The standard deviation of $11.03/\sqrt{20.68} = 2.42$ days is quite short, so that there should be relatively little overlap between generations. In what follows we shall assume that the serial interval follows a gamma distribution with these parameter values. The models fitted to the 41 outbreak durations gave the profile log likelihoods and posterior distributions shown in Figure 9. The estimates and 95% profile confidence intervals are $\widehat{\lambda} = 0.53$ $(0.40, 0.68)$ based on Poisson offspring, and $\widehat{\lambda} = 0.56$ $(0.42, 0.73)$ based on geometric offspring. The posterior probability that $\lambda > 1$ is negligible; again, the corresponding prior probability was 0.5. As with the analysis based on outbreak sizes, we conclude that elimination of measles was maintained in this population.

We also fitted a simpler model in which the numbers of generations were imputed directly from the mean serial interval of 11 days, outbreak durations between 5.5 and 16.5 days corresponding to one generation of spread, those between 16.5 and 27.5 days corresponding to two generations of spread and so on. The results were very similar.

5. Discussion

The statistical methods we have described provide a simple framework for analysing surveillance data for infectious diseases after elimination of sustained endemic transmission, a situation characterized by maintenance of the effective reproduction number λ below 1. Epidemiological aspects of the methods and their practical implementation have been discussed in De Serres *et al.* (2000). In the measles application

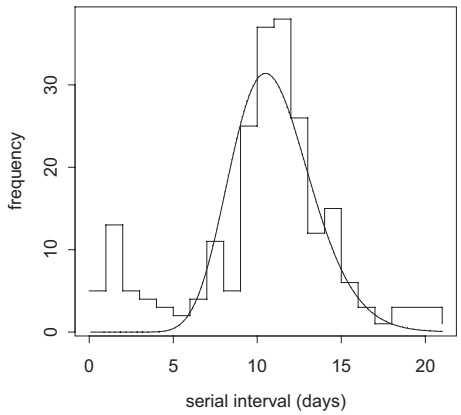

Fig. 8. Distribution of serial interval for measles: observed frequencies, gamma (——) and shifted gamma (- - -) fitted to intervals > 5 days.

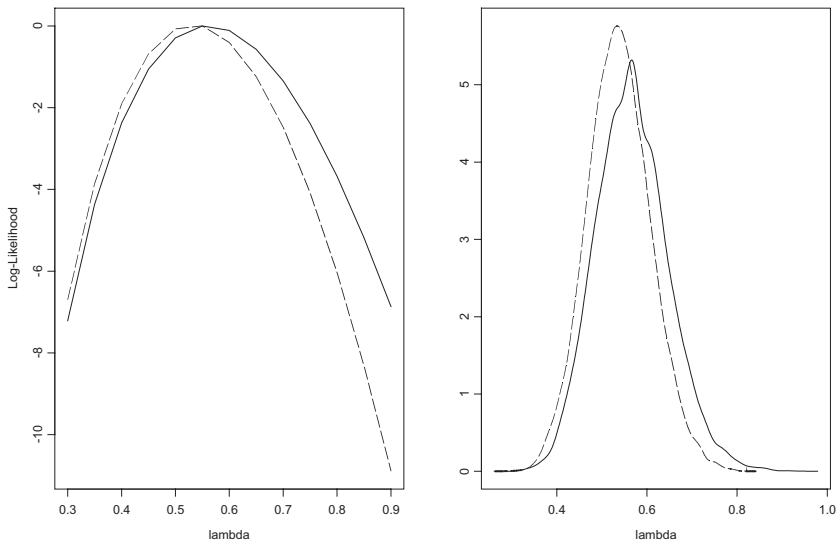

Fig. 9. Estimation from outbreak duration:(a) log likelihoods for λ ; (b) posterior distribution for λ. —— geometric offspring, - - - Poisson offspring.

described above, all three methods, based on outbreak size and outbreak duration, yield the same broad results: measles is well-controlled in this population.

Branching processes have a long pedigree in infectious disease epidemiology, but are seldom used directly for statistical modelling in public health applications. The particular context of estimation in the presence of mass vaccination programmes gives these methods new relevance. In this paper we used a modelling framework based on outbreak sizes and durations. By avoiding conditioning on extinction, the cases $\lambda \leqslant 1$ and $\lambda > 1$ can be considered together. Our models are similar to censored survival models.

