Biometrika (2004), **91,** 2, *pp.* 345–362

Printed in Great Britain

Stochastic multitype epidemics in a community of households: Estimation of threshold parameter R_* and secure vaccination coverage

BY FRANK G. BALL

School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, U.K.

frank.ball@nottingham.ac.uk

TOM BRITTON

Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden

tom.britton@math.su.se

AND OWEN D. LYNE

School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, U.K.

owen.lyne@nottingham.ac.uk

SUMMARY

This paper is concerned with estimation of the threshold parameter R_* for a stochastic model for the spread of a susceptible → infective → removed epidemic among a closed, finite population that contains several types of individual and is partitioned into households. It turns out that R_* cannot be estimated consistently from final outcome data, so a Perron–Frobenius argument is used to obtain sharp lower and upper bounds for R_*, which can be estimated consistently. Determining the allocation of vaccines that reduces the upper bound for R_* to its threshold value of one, thus preventing the occurrence of a major outbreak, with minimum vaccine coverage is shown to be a linear programming problem. The estimates of R_*, before and after vaccination, and of the secure vaccination coverage, i.e. the proportion of individuals that have to be vaccinated to reduce the upper bound for R_* to 1 assuming an optimal vaccination scheme, are equipped with standard errors, thus yielding conservative confidence bounds for these key epidemiological parameters. The methodology is illustrated by application to data on influenza outbreaks in Tecumseh, Michigan.

Some key words: Estimation; Household epidemic; Multitype epidemic; Outbreak data; Stochastic epidemic; Threshold parameter; Vaccination.

1. INTRODUCTION

Epidemic models have a long history going back at least to Bernoulli (1760), who used a mathematical method to evaluate the effectiveness of variolation against smallpox, with the aim of influencing public health policy. By far the most important result to come out of mathematical epidemic theory is the celebrated threshold theorem, which

dates back to the pioneering work by Kermack & McKendrick (1927), and, in modern terminology, states that a major epidemic can occur only if the basic reproduction number R_0 (Heesterbeek & Dietz, 1996) is larger than its threshold value of one. The result is important because it implies the critical vaccination coverage, i.e. the proportion of susceptible individuals that need to be vaccinated in order to prevent an epidemic occurring. However, for it to be practically relevant, it is necessary that modelling assumptions adequately reflect what happens in real-life epidemics. The early models were deterministic and assumed a community of homogeneous individuals who mix uniformly. Subsequently, these models have been extended to take account of stochasticity, individual heterogeneities and social structures that yield non-uniform mixing; see for example Bailey (1975), Anderson & May (1991) and Andersson (1999). In order to determine the critical vaccination coverage in practice, estimates of model parameters are required. Thus procedures for statistical inference have been developed, often focusing on estimation of epidemiologically important parameters, such as the basic reproduction number R_0, both before and after vaccinating a specified proportion of individuals, and the critical vaccination coverage; see for example Anderson & May (1991, Ch. 5, 7), Becker (1989, Ch. 8), Becker & Britton (1999) and Andersson & Britton (2000, Ch. 12).

One departure from homogeneous mixing, that has received considerable interest recently and has an important impact on model behaviour, is that attributable to the household structure of most human populations (Becker & Dietz, 1995; Ball et al., 1997). Most of the work on the so-called households model has assumed only one type of individual but with different rates for within- and between-households infections. However, heterogeneities, such as those corresponding to age, sex and response to vaccine, can have a significant effect on disease spread. Ball & Lyne (2001) studied the probabilistic behaviour of a stochastic multitype susceptible $\rightarrow$ infective $\rightarrow$ removed households model and, in particular, derived a threshold parameter R_*, the households model equivalent of R_0, that determines whether or not a major outbreak can occur; see also Becker & Hall (1996). Statistical inference for this model from final outcome data, possibly only for a sample of households in the community, will be considered in a future paper by Ball and Lyne. However, the between-households infection rates are not identifiable from such data, and consequently neither are the epidemiologically important parameters R_*, before and after a vaccination policy, and the associated critical vaccination coverage. Similar phenomena have previously been observed by Greenhalgh & Dietz (1994) and Britton (2001) for multitype epidemics without household structure.

In the present paper, estimation of the above epidemiologically important parameters is studied for a stochastic model incorporating both household structure and individual heterogeneity, using two different models for vaccine action. In the first model, a vaccinated individual is either rendered completely immune or the vaccine has no effect. In the second model, vaccinated individuals have a reduced probability of infection given exposure to infection. These models are defined in Smith et al. (1984) and, following Halloran et al. (1992), are referred to as 'all-or-nothing' and 'leaky', respectively. The above mentioned identifiability problems are overcome by deriving sharp upper and lower bounds for R_*, both before and after a vaccination scheme, which can be estimated consistently from final outcome data, thus enabling one to estimate the secure vaccination coverage, that reduces the upper bound for R_* to one. Furthermore, all of these estimates are equipped with asymptotic standard errors, as the number of households in the community becomes large, yielding asymptotically conservative confidence intervals. Determination of the allocation of vaccines that reduces the upper bound for R_* to one with minimal vaccine

coverage is shown to be a linear programming problem, in contrast to the case where the infection rates are known, when a complex nonlinear optimisation problem has to be solved, unless between-households infection is proportionate mixing; see § 2·3.

The stochastic multitype susceptible → infective → removed households epidemic model is described in § 2, where its threshold behaviour and final outcome are outlined. The threshold parameters following a vaccination scheme, using the two models for vaccine action, are determined and optimal vaccination schemes are briefly discussed. Estimation of the epidemiologically important parameters is considered in §§ 3 and 4, with point estimates being given in § 3 and uncertainty being treated in § 4. The methodology is illustrated in § 5 by an application to data on influenza outbreaks in Tecumseh, Michigan, and the paper concludes with a brief discussion in § 6.

2. Model, threshold, behaviour and vaccination

2·1. *Model*

The model under consideration in this paper is that of Ball & Lyne (2001) for the spread of a susceptible → infective → removed epidemic among a closed, finite population that contains J classes of individuals, labelled $1, \ldots, J$, and is partitioned into households. Let $\mathscr{J} = \{1, \ldots, J\}$ and

$$\mathscr{N}_0 = \left\{ n = (n_1, \ldots, n_J) \in \mathbb{Z}^J : n_j \geqslant 0 (j \in \mathscr{J}),\ \|n\| = \sum_{j \in \mathscr{J}} n_j \geqslant 1 \right\}.$$

Suppose that, for $n \in \mathscr{N}_0$, the population contains m_n households of category n, where a household of category n contains n_j individuals of class j ($j \in \mathscr{J}$). Let $m = \sum_{n \in \mathscr{N}_0} m_n$ denote the total number of households in the population, let $N_j = \sum_{n \in \mathscr{N}_0} n_j m_n$ denote the total number of individuals of class j in the population ($j \in \mathscr{J}$) and let $N = \sum_{n \in \mathscr{N}_0} \|n\| m_n$ denote the total number of individuals in the population. Assume that N, and hence N_j ($j \in \mathscr{J}$) and m, are finite. Thus $\mathscr{N} = \{n \in \mathscr{N}_0 : m_n > 0\}$ is finite.

The epidemic is initiated by some individuals becoming infected at time $t = 0$, with the remaining individuals in the population all assumed to be susceptible. For $j \in \mathscr{J}$, the infectious periods of class-j infectives are each distributed according to a finite random variable $T_I^{(j)}$, having an arbitrary but specified distribution with mean μ_j. For $i, j \in \mathscr{J}$, throughout his or her infectious period a given class-i infective makes global contacts with any given susceptible of class j in the population at the points of a homogeneous Poisson process having rate λ_{ij}^G / N_j and, additionally, he or she makes local contacts with any given susceptible of class j in his or her own household at the points of a homogeneous Poisson process having rate λ_{ij}^L. Let $\Lambda^L = (\lambda_{ij}^L)$ and $\Lambda^G = (\lambda_{ij}^G)$. All the Poisson processes describing infectious contacts, whether or not either or both of the individuals involved are the same, as well as the random variables describing infectious periods, are assumed to be mutually independent. A susceptible becomes infective as soon as it is contacted by an infective and is removed, and plays no further part in the epidemic, at the end of its infectious period. The epidemic ceases as soon as there is no infective present in the population; see Ball & Lyne (2001) for further discussion of this model and for rigorous proofs of the informal arguments outlined below.

2·2. *Threshold behaviour and final outcome*

Suppose that the number of households m is large. Then, during the early stages of an epidemic initiated by a small number of infectives, the probability that a global contact

is with an individual residing in a previously infected household is small. Thus the initial growth of the epidemic can be approximated by a process in which each global contact is with an individual in an otherwise completely susceptible household. The process of infected households in this approximating process follows a multitype branching process, with type space $\mathscr{J}$, where the type of an infected household is given by the class of its initial, globally contacted, infective. Let $M=(m_{ij})$, where, for $i,j\in\mathscr{J}$, m_{ij} is the mean number of class-j global contacts that emanate from a typical type-i infected household. Suppose that M is positively regular; that is $0\leqslant m_{ij}<\infty$ $(i,j\in\mathscr{J})$ and there exists $n\in\mathbb{N}$ such that all the elements of M^n are strictly positive. Let R_* denote the maximal eigenvalue of M. Then the approximating branching process has extinction probability strictly less than one if and only if $R_*>1$ (Mode, 1971, Ch. 1, Theorem 7.1). Thus, for large m, the probability of a global epidemic, i.e. one infecting a large number of households, is nonzero if and only if $R_*>1$.

In order to compute R_*, expressions for m_{ij} $(i,j\in\mathscr{J})$ are required. For $n\in\mathscr{N}$, let $\alpha_n=m_n/m$ denote the proportion of households in the population that have category n and, for $i\in\mathscr{J}$ and $n\in\mathscr{N}$, let $\alpha_i(n)=n_i m_n/N_i$ be the probability that a class-i individual chosen at random in the population resides in a household of category n. Consider a completely susceptible household of category n and suppose that a class-i individual residing in that household is contacted globally. That class-i individual will start a realisation of a single-household epidemic, whose internal dynamics are determined purely by local infection since, in the branching process approximation, all global contacts are with individuals in completely susceptible households. For $j\in\mathscr{J}$, let Y_j denote the number of class-j individuals that are ultimately infected by this single-household epidemic, including the initial infective if $j=i$, and let T_j^A denote the sum of the infectious periods of those Y_j class-j infectives. Let $\mu_{n,i,j}(\Lambda^L)=E(Y_j)$ and note that, by Wald's identity for multitype susceptible $\rightarrow$ infective $\rightarrow$ removed epidemics (Ball, 1986, Corollary 3.2), $E(T_j^A)=E(T_I^{(j)})\mu_{n,i,j}(\Lambda^L)=\mu_j\mu_{n,i,j}(\Lambda^L)$. During the above single-household epidemic, for $k\in\mathscr{J}$, each class-k infective makes class-j global contacts at total rate λ_{kj}^G, so the expected total number of class-j global contacts that emanate from this single-household epidemic is $\sum_{k\in\mathscr{J}}\mu_k\mu_{n,i,k}(\Lambda^L)\lambda_{kj}^G$. Finally, conditioning on the household category of a typical type i infected household yields

$$m_{ij}=\sum_{n\in\mathscr{N}}\alpha_i(n)\sum_{k\in\mathscr{J}}\mu_{n,i,k}(\Lambda^L)\mu_k\lambda_{kj}^G\quad(i,j\in\mathscr{J}).$$

An algorithm for computing $\mu_{n,i,j}(\Lambda^L)$ $(n\in\mathscr{N}_0;\,i,j\in\mathscr{J})$ is given in the Appendix.

Suppose that the number of households m is large, the number of initial infectives is small and a global epidemic occurs. For $j\in\mathscr{J}$, let z_j denote the expected proportion of class-j susceptibles that are ultimately infected and let T_j denote the sum of the infectious periods of all the class-j infectives during the epidemic. Now $T_j\eqsim N_j z_j E(T_I^{(j)})=N_j z_j\mu_j$, so $T_j/N_i\eqsim\gamma_j z_j\mu_j/\gamma_i$, where $\gamma_i=N_i/N$ is the proportion of class-i individuals in the population $(i\in\mathscr{J})$. Thus the probability that a given class-i susceptible avoids global infection during the epidemic is approximately given by

$$\pi_i=\exp\left(-\sum_{j\in\mathscr{J}}\gamma_j z_j\mu_j\lambda_{ji}^G/\gamma_i\right)\quad(i\in\mathscr{J}),\tag{2·1}$$

since this individual is contacted globally by a given class-j infective at rate $N_i^{-1}\lambda_{ji}^G$. For large m, distinct individuals avoid global infection approximately independently of each other. Thus the ultimate spread of infection within a household having category n, that

did not contain any initial infectives, is approximately distributed as that of a multitype single-household epidemic model, $E_n(\Lambda^L, \pi)$ say, where $\pi = (\pi_1, \ldots, \pi_J)$, studied by Addy et al. (1991), in which, in addition to local infection, during the course of the epidemic individuals avoid infection from outside the household independently and with probability π_i for a class-i susceptible ($i \in \mathscr{J}$). For $i \in \mathscr{J}$, let $\mu_{n,i}(\Lambda^L, \pi)$ be the expected number of class-i individuals that are ultimately infected by $E_n(\Lambda^L, \pi)$. An algorithm for computing $\mu_{n,i}(\Lambda^L, \pi)$ ($n \in \mathscr{N}_0$, $i \in \mathscr{J}$) is given in the Appendix.

For $i \in \mathscr{J}$, z_i can be interpreted as the probability that a randomly chosen initial class-i susceptible is ultimately infected by the epidemic. If we condition on the category of household in which this initial susceptible resides and note that if it resides in a household of category n then its chance of ultimate infection is $\mu_{n,i}(\Lambda^L, \pi)/n_i$, it follows that

$$z_i = \sum_{n \in \mathscr{N}} \alpha_i(n) \mu_{n,i}(\Lambda^L, \pi)/n_i \quad (i \in \mathscr{J}), \tag{2·2}$$

which, together with (2·1), is a set of J implicit equations for $z = (z_1, \ldots, z_J)$. Note that $z = 0$ is a root of (2·2). It is shown in Ball & Lyne (2001, § 5.2), that, provided the $J \times J$ matrix A having elements

$$a_{ij} = \sum_{k \in \mathscr{J}} \mu_i \lambda^G_{ik} \sum_{n \in \mathscr{N}} \alpha_k(n) \mu_{n,k,j}(\Lambda^L) \quad (i, j \in \mathscr{J})$$

is positively regular, if $R_* \leqslant 1$ then $z = 0$ is the only solution of (2·2) in $[0, 1]^J$, while if $R_* > 1$ then there is a unique second root, with $z_i > 0$ ($i \in \mathscr{J}$), yielding the expected proportion of individuals of different classes that are infected by a global epidemic.

2·3. *Vaccination*

Consider first the case of all-or-nothing vaccines and suppose that vaccinated individuals are rendered immune independently, with probability ε_i for a class-i individual ($i \in \mathscr{J}$). For $n \in \mathscr{N}$ and $0 \leqslant r = (r_1, \ldots, r_J) \leqslant n$, where inequalities between vectors are to be interpreted elementwise, let $v_{n,r}$ denote the proportion of households of category n that have had r members vaccinated, and let $v = \{v_{n,r} : n \in \mathscr{N}, 0 \leqslant r \leqslant n\}$.

For $i, j \in \mathscr{J}$, let $m_{ij}(v)$ denote the expected number of class-j global contacts that emanate from a single-household epidemic that is initiated by a randomly chosen class-i individual being contacted globally. The probability that a randomly chosen class-i individual resides in a household of category n having r members vaccinated is $\alpha_i(n) v_{n,r}$. For $n - r \leqslant s \leqslant n$, such a household has s susceptible individuals if $n - s$ of the vaccinations are successful, which happens with probability $C(r, n-s)\varepsilon^{n-s}(1-\varepsilon)^{r-n+s}$, where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_J)$, 1 is a vector of J ones, $C(r, n-s) = \prod_{l \in \mathscr{J}} {}^{r_l}C_{n_l - s_l}$ with ${}^{r_l}C_{n_l - s_l}$ denoting the usual binomial coefficient and, for two vectors x, y of length J, $x^y = \prod_{l \in \mathscr{J}} x_l^{y_l}$. Furthermore, given that s individuals in the household are susceptible, the probability that a global contact with a class-i individual in that household is with a susceptible, and thus triggers a local household epidemic, is s_i/n_i. Hence, for $i, j \in \mathscr{J}$,

$$m_{ij}(v) = \sum_{n \in \mathscr{N}} \alpha_i(n) \sum_{r=0}^{n} v_{n,r} \sum_{s=n-r}^{n} C(r, n-s)\varepsilon^{n-s}(1-\varepsilon)^{r-n+s} \frac{s_i}{n_i} \sum_{l \in \mathscr{J}} \mu_{s,i,l}(\Lambda^L) \mu_l \lambda^G_{lj}, \tag{2·3}$$

where for example $\sum_{r=0}^{n} = \sum_{r_1=0}^{n_1} \cdots \sum_{r_J=0}^{n_J}$.