Several methodological issues are worthy of further investigation. (a) De Serres *et al.* (2000) also proposed basing surveillance of λ on the proportion of cases that are imported. Suppose that in a time

interval $[0, \tau]$ a total x cases are observed, of which s are found to have been imported. If $\lambda \leqslant 1$, all cases must originate from importations. If end effects due to censoring at 0 and τ are ignorable, we may regard x as corresponding approximately to an observation on the total outbreak size from s cases, and hence estimate $\widehat{\lambda} = 1 - s/x$. It would be useful to cast this method in a framework valid for all values of λ so as to derive surveillance thresholds. (b) In his elegant paper Yanev (1975) described, amongst other things, the asymptotic properties of the MLE $\widehat{\lambda} = 1 - s/x$ based on outbreak size in the absence of censoring when $\lambda = 1$. In this special case, standard regularity conditions no longer hold. The MLE $\widehat{\lambda}$ is still asymptotically unbiased and consistent, and $s(1 - \widehat{\lambda}) \longrightarrow_D \sigma^2 \chi_1^2$ as $s \longrightarrow \infty$, where σ^2 is the variance of the offspring distribution. For example, for Poisson offspring, when $\lambda = 1$,

$$\mathbb{E}(\widehat{\lambda}) = \frac{s}{s+1} \quad \text{and} \quad V(\widehat{\lambda}) = \frac{2s}{(s+1)^2(s+2)}.$$

It would be interesting to obtain corresponding asymptotic results for the MLE based on numbers of generations of spread and durations with $\lambda = 1$ when the number of outbreaks $n \longrightarrow \infty$. (c) Our derivation of the likelihood for outbreak durations was based on the assumption that the first and last case are necessarily those separated by the maximum number of generations. It would be useful to relax this assumption. (d) The presence of censoring destroys the sufficiency of outbreak size. It would be interesting to know, in situations where censoring occurs frequently, how much gain in efficiency would be achieved by jointly modelling outbreak size and duration. (e) Finally, it would be useful to extend the methods to incorporate heterogeneity in contact rates. For example, λ could be allowed to vary between outbreaks according to some distribution reflecting the variation in contact rates, and hence λ, between communities. This would be useful if, for instance, the model with fixed λ were found not to provide a good fit to the data. Note that in this case a new surveillance criterion would need to be specified, possibly based on an upper quantile of the distribution of λ.

Clearly, our methods are limited in scope, in that they do not take account of depletion of susceptibles and heterogeneities in contact rates. However, they provide a simple surveillance tool, and as such have been welcomed by practitioners (Hinman *et al.*, 2000). As we have shown, bias due to ignoring the depletion of susceptibles is only really a problem in communities with fewer than 100 susceptibles. If, say, 10% of the population are susceptible, then the methods are applicable with little bias in communities in excess of 1000. Allowing for heterogeneities in contact rates, on the other hand, is difficult, though the random λ model suggested above provides one simple method. Various approaches to take account of mixing within and between households have recently been suggested (Ball *et al.*, 1997; Ball and Lyne, 2001), though it seems rather unlikely that they can be applied in a surveillance setting. Alternatively, serological survey methods provide a contrasting population-based approach to the surveillance of vaccination programmes (Gay *et al.*, 1995; Farrington *et al.*, 2001).

ACKNOWLEDGEMENTS

We are grateful to Kevin McConway and Pia Veldt Larsen for discussing some of the issues raised in this paper. We also thank the referees of this and an earlier version of the paper for useful comments. This research was supported by Wellcome Trust project grant 061830.

REFERENCES

ANDERSON, R. M. AND MAY, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

BAILEY, N. T. J. (1975). *Mathematical Theory of Infectious Diseases, 2nd edn*. London: Griffin.

BALL, F. AND DONNELLY, P. (1995). Strong approximations for epidemic models. *Stochastic Processes and their Applications* **55**, 1–21.

BALL, F. AND LYNE, O. (2001). Stochastic multi-type SIR epidemics among a population partitioned into households. *Advances in Applied Probability* **33**, 99–123.

BALL, F., MOLLISON, D. AND SCALIA-TOMBA, G. (1997). Epidemics with two levels of mixing. *The Annals of Applied Probability* **7**, 46–89.

BECKER, N. (1974a). On parametric estimation for mortal branching processes. *Biometrika* **61**, 393–399.

BECKER, N. (1974b). On assessing the progress of epidemics. Williams, E. (ed.), *Studies in Probability and Statistics: Papers in Honour of Edwin J. G. Pitman*. Amsterdam: North-Holland, pp. 135–141.

BECKER, N. (1976). Estimation for an epidemic model. *Biometrics* **32**, 769–777.