Let $M(v) = (m_{ij}(v))$ and $R_*^{\mathrm{AoN}}(v)$ be the maximal eigenvalue of $M(v)$. Then $R_*^{\mathrm{AoN}}(v)$ is a threshold parameter for the epidemic after vaccination with an all-or-nothing vaccine, in the sense that a global epidemic can occur only if $R_*^{\mathrm{AoN}}(v) > 1$. Consequently, a vaccination scheme v having $R_*^{\mathrm{AoN}}(v) \leqslant 1$ is protective for the whole community.

In general, there is no closed-form expression for $R_*^{\mathrm{AoN}}(v)$. However, if the global infection rates take the proportionate mixing form (Hethcote & Van Ark, 1987; Becker & Marschner, 1990) $\lambda_{ij}^G = \alpha_i^G \beta_j^G$ $(i, j \in \mathscr{J})$, then the matrix $M(v)$ has rank one, so $R_*^{\mathrm{AoN}}(v)$ is given by its trace, i.e.

$$R_*^{\mathrm{AoN}}(v) = \sum_{i \in \mathscr{J}} \sum_{n \in \mathscr{N}} \alpha_i(n) \sum_{r=0}^{n} v_{n,r} \sum_{s=n-r}^{n} C(r, n-s) \varepsilon^{n-s} (1-\varepsilon)^{r-n+s} \frac{S_i}{n_i} \sum_{i \in \mathscr{J}} \mu_{s,i,l}(\Lambda^L) \mu_l \alpha_l^G \beta_i^G.$$

Consider now the case of leaky vaccines and suppose that, for all $j \in \mathscr{J}$, all the infection rates to vaccinated class-j individuals are reduced by a factor ε_j. Hence, for $i, j \in \mathscr{J}$, the rate at which a class-i infective has global, respectively local, contact with a vaccinated class-j individual is $\lambda_{ij}^G(1-\varepsilon_j)/N_j$, respectively $\lambda_{ij}^L(1-\varepsilon_j)$. Note that the average vaccine efficacy is the same as in the all-or-nothing case.

After a vaccination scheme, there may be $2J$ types of individual in the population, i.e. vaccinated and unvaccinated individuals for each of the J original classes. Let $\mu_{n-r,r,\mathrm{u}:i,l}(\Lambda^L, \varepsilon)$, respectively $\mu_{n-r,r,\mathrm{v}:i,l}(\Lambda^L, \varepsilon)$, denote the expected number of infected class-l individuals, counting both vaccinated and unvaccinated individuals, in a category-n household having r vaccinated, and hence $n-r$ unvaccinated, individuals, initiated by an infectious unvaccinated, respectively vaccinated, class-i individual, neglecting further outside infections.

As for all-or-nothing vaccines, for $i, j \in \mathscr{J}$, let $m_{ij}(v)$ be the expected number of global contacts with class-j individuals that emanate from a single-household epidemic that is initiated by a randomly chosen class-i individual being contacted globally. If such a globally contacted class-i individual happens to be vaccinated, the chance that he or she will actually become infected is $(1-\varepsilon_i)$, whereas this chance is 1 if the individual is unvaccinated. Thus, in a household having r_i vaccinated and $n_i - r_i$ unvaccinated class-i individuals, such a contact will result in infection of an unvaccinated individual with probability $(n_i - r_i)/n_i$ and in infection of a vaccinated individual with probability $r_i(1-\varepsilon_i)/n_i$. Consequently

$$m_{ij}(v) = \sum_{n \in \mathscr{N}} \alpha_i(n) \sum_{r=0}^{n} v_{n,r} \sum_{k \in \mathscr{J}} \left\{ \frac{n_i - r_i}{n_i} \mu_{n-r,r,\mathrm{u}:i,k}(\Lambda^L, \varepsilon) + \frac{r_i(1-\varepsilon_i)}{n_i} \mu_{n-r,r,\mathrm{v}:i,k}(\Lambda^L, \varepsilon) \right\} \mu_k \lambda_{kj}^G, \tag{2·4}$$

and $R_*^{\mathrm{Le}}(v)$, the threshold parameter after vaccination with a leaky vaccine, is the maximal eigenvalue of the matrix $M(v) = (m_{ij}(v))$. As before, an explicit expression for $R_*^{\mathrm{Le}}(v)$ is available when Λ^G takes the proportionate mixing form.

As noted above, the main aim of any vaccination scheme is to bring the threshold parameter below one, i.e. to ensure that $R_*(v) \leqslant 1$. The threshold parameter following the vaccination scheme v is referred to generically as $R_*(v)$; $R_*(v)$ is $R_*^{\mathrm{AoN}}(v)$ if the vaccine is all-or-nothing and $R_*^{\mathrm{Le}}(v)$ if it is leaky. Therefore, for a given community and a given vaccine response, the vaccination scheme v is said to be preventive, written $v \in P$, if the induced threshold parameter satisfies $R_*(v) \leqslant 1$.

If the vaccine response, or efficacy, ε is not large enough, it could happen that no vaccination scheme is preventive; that is $R_*(v) > 1$ even when the whole population is vaccinated. On the other hand, if the vaccine response is large enough there will be many different vaccination schemes v satisfying $R_*(v) \leqslant 1$. It is then important to determine which such scheme is the best in the sense that it requires the fewest vaccinations.

Accordingly, if

$$S(v) = \frac{\sum_{n \in \mathscr{N}} \sum_{r=0}^{n} \|r\| v_{n,r} \alpha_n}{\sum_{n \in \mathscr{N}} \|n\| \alpha_n} \tag{2·5}$$

denotes the proportion of the population that are vaccinated, i.e. the overall vaccination coverage, under the scheme v, then any scheme, v_{opt} say, satisfying $S(v_{\mathrm{opt}}) \leqslant S(v)$ for all $v \in P$ is optimal. The definition of v_{opt} could be generalised to incorporate costs associated with the practical implementation of a vaccination scheme, for example by including an additional cost per household of having individuals vaccinated (Ball & Lyne, 2002).

It is a nontrivial problem to derive v_{opt}, particularly since, in general, $R_*(v)$ does not admit a closed-form expression. However, if the global infection rates take the proportionate mixing form then $R_*(v)$ and $S(v)$ are both linear functions of v, so determining the allocation of vaccines which (a) minimises $R_*(v)$ subject to an upper bound on $S(v)$ or (b) minimises $S(v)$ subject to $R_*(v) \leqslant 1$ is in each case a linear programming problem; compare Becker & Starczak (1997). Note that there are further, linear constraints on v implicit in the above formulations, specifically that, for $n \in \mathscr{N}$, $v_{n,r} \geqslant 0$ $(0 \leqslant r \leqslant n)$ and $\sum_{r=0}^{n} v_{n,r} = 1$.

3. Estimation

3·1. *Estimation of local and global infection parameters*

In order to estimate the threshold parameter $R_*(v)$ associated with any given vaccination scheme, and to design vaccination strategies that prevent global epidemics with minimal vaccination coverage, it is necessary to have estimates of the local and global infection parameters. These parameters are assumed to be unknown and are to be estimated from data on one previous outbreak in the population. The distributions of $T_I^{(i)}$ $(i \in \mathscr{J})$ are assumed known from previous epidemiological studies.

Suppose that the final outcome of the previous outbreak is observed in a sample of households. Label the m households in the population $1, \ldots, m$. For $i = 1, \ldots, m$, let $t_i = (t_{i1}, \ldots, t_{iJ})$, where t_{ij} is the number of class-j susceptibles ultimately infected in household i, let $n(i)$ be the category of household i and let δ_i be 1 or 0 according as household i is observed or unobserved. For $n \in \mathscr{N}$ and $0 \leqslant t \leqslant n$, let $p_n(t|\Lambda^L, \pi)$ be the probability that the epidemic $E_n(\Lambda^L, \pi)$, described in § 2·2, has final outcome t. For $n \in \mathscr{N}_0$, a triangular system of linear equations governing $p_n(t|\Lambda^L, \pi)$ $(0 \leqslant t \leqslant n)$ is given in the Appendix.

Let $t_D = \{t_i : \delta_i = 1\}$ denote the observed data. There does not exist a feasible method for computing the likelihood of (Λ^L, Λ^G) given t_D. However, suppose that the previous outbreak resulted in a global epidemic, so that $R_* > 1$. Then (2·1) and (2·2) implicitly determine π as a function of (Λ^L, Λ^G), so write $\pi = \pi(\Lambda^L, \Lambda^G)$. For large m, the argument of § 2·2 shows that the marginal probability of observing the outcome t_i in household i is $p_{n(i)}\{t_i|\Lambda^L, \pi(\Lambda^L, \Lambda^G)\}$. The outcomes t_i $(i = 1, \ldots, m)$ are not mutually independent. However, the central limit theorem of Ball & Lyne (2001) implies that their covariances are of order $1/m$ for large m, so, as in Lyne & Ball (1999), consider estimating (Λ^L, Λ^G) by maximising the pseudolikelihood

$$L(\Lambda^L, \Lambda^G|t_D) = \prod_{i=1}^{m} [p_{n(i)}\{t_i|\Lambda^L, \pi(\Lambda^L, \Lambda^G)\}]^{\delta_i}, \tag{3·1}$$

obtained by assuming that the outcomes in different households are independent.

The pseudolikelihood (3·1) can be maximised by first maximising it as a function of (Λ^L, π), to yield the estimator $(\hat{\Lambda}^L, \hat{\pi})$, then obtaining an estimator, $\hat{z}$ say, of z by substituting $(\hat{\Lambda}^L, \hat{\pi})$ in the right-hand side of (2·2), and finally solving (2·1), with (π, z) replaced by $(\hat{\pi}, \hat{z})$, for Λ^G. For single-type epidemics ($J = 1$), the resulting estimator of (Λ^L, Λ^G), which are both scalars, corresponds to that described in § 5.1 of Ball et al. (1997) although the pseudolikelihood interpretation was not present in that paper. However, for $J > 1$, the final step in the above procedure involves solving J linear equations in the J^2 unknown quantities λ^G_{ij} $(i, j \in \mathscr{J})$, so Λ^G is not identifiable from the observed data using this approach and the threshold parameters before and after vaccination, R_* and $R_*(v)$, cannot be estimated consistently. It is possible that the local infection rates Λ^L may also be unidentifiable, for example if for some $i, j \in \mathscr{J}$ there is no household in the sample that contains individuals of classes i and j, but this can be avoided by choosing the sample of households suitably. Note that if there is no household in the population that contains individuals of classes i and j then the parameters λ^L_{ij} and λ^L_{ji} are redundant.

3·2. *Estimation of R_*, $R_*^{\mathrm{AoN}}(v)$ and $R_*^{\mathrm{Le}}(v)$*

In § 3·1 final-size data, from a sample of households of one epidemic outbreak, were used to derive estimators of the matrix Λ^L and the vectors π and z. Estimation of the epidemiologically more important parameters R_* and $R_*(v)$ is considered now, under the assumption that the population structure is sufficiently rich and the sample is chosen suitably for Λ^L to be identifiable. The vaccination effect ε and the type, all-or-nothing or leaky, of the vaccine are assumed known, as are the distributions of $T_I^{(i)}$ $(i \in \mathscr{J})$. If the latter are unknown, then parameters of these distributions can be estimated if some parametric family is assumed, although note that with estimation from final-outcome data the scale of these distributions is confounded with the infection rates.

The method permits estimation of R_* and $R_*(v)$ for some future epidemic in a community with different household structure, provided it is considered reasonable to extrapolate parameter estimates from the sample to the future population. Let $\tilde{\alpha}_n$ denote the proportion of households in the future population that have category n, for $n \in \tilde{\mathscr{N}}$ with the obvious definition of $\tilde{\mathscr{N}}$, and, for $i \in \mathscr{J}$ and $n \in \tilde{\mathscr{N}}$, let $\tilde{\alpha}_i(n)$ be the probability that a class-i individual chosen at random in the future population resides in a household of category n. For the remainder of § 3 and for § 4, it is assumed that estimation of R_* and $R_*(v)$ is for a population with household structure given by $\tilde{\alpha}_n$ $(n \in \tilde{\mathscr{N}})$. Furthermore, when referring to formulae in § 2 for the mean matrices M and $M(v)$, we assume implicitly that $\alpha_i(n)$ has been replaced by $\tilde{\alpha}_i(n)$.

Note that $R_* = R_*(v_0)$, where v_0 denotes the null vaccination scheme given by $v_{n,r} = 1$ if $r = 0$ and $v_{n,r} = 0$ otherwise $(n \in \mathscr{N})$. Thus it is sufficient to consider estimation of $R_*(v)$, which, from (2·3) and (2·4), is given by the maximal eigenvalue of $M(v) = (m_{ij}(v))$, where

$$m_{ij}(v) = \sum_{l \in \mathscr{J}} b_{il}(v, \Lambda^L, \varepsilon)\mu_l \lambda^G_{lj},$$

with

$$b_{il}(v, \Lambda^L, \varepsilon) = \sum_{n \in \tilde{\mathscr{N}}} \tilde{\alpha}_i(n) \sum_{r=0}^{n} v_{n,r} \sum_{s=n-r}^{n} C(r, n-s)\varepsilon^{n-s}(1-\varepsilon)^{r-n+s} \frac{s_i}{n_i} \mu_{s,i,l}(\Lambda^L) \qquad (3{\cdot}2)$$

if the vaccine is all-or-nothing, and

$$b_{il}(v, \Lambda^L, \varepsilon) = \sum_{n \in \tilde{\mathscr{N}}} \tilde{\alpha}_i(n) \sum_{r=0}^{n} v_{n,r} \left\{ \frac{n_i - r_i}{n_i} \mu_{n-r,r,\mathrm{u}:i,l}(\Lambda^L, \varepsilon) + \frac{r_i(1-\varepsilon_i)}{n_i} \mu_{n-r,r,\mathrm{v}:i,l}(\Lambda^L, \varepsilon) \right\} \tag{3·3}$$

if the vaccine is leaky.

In the expression for $m_{ij}(v)$ the quantities $\tilde{\alpha}_i(n)$ and $\mu_l = E(T_I^{(l)})$ are known, and $\mu_{s,i,l}(\Lambda^L)$, $\mu_{n-r,r,\mathrm{u}:i,l}(\Lambda^L, \varepsilon)$ and $\mu_{n-r,r,\mathrm{v}:i,l}(\Lambda^L, \varepsilon)$ are estimated consistently by $\mu_{s,i,l}(\hat{\Lambda}^L)$, $\mu_{n-r,r,\mathrm{u}:i,l}(\hat{\Lambda}^L, \varepsilon)$ and $\mu_{n-r,r,\mathrm{v}:i,l}(\hat{\Lambda}^L, \varepsilon)$, respectively. However, for $J > 1$, the matrix Λ^G, and hence $R_*(v)$, cannot be estimated consistently. Nevertheless, Λ^G is known to satisfy the constraints given by (2·1), where π and z can be estimated consistently. Thus the Perron–Frobenius theorem is used to obtain bounds on $R_*(v)$, which are functions of (π, z) and thus can be estimated consistently. Similar methods were used for a multitype epidemic model without household structure by Britton (2001).

By the Perron–Frobenius theorem (Jagers, 1975, p. 92) it follows that there is a unique, up to normalisation, strictly positive vector $(x_1, \ldots, x_J)$ satisfying

$$R_*(v) x_j = \sum_{i \in \mathscr{J}} x_i m_{ij}(v) \quad (j \in \mathscr{J}).$$

Then

$$R_*(v) x_j = \sum_{i,k \in \mathscr{J}} x_i \mu_k b_{ik}(v, \Lambda^L, \varepsilon) \lambda^G_{kj} = \sum_{k \in \mathscr{J}} \frac{1}{\gamma_k z_k} \gamma_k z_k \mu_k \lambda^G_{kj} \sum_{i \in \mathscr{J}} x_i b_{ik}(v, \Lambda^L, \varepsilon).$$

Define the final sum by $r_k = \sum_{i \in \mathscr{J}} x_i b_{ik}(v, \Lambda^L, \varepsilon)$. Furthermore, let $A = \max_k \{r_k/(\gamma_k z_k)\}$ and assume that the maximum is attained for $k = k_0$. Then

$$R_*(v) x_j = \gamma_j \sum_{k \in \mathscr{J}} \frac{r_k}{\gamma_k z_k} \gamma_k z_k \mu_k \lambda^G_{kj}/\gamma_j \leqslant A \gamma_j \sum_{k \in \mathscr{J}} \gamma_k z_k \mu_k \lambda^G_{kj}/\gamma_j = A \gamma_j (-\log \pi_j) \quad (j \in \mathscr{J}),$$

where the final equality follows from (2·1). Hence, if we recall the definition of r_k,

$$\frac{R_*(v) r_k}{\gamma_k z_k} = \frac{1}{\gamma_k z_k} \sum_{i \in \mathscr{J}} R_*(v) x_i b_{ik}(v, \Lambda^L, \varepsilon) \leqslant A R^{(k)}_*(v),$$

where

$$R^{(k)}_*(v) = \frac{1}{\gamma_k z_k} \sum_{i \in \mathscr{J}} \gamma_i (-\log \pi_i) b_{ik}(v, \Lambda^L, \varepsilon) \quad (k \in \mathscr{J}). \tag{3·4}$$

Setting $k = k_0$ yields $R_*(v) \leqslant R^{(k_0)}_*(v) \leqslant \max_k R^{(k)}_*(v)$. Identical arguments yield a similar lower bound for $R_*(v)$, so that

$$\min_k R^{(k)}_*(v) \leqslant R_*(v) \leqslant \max_k R^{(k)}_*(v). \tag{3·5}$$

Just as in the multitype case without household structure treated in Britton (2001), these bounds are sharp, in that there exists Λ^G satisfying (2·1), such that the corresponding maximal eigenvalue equals the right-hand side of (3·5), and similarly for the lower bound. To see this, suppose the maximum on the right-hand side of (3·5) is obtained for $k = k_1$. For each j, define $\lambda^G_{k_1 j} = (-\log \pi_j)\gamma_j/(\gamma_{k_1} z_{k_1} \mu_{k_1})$ and $\lambda^G_{kj} = 0$ for all other k. First,

note that this choice for λ_{ij}^G satisfies (2·1). Secondly, observe that this choice for λ_{ij}^G is proportionate mixing and consequently it is easily verified that the maximal eigenvalue of the corresponding matrix M is the upper bound for $R_*(v)$. A similar argument shows that the lower bound can also be attained. Furthermore, linear interpolation between the two extreme choices for Λ^G shows that $R_*(v)$ can take any value in the interval given by (3·5).