BECKER, N. (1977). Estimation for discrete time branching processes with applicaqtions to epidemics. *Biometrics* **33**, 515–522.

BECKER, N. G. (1989). *Analysis of Infectious Disease Data*. London: Chapman and Hall.

CARACO, T., DURYEA, M., GARDNER, G., MANIATTY, W. AND SZYMANSKI, B. (1998). Host spatial heterogeneity and extinction of an SIS epidemic. *Journal of Theoretical Biology* **192**, 351–361.

CLANCY, D. AND O'NEILL, P. (1998). Approximation of epidemics by inhomogeneous birth-and-death processes. *Stochastic processes and their Applications* **73**, 233–245.

DE SERRES, G., GAY, N. J. AND FARRINGTON, C. P. (2000). Epidemiology of transmissible diseases after elimination. *American Journal of Epidemiology* **151**, 1039–1048.

DIETZ, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research* **2**, 23–41.

FARRINGTON, C. P. AND GRANT, A. D. (1999). The distribution of time to extinction in subcritical branching processes: Applications to outbreaks of infectious disease. *Journal of Applied Probability* **36**, 771–779.

FARRINGTON, C. P., KANAAN, M. N. AND GAY, N. J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Applied Statistics* **50**, 251–292.

FROST, W. (1976). Some conceptions of epidemics in general. *American Journal of Epidemiology* **103**, 141–151.

GAY, N. J., DE SERRES, G., FARRINGTON, C. P., REDD, S. AND PAPANIA, M. (2002). Elimination of measles from the United States: An assessement through basic surveillance data. *Journal of Infectious Diseases Supplement*, to appear

GAY, N. J., HESKETH, L. M., MORGAN-CAPNER, P. AND MILLER, E. (1995). Interpretation of serological surveillance data for measles using mathematical models: Implications for vaccine strategy. *Epidemiology and Infection* **115**, 139–156.

GUTTORP, P. (1991). *Statistical Inference for Branching Processes*. New York: Wiley.

HAIGHT, F. AND BREUER, M. A. (1960). The Borel–Tanner distribution. *Biometrika* **47**, 143–150.

HEYDE, C. C. (1979). On assessing the potential severity of an outbreak of a rare infectious disease: A Bayesian approach. *Australian Journal of Statisitics* **21**, 282–292.

HINMAN, A. R., ORENSTEIN, W. A. AND PAPANIA, M. J. (2000). Invited commentary: Epidemiology of transmissible diseases after elimination. *American Journal of Epidemiology* **151**, 1049–1052.

JACOB, C. AND PECCOUD, J. (1998). Estimation of the parameters of a branching process from migrating binomial observations. *Advances in Applied Probability* **30**, 948–967.

JOHNSON, N. L., KOTZ, S. AND KEMP, A. W. (1992). *Univariate Discrete Distributions, 2nd edn*. Chichester: Wiley.

KOTZ, S. AND JOHNSON, N. L. (EDS) (1986). *Encyclopedia of Statistical Sciences, Vol. 7*. Chichester: Wiley.

LOCKHART, R. (1982). On the nonexistence of consistent estimates in Galton–Watson branching processes. *Journal of Applied Probability* **19**, 842–846.

MARSCHNER, I. C. (1992). The effect of preferential mixing on the growth of an epidemic. *Mathematical Biosciences* **109**, 39–67.

MULLER, J. S. B. AND KIRKILIONIS, M. (2000). Ring vaccination. *Journal of Mathematical Biology* **41**, 143–171.

O'NEILL, P. D., BALDING, D. J., BECKER, N. G., EEROLA, M. AND MOLLISON, D. (2000). Analyses of infectious disease data from household outbreaks by Markov Chain Monte Carlo methods. *Applied Statistics* **49**, 517–542.

SRIRAM, T. N. (1991). On the uniform strong consistency of an estimator of the offspring mean in a branching process with immigration. *Statistics and Probability Letters* **12**, 151–155.

WAUGH, W. A. O. (1958). Conditioned Markov processes. *Biometrika* **45**, 241–249.

YANEV, G. P. AND TSOKOS, C. P. (1999). Decision-theoretic estimation of the offspring mean in mortal branching processes. *Communications in Statistics: Stochastic Models* **15**, 889–902.

YANEV, N. M. (1975). On the statistics of branching processes. *Theory of Probability and its Applications* **20**, 612–622.

[*Received 17 April, 2002; revised 16 July, 2002; accepted for publication 15 August, 2002*]