3·3. *Estimation of the optimal vaccination scheme*

Let $R_*^{\max}(v) = \max_k R_*^{(k)}(v)$. Then, from (3·5), any vaccination scheme v with $R_*^{\max}(v) \leqslant 1$ is preventive, irrespective of the underlying parameter Λ^G consistent with the data, whilst for any vaccination scheme v with $R_*^{\max}(v) > 1$ there exists Λ^G, consistent with the data, so that $R_*(v) > 1$. Thus it is appropriate to consider minimisation of the vaccine coverage $S(v)$ subject to the constraints $R_*^{(k)}(v) \leqslant 1$ $(k \in \mathscr{J})$. Note that this is a linear programming problem since, by (2·5), (3·2), (3·3) and (3·4), the objective function $S(v)$ and the constraints $R_*^{(k)}(v) \leqslant 1$ $(k \in \mathscr{J})$ are all linear functions of the optimising variables v; a paper by the authors discussing the form of associated optimal vaccination schemes is currently under review. Let v^* denote a solution to this minimisation problem and let $c_v = S(v^*)$ be the corresponding vaccination coverage. Thus, if we assume optimal allocation of vaccines, c_v is the secure vaccination coverage required to be sure of preventing a future global outbreak. Note that $R_*^{(k)}(v)$ is estimated by replacing the unknown parameters in the right-hand side of (3·4) by their estimates, yielding $\hat{R}_*^{(k)}(v)$ say; (Λ^L, π) determines z by (2·2). Thus, $R_*^{\max}(v)$ is estimated by $\hat{R}_*^{\max}(v) = \max_k \hat{R}_*^{(k)}(v)$ and c_v is estimated by solving the above linear programming problem, with $R_*^{(k)}(v)$ replaced by $\hat{R}_*^{(k)}(v)$ $(k \in \mathscr{J})$, yielding $\hat{v}^*$ and $\hat{c}_v = S(\hat{v}^*)$.

4. Uncertainty

Let $\theta = (\mathrm{vec}\,(\Lambda^L), \pi) = (\theta_1, \ldots, \theta_{J(J+1)})$, where $\mathrm{vec}\,(\Lambda^L)$ is the row vector representation of Λ^L, and let $\hat{\theta}$ be the maximum pseudolikelihood estimator of θ. The central limit theorem of Ball & Lyne (2001), extended to the situation when not all households are observed, can be used to obtain, given the occurrence of a global epidemic, a central limit theorem for the pseudoscore function associated with (3·1), and weak laws of large numbers for the second derivatives of the corresponding pseudo-loglikelihood function and for locally uniform bounds on the third derivatives, as the population and sample sizes tend to infinity in an appropriate fashion. The argument of Cramér (1946, pp. 500–4), can then be used to show that $\hat{\theta}$ is a consistent estimator of θ and

$$m^{\frac{1}{2}}(\hat{\theta} - \theta) \to N\{0, \Sigma\,(\Lambda^L, \Lambda^G)\}, \tag{4·1}$$

in distribution, as $m \to \infty$.

Note that the variance matrix $\Sigma(\Lambda^L, \Lambda^G)$ depends on Λ^G rather than on π. The matrix $\Sigma(\Lambda^L, \Lambda^G)$ also depends on α_n, β_n $(n \in \mathscr{N})$, where, for $n \in \mathscr{N}$, β_n denotes the proportion of households of category n in the population that are in the observed sample. This latter dependence is suppressed for ease of notation. Calculation of $\Sigma(\Lambda^L, \Lambda^G)$ will be described in a future paper by Ball and Lyne and is not described here as it is lengthy and involved. Sufficient conditions for the limit (4·1) to hold will also be given in that paper. These include the important practical case when the proportions α_n, β_n $(n \in \mathscr{N})$ are held fixed as the number of households $m \to \infty$.

Recall that $R_*^{\max}(v) = \max_k R_*^{(k)}(v, \Lambda^L, \pi)$, where $R_*^{(k)}(v, \Lambda^L, \pi)$ $(k \in \mathscr{J})$ are given by (3·4), and their dependence on the unknown parameters Λ^L and π is shown explicitly. Let k_1 denote the k which maximises $R_*^{(k)}(v, \Lambda^L, \pi)$ $(k \in \mathscr{J})$, and suppose, for ease of exposition, that k_1 is unique. Then $R_*^{\max}(v)$ is estimated consistently by $\hat{R}_*^{\max}(v) = R_*^{(\hat{k}_1)}(v, \hat{\Lambda}^L, \hat{\pi})$, where $\hat{k}_1$ maximises $\hat{R}_*^{(k)}(v)$ $(k \in \mathscr{J})$. Furthermore, let $d(v, \Lambda^L, \pi)$ be the $J(J+1)$-dimensional column vector whose ith element is the partial derivative of $R_*^{(k_1)}(v, \Lambda^L, \pi)$ with respect to θ_i. Then an application of the delta method (Andersen et al., 1993, pp. 109–10) shows that

$$m^{\frac{1}{2}}\{\hat{R}_*^{\max}(v) - R_*^{\max}(v)\} \to N\{0, \sigma^2(v, \Lambda^L, \Lambda^G)\}, \tag{4·2}$$

in distribution, as $m \to \infty$, where, with T denoting transpose,

$$\sigma^2(v, \Lambda^L, \Lambda^G) = d(v, \Lambda^L, \pi)^{\mathrm{T}} \Sigma(\Lambda^L, \Lambda^G) d(v, \Lambda^L, \pi).$$

In order to use (4·2) to obtain a confidence interval for $R_*^{\max}(v)$, an estimator of Λ^G is required. Now $R_*(v)$ is maximised when class-k_1 individuals are responsible for all global infections, so Λ^G is estimated by setting $\hat{\lambda}_{ij}^G = 0$ if $i \neq k_1$ and

$$\hat{\lambda}_{k_1 j}^G = (-\log \hat{\pi}_j)\gamma_j / (\gamma_{k_1} \hat{z}_{k_1} \mu_{k_1}) \quad (j \in \mathscr{J}), \tag{4·3}$$

where $\hat{z}_{k_1}$ is obtained by setting $(\Lambda^L, \pi) = (\hat{\Lambda}^L, \hat{\pi})$ in (2·2). A one-sided $1 - \alpha$ confidence interval for $R_*^{\max}(v)$ is then given by $(0, \hat{R}_*^{\max}(v) + m^{-\frac{1}{2}} z_\alpha \sigma(v, \hat{\Lambda}^L, \hat{\Lambda}^G))$, where z_α is the $(1 - \alpha)$-quantile of the standard normal distribution. The asymptotic variance $\sigma^2(v, \Lambda^L, \Lambda^G)$ may be larger for other choices of Λ^G consistent with (Λ^L, π), but such choices will have $R_*(v) < R_*^{\max}(v)$. Thus the above confidence interval is asymptotically conservative.

We turn now to estimation of the secure vaccination coverage c_v under the optimal vaccination strategy, outlined in § 3·3. For $c \in (0, 1)$, let $v^{\mathrm{opt}}(c, \theta)$ denote an optimal vaccination scheme, given that a proportion c of the population are to be vaccinated, where dependence on the unknown parameters $\theta = (\mathrm{vec}(\Lambda^L), \pi)$ is shown explicitly; that is $v^{\mathrm{opt}}(c, \theta)$ minimises $R_*^{\max}(v)$ subject to $S(v) \leqslant c$. Let $R(c, \theta) = R_*^{\max}\{v^{\mathrm{opt}}(c, \theta)\}$. Then, for fixed θ, $R(c, \theta)$ is a continuous, decreasing, piecewise linear function of c; moreover, it is strictly decreasing for $c \in [0, c_q(\theta)]$, where $c_q(\theta) = \min\{c : R(c, \theta) = R(1, \theta)\}$. Note that $c_q(\theta)$ may be strictly less than 1, if, for example, some classes of individuals are insensitive to the vaccine. Thus, in the practically relevant situation when $R(0, \theta) > 1 > R(1, \theta)$, the secure vaccination coverage $c_v = c_v(\theta)$ is obtained by solving $R(c, \theta) = 1$. In practice, θ is unknown and c_v is estimated by $\hat{c}_v = c_v(\hat{\theta})$.

Let $d_c(\Lambda^L, \pi)$ be the $J(J+1)$-dimensional column vector whose ith element is $\partial c_v(\theta)/\partial \theta_i$ and let

$$\sigma_c^2(\Lambda^L, \Lambda^G) = d_c(\Lambda^L, \pi)^{\mathrm{T}} \Sigma(\Lambda^L, \Lambda^G) d_c(\Lambda^L, \pi).$$

Then, by the delta method,

$$m^{\frac{1}{2}}(\hat{c}_v - c_v) \to N\{0, \sigma_c^2(\Lambda^L, \Lambda^G)\},$$

in distribution, as $m \to \infty$.

A one-sided $1 - \alpha$ confidence interval for c_v is then given by $(0, \hat{c}_v + m^{-\frac{1}{2}} z_\alpha \sigma_c(\hat{\Lambda}^L, \hat{\Lambda}^G))$, with $\hat{\Lambda}^G$ being given by (4·3). The lack of an explicit expression for $R(c, \theta)$ means that, unless $J = 1$ (Britton & Becker, 2000), the derivatives $\partial c_v(\theta)/\partial \theta_i$ need to be evaluated numerically, which is straightforward since $c_v(\theta)$ arises from the solution of a linear programming problem.

The above confidence interval for $R_*^{\max}(v)$ assumes that the vaccination scheme v is fixed, whereas, in practice, for a given coverage c, a confidence interval may be required for $R(c, \theta)$, the post-vaccination threshold parameter upper bound, assuming that the vaccines are allocated optimally. To obtain such an interval using the delta method, the partial derivatives $\partial R(c, \theta)/\partial \theta_i$ $(i = 1, \ldots, J(J+1))$ are required. It is difficult to obtain these directly, even numerically, since the optimisation problem underlying $R(c, \theta)$ is a linear programming problem only when $J = 1$. When $J > 1$, the objective function $R_*^{\max}(v) = \max_k R_*^{(k)}(v)$ is nonlinear in v. For fixed $r \in (0, \theta)$, let $c_v^{(r)} = c_v^{(r)}(\theta)$ satisfy $R(c_v^{(r)}, \theta) = r$. Then, by the implicit function theorem,

$$\frac{\partial c_v^{(r)}(\theta)}{\partial \theta_i} = -\frac{\partial R(c_v^{(r)}, \theta)}{\partial \theta_i} \Bigg/ \frac{\partial R(c_v^{(r)}, \theta)}{\partial c} \quad (i = 1, \ldots, J(J+1)). \tag{4·4}$$

Now $c_v^{(r)}$ arises from the solution of the linear programming problem 'minimise $S(v)$ subject to $R_*^{(k)} \leqslant r$ $(k \in \mathscr{J})$', enabling $\partial c_v^{(r)}(\theta)/\partial \theta_i$ and $\partial c_v^{(r)}(\theta)/\partial r$ to be calculated numerically. Also, $\partial R(c_v^{(r)}, \theta)/\partial c = 1/\{\partial c_v^{(r)}(\theta)/\partial r\}$, so $\partial R(c, \theta)/\partial \theta_i$ can be found by letting $r = R(c, \theta)$ and using (4·4).

It has been assumed in the above that k_1, which maximises $R_*^{(k)}(v, \Lambda^L, \pi)$ $(k \in \mathscr{J})$, is unique. If that is not the case then $\hat{R}_*^{\max}(v)$ is still a consistent estimator of $R_*^{\max}(v)$ but the above confidence interval for $R_*^{\max}(v)$ may no longer have the required asymptotic coverage probability. For $k \in \mathscr{J}$, let $\hat{\Lambda}_k^G$ denote the estimator of Λ^G obtained by assuming that class-k individuals are responsible for all global infections, and let

$$\hat{R}_*^U(v, \alpha) = \max_k \{R_*^{(k)}(v, \hat{\Lambda}^L, \hat{\pi}) + m^{-\frac{1}{2}} z_\alpha \sigma(v, \hat{\Lambda}^L, \hat{\Lambda}_k^G)\}.$$

Then $(0, \hat{R}_*^U(v, \alpha))$ is a $1 - \alpha$ confidence interval for $R_*^{\max}(v)$ that is asymptotically conservative. Moreover, when k_1 is unique, the probability that this confidence interval coincides with the earlier one tends to 1 as $m \to \infty$. Thus, it is recommended that the interval $(0, \hat{R}_*^U(v, \alpha))$ be used in practice. A similar comment applies to the confidence intervals for c_v and $R(c, \theta)$. Indeed, the numerical examples in § 5 indicate that in these latter two cases k_1 is usually not unique.

5. Numerical examples

The techniques developed in this paper are illustrated by application to data on influenza epidemics in Tecumseh, Michigan (Monto et al., 1985), kindly made available by Ira M. Longini. These data are from a continuous epidemiological survey from 1976 to 1981, representing a 10% cross-sectional sample of households that were followed prospectively. There were two main epidemics, in 1977–78 and 1980–81, infecting 130 and 128 out of the 685 and 795 individuals monitored, respectively. The data from the 1977–78 outbreak are considered here, since by the 1980–81 outbreak additional recruitment of families with infants into the survey meant that the observed sample was not representative of the underlying population structure. Individuals in the survey underwent a haemagglutination test before and after each epidemic season. The pre-season results were used to classify individuals into those possessing low (highly susceptible), higher (less susceptible) and very high (immune) levels of antibodies, and the post-season results were used to determine whether a susceptible individual had been infected. Several other covariates

were also recorded, including age, so individuals can be classified as adults ($\geqslant 18$ years) or children (<18 years). The data are too numerous to present in detail. For the 1977–78 epidemic, there were 289 households in the survey, of which 77 were of size 1, 106 of size 2, 47 of size 3, 44 of size 4, 12 of size 5, 2 of size 6 and 1 of size 7, where the size of a household is the number of susceptibles in it at the start of the epidemic season, counting both low- and high-titre individuals, but neglecting immunes. These households contained a total of 685 individuals, 308 low-titre adults, 136 low-titre children, 184 high-titre adults and 57 high-titre children, of whom 48, 56, 17 and 9 were infected, respectively. In the following examples, the observed households are assumed to form an exact 10% sample from the population and, following Addy et al. (1991), the infectious period of all individuals is assumed to follow a gamma distribution with mean 4·1 days and shape parameter 2.

Consider first the case when age is ignored, so there are two classes of susceptible individuals and the category of a household is determined by the number of low-titre (class 1) and high-titre (class 2) individuals it contains. The following estimates are obtained:

$$\hat{\Lambda}^L = \begin{pmatrix} 0{\cdot}0536 & 0{\cdot}0291 \\ 0{\cdot}0000 & 0{\cdot}0052 \end{pmatrix}, \quad \hat{\pi} = (0{\cdot}8172 \;\; 0{\cdot}9196), \quad \hat{z} = (0{\cdot}2339 \;\; 0{\cdot}1015).$$

Define $R_*^{(k)}(\hat{\Lambda}^L, \hat{\pi}) = R_*^{(k)}(v_0, \hat{\Lambda}^L, \hat{\pi})$, where v_0 is the null vaccination scheme defined in § 3·2. Then $R_*^{(1)}(\hat{\Lambda}^L, \hat{\pi}) = 1{\cdot}0868$ and $R_*^{(2)}(\hat{\Lambda}^L, \hat{\pi}) = 1{\cdot}1801$, so $\hat{R}_*^{\max}(v_0) = 1{\cdot}1801$ and $\hat{R}_*^U(v_0, 0{\cdot}05) = 1{\cdot}2832$, leading to a 95% confidence interval of $(0, 1{\cdot}2832)$ for $R_*^{\max}(v_0)$. The current killed influenza vaccine has an efficacy of about 0·7, irrespective of prior immunity, according to a personal communication from Ira M. Longini. If we assume an all-or-nothing vaccine with $\varepsilon = (0{\cdot}7 \;\; 0{\cdot}7)$, the secure vaccination coverage c_v is estimated to be 0·0877 with a 95% confidence interval of $(0, 0{\cdot}2424)$. For the leaky case, $\hat{c}_v = 0{\cdot}0897$ with 95% confidence interval $(0, 0{\cdot}2441)$, so slightly more vaccine is required than in the all-or-nothing case. The leaky case is illustrated in Fig. 1(a), in which $R_*^{(k)}(v_c, \hat{\Lambda}^L, \hat{\pi})$ $(k = 1, 2)$ and $\hat{R}_*^U(v_c, 0{\cdot}05)$ are plotted against vaccine coverage c, where $v_c = v^{\mathrm{opt}}(c, \hat{\theta})$; see § 4. Figure 1(a) also shows the estimate of the secure vaccination coverage $\hat{c}_v$ and the upper limit of its associated 95% confidence interval. The figures are drawn by solving the linear programming problem 'minimise $S(v)$ subject to $R_*^{(k)}(v, \hat{\Lambda}^L, \hat{\pi}) \leqslant r$ $(k \in \mathscr{J})$', having solution $v^*(r)$ say, for a grid of values of r, and then plotting $R_*^{(k)}\{v^*(r), \hat{\Lambda}^L, \hat{\pi}\}$ $(k \in \mathscr{J})$ and $\hat{R}_*^U\{v^*(r), 0{\cdot}05\}$ against $S\{v^*(r)\}$, using linear interpolation. Note that, since $R(c, \hat{\theta})$ is continuous and decreasing in c, $R[S\{v^*(r)\}, \hat{\theta}] = r$. Thus, if $S\{v^*(r)\} = c$ then $R(c, \hat{\theta}) = r$, implying that $R_*^{\max}(v_c, \hat{\theta}) = r$, so v_c may be taken to be $v^*(r)$ as $R_*^{\max}\{v^*(r), \hat{\theta}\} = r$. The confidence intervals indicated by the dashed line in Fig. 1(a) are calculated under the assumption that the vaccination scheme v_c is known. For comparison, the confidence interval for $R(c, \theta)$ when $c = \hat{c}_v$ is also shown, it being numerically prohibitive to calculate the latter confidence interval for all values of c. Note that, except for c close to zero or one, the optimal strategy results in $R_*^{(1)}(v_c, \hat{\Lambda}^L, \hat{\pi}) = R_*^{(2)}(v_c, \hat{\Lambda}^L, \hat{\pi})$. The vaccination scheme that results in $R_*^{(1)}(v_c, \hat{\Lambda}^L, \hat{\pi}) = R_*^{(2)}(v_c, \hat{\Lambda}^L, \hat{\pi}) = 1$ concentrates most vaccination on low-titre individuals residing in large households.

It is of interest to consider the vaccination problem if all of the population were in fact low antibody level, since if a vaccination scheme is now implemented there will be no significant immunity due to disease, using the estimates for $\hat{\Lambda}^L$, $\hat{\pi}$ and $\hat{z}$ obtained above. The bounds $R_*^{(k)}(\hat{\Lambda}^L, \hat{\pi})$ can then be calculated as 1·2810 and 0 for $k = 1, 2$, respectively,

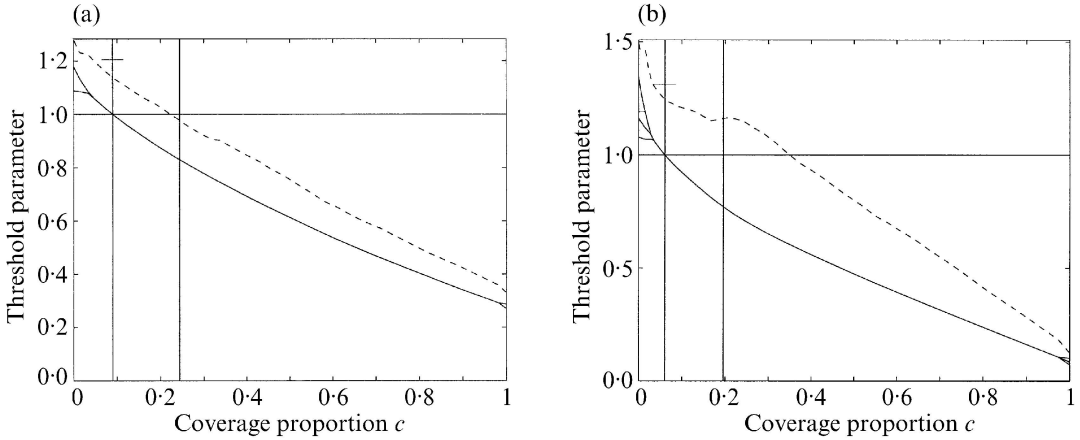

Fig. 1. (a) Reduction of threshold parameter, as a function of vaccination coverage c, for two-type model with a leaky vaccine of efficacy $\varepsilon = (0{\cdot}7\ 0{\cdot}7)$. (b) Reduction of threshold parameter through optimal vaccination for four-type model, with an all-or-nothing vaccine of efficacy $\varepsilon = (0{\cdot}9, 0{\cdot}9, 0{\cdot}9, 0{\cdot}9)$. The solid lines, which coincide for most values of c, are $R_*^{(k)}(v_c, \hat{\Lambda}^L, \hat{\pi})$, where $k = 1, 2$ in (a) and $k = 1, 2, 3, 4$ in (b), and the dashed line is $\hat{R}_*^U(v_c, 0{\cdot}05)$, where v_c denotes the best vaccination strategy for coverage level c; see text for further details. The left and right vertical lines mark $\hat{c}_v$ and the upper limit of its associated 95% confidence interval, respectively. The horizontal bar marks $\hat{R}_*^U(\hat{v}^{\text{opt}}, 0{\cdot}05)$ for the estimated optimal vaccination scheme associated with this coverage, i.e. the upper limit of the 95% confidence interval for $R(c, \hat{\theta})$ when $c = \hat{c}_v$; recall that $R(\hat{c}_v, \hat{\theta}) = 1$.

and the upper limit of the 95% confidence interval for $R_*^{\max}(v_0)$ is $\hat{R}_*^U(v_0, 0{\cdot}05) = 1{\cdot}4430$. If we assume an all-or-nothing [leaky] vaccine with $\varepsilon_1 = 0{\cdot}7$, the secure vaccination coverage c_v is estimated to be 0·1631 [0·1673] with a 95% confidence interval of (0, 0·2491) [(0, 0·2544)]. Note that the assumption that the population is all low-titre has increased appreciably the estimate of c_v for both kinds of vaccine. However, the upper limits of the associated confidence intervals have increased only marginally, which at first sight seems surprising. Note, as can be seen in Fig. 1(a) for the leaky case, that, when $c = \hat{c}_v$, $R_*^{(1)}(v_c, \hat{\Lambda}^L, \hat{\pi}) = R_*^{(2)}(v_c, \hat{\Lambda}^L, \hat{\pi})$, so that, if we recall the discussion at the end of § 4, the upper end of the confidence interval for c_v is given by $\hat{c}_v + m^{-\frac{1}{2}} z_{0{\cdot}05} \max_k \sigma_c(\hat{\Lambda}^L, \hat{\Lambda}_k^G)$. Now $\sigma_c(\hat{\Lambda}^L, \hat{\Lambda}_1^G) < \sigma_c(\hat{\Lambda}^L, \hat{\Lambda}_2^G)$, as is intuitively plausible since far fewer high-titre individuals than low-titre individuals were infected by the epidemic, so $k = 2$ is used when calculating the confidence interval for c_v, under the assumption that the population structure remains constant, and $m^{-\frac{1}{2}} \sigma_c(\hat{\Lambda}^L, \hat{\Lambda}_2^G) = 0{\cdot}0941$ for the all-or-nothing vaccine, and 0·0939 for the leaky. However, when the population is assumed to be all low-titre, only $k = 1$ is relevant and, although $m^{-\frac{1}{2}} \sigma_c(\hat{\Lambda}^L, \hat{\Lambda}_1^G)$ has increased from 0·0295 to 0·0522 for the all-or-nothing vaccine, or from 0·0294 to 0·0529 for the leaky vaccine, the upper limit of the confidence interval has increased only slightly.

Consider now the case when individuals are also classified as adults or children. This gives four classes of individuals and estimates of $\hat{\Lambda}^L$, $\hat{\pi}$ and $\hat{z}$ can be obtained. Figure 1(b) shows the reduction of threshold parameter with an all-or-nothing vaccine with $\varepsilon = (0{\cdot}9, 0{\cdot}9, 0{\cdot}9, 0{\cdot}9)$, chosen as it illustrates the following points more clearly than $\varepsilon = (0{\cdot}7, 0{\cdot}7, 0{\cdot}7, 0{\cdot}7)$. Note that the upper limit of the 95% confidence interval for c_v is 0·1958, which is somewhat smaller than the coverage required for $\hat{R}_*^U(v_c, 0{\cdot}05)$ to be below 1, namely 0·3495. Also note that $\hat{R}_*^U(v_c)$ is not monotonic.

If all the individuals in a future population were in fact low-titre and the vaccine was all-or-nothing with $\varepsilon = (0{\cdot}7, 0{\cdot}7, 0{\cdot}7, 0{\cdot}7)$, the required coverage is 0·2567, with 95%

confidence interval $(0, 0{\cdot}3356)$, and is achieved entirely through vaccinating children. A leaky vaccine of the same efficacy would require coverage 0·2881, with 95% confidence interval $(0, 0{\cdot}3995)$. Note that the four-type model used here to predict on low-titre individuals estimates that rather more vaccine is required than in the two-type models above.

6. Discussion

The data needed to perform the analysis of this paper require information at the household level, where individuals are also categorised into different types according to knowledge of some individual covariates, such as age, sex and previous history of disease and/or vaccination. In large outbreaks such information is rarely available for the whole community as it more or less requires visits to each household separately. Still, such information can, and it is recommended should, be collected for a sample of households, and this is all that is required for the present analysis. In case the sample is not representative in terms of the household structure of the population, information about the community distribution of various household categories is also needed, but this can often be obtained from census data. In order to derive preventive vaccination schemes the type and efficacy of the vaccine must also be known. Methods for estimating efficacy of vaccines is a topic in its own right; see for example the review by Halloran et al. (1999).

The model used in the paper allows for heterogeneities associated with observable, and hence classifiable, individual characteristics and also for departures from homogeneous mixing caused by the presence of households. Of course, there are other heterogeneities present in any community. For example, individuals may differ in a way which cannot be known by epidemiologists collecting the data. Furthermore, there are other social structures which surely affect the spread of disease, such as schools and workplaces, which clearly act as clusters. Nevertheless, it is believed that households, in combination with having different types of individuals, capture the most important departures from homogeneity. Needless to say, it is impossible to capture all heterogeneities in a community in a mathematical model.

As noted already in § 1, consistent estimation of threshold parameters and associated optimal vaccination schemes is not feasible because the global infectivity rates are unidentifiable from final outcome data. In applications some knowledge of these parameters may be available, expressed either in deterministic terms or in the form of prior distributions. Such prior knowledge should narrow the lower and upper bounds of the estimates, thus giving less conservative estimates. In the Bayesian framework, the complexity of the model suggests that such inferences will most likely be performed using Markov chain Monte Carlo methods; see O'Neill et al. (2000) for an application of these methods in a simpler epidemic setting.

Finally, although linear programming provides a means of computing optimal vaccination allocations, it is useful to have an explicit characterisation of the resulting solution and thereby gain insight into the form of optimal vaccination schemes. In the single class case ($J = 1$) with all-or-nothing vaccines, Ball & Lyne (2002) show that, provided a certain convexity conjecture holds, successive vaccinations within the same household yield diminishing reductions in the threshold parameter R_*, leading to simple characterisations for the form of optimal vaccination allocations. In particular, if the vaccine is perfect, the optimal vaccination scheme is the so-called equalising strategy of Ball et al. (1997), in which vaccines are allocated sequentially, always to a household

that contains the greatest number of unvaccinated individuals. The form of the optimal vaccination scheme in the multitype case is investigated in a paper by the authors that is currently under review, where it is shown that the multitype case does not admit such a simple characterisation and that the leaky vaccine leads to less reduction in the spread of disease than the corresponding all-or-nothing vaccine.

Acknowledgement

This research was supported by the UK Engineering and Physical Sciences Research Council, the Swedish Research Council and the Royal Swedish Academy of Sciences.

Appendix

Properties of single household epidemics

We present algorithms for calculating $\mu_{n,i,j}(\Lambda^L)$, $\mu_{n,i}(\Lambda^L, \pi)$ and $p_n(t|\Lambda^L, \pi)$. For $n=(n_1, \ldots, n_J)$ and $a=(a_1, \ldots, a_J)$, let $E_{n,a}(\Lambda^L, \pi)$ denote the multitype single-household epidemic model studied by Addy et al. (1991), in which initially there are a_i infectives and n_i susceptibles of class i ($i \in \mathscr{J}$), and, during the course of the epidemic, initially susceptible individuals avoid infection from outside the household independently and with probability π_i for a class-i individual. The infectious periods of different infectives are independent, with that of a class-i infective following a random variable $T_I^{(i)}$, having an arbitrary but specified distribution with moment generating function $\phi_i(\theta)=E\{\exp(-\theta T_I^{(i)})\}$ ($\theta \geqslant 0$). Throughout its infectious period, a given class-i infective contacts a given class-j susceptible at the points of a homogenous Poisson process with rate λ_{ij}^L. For $i \in \mathscr{J}$, let $\tilde{S}_i$ denote the number of initial class-i susceptibles that are uninfected at the end of the epidemic and let $\mu_{n,a,i}(\Lambda^L, \pi)=E(\tilde{S}_i)$. A recursive expression for $\mu_{n,a,i}(\Lambda^L, \pi)$ is presented, from which expressions for $\mu_{n,i,j}(\Lambda^L)$ and $\mu_{n,i}(\Lambda^L, \pi)$ are easily obtained. To be specific, if, for $i \in \mathscr{J}$, $a^{(i)}$ denotes the J-dimensional vector whose ith element is one and all of whose other elements are zero, then

$$\mu_{n,i,j}(\Lambda^L)=n_j-\mu_{n-a^{(i)},a^{(i)},j}(\Lambda^L, 1) \quad (i, j \in \mathscr{J}), \quad \mu_{n,i}(\Lambda^L, \pi)=n_i-\mu_{n,0,i}(\Lambda^L, \pi) \quad (i \in \mathscr{J}).$$

An expression for the joint probability generating function of $S=(S_1, \ldots, S_J)$ for the epidemic $E_{n,a}(\Lambda^L, 1)$ is given, in different notation, by Theorem 3.5 of Ball (1986). Appropriate differentiation of that expression shows that, for $i \in \mathscr{J}$,

$$\mu_{n,a,i}(\Lambda^L, 1)=\sum_{s=0}^{n} C(n, s)\alpha_s^{(i)}\phi\{h(s)\}^{a+n-s}, \tag{A·1}$$

where $\alpha_s^{(i)}$ ($s \geqslant 0$) are determined by

$$\sum_{s=0}^{n} C(n, s)\alpha_s^{(i)}\phi\{h(s)\}^{n-s}=n_i \quad (n \geqslant 0).$$

Here, for $r \in \mathscr{N}_0 \cup \{0\}$, $h(r)=(h_1(r), \ldots, h_J(r))$, where $h_j(r)=\sum_{k \in \mathscr{J}} r_k \lambda_{jk}^L$, and

$$\phi(\theta)=(\phi_1(\theta_1), \ldots, \phi_J(\theta_J)) \ (\theta \in \mathbb{R}^J).$$

The distribution of the ultimate spread of the epidemic $E_{n,a}(\Lambda^L, \pi)$ can be obtained by conditioning on the numbers of initial susceptibles of the J classes that avoid infection from outside the household, $Y=(Y_1, \ldots, Y_J)$ say, and considering the epidemic $E_{Y,n+a-Y}(\Lambda^L, 1)$ in which there is no outside infection. Hence, for $i \in \mathscr{J}$,

$$\mu_{n,a,i}(\Lambda^L, \pi)=\sum_{r=0}^{n} C(n, r)\pi^r(1-\pi)^{n-r}\mu_{r,n+a-r,i}(\Lambda^L, 1). \tag{A·2}$$

Substituting (A·1) into (A·2) and reversing the order of summation yields

$$\mu_{n,a,i}(\Lambda^L, \pi)=\sum_{s=0}^{n} C(n, s)\alpha_s^{(i)}\phi\{h(s)\}^{a+n-s}\pi^s \quad (i \in \mathscr{J}).$$

Recall from § 3·1 that $p_n(t|\Lambda^L, \pi)$ $(0 \leqslant t \leqslant n)$ is the total size distribution of the epidemic $E_{n,0}(\Lambda^L, \pi)$. It follows, using equation (4) of Addy et al. (1991), that

$$\sum_{t=0}^{s} C(n-t, s-t)p_n(t|\Lambda^L, \pi)/([\phi\{h(n-s)\}]^t \pi^{n-s}) = C(n, s) \quad (0 \leqslant s \leqslant n). \tag{A·3}$$

The triangular system of linear equations (A·3) determines $p_n(t|\Lambda^L, \pi)$ $(0 \leqslant t \leqslant n)$.

References

Addy, C. L., Longini, I. M. & Haber, M. (1991). A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics* **47**, 961–74.

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Anderson, R. M. & May, R. M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Andersson, H. (1999). Epidemic models and social networks. *Math. Scientist* **24**, 128–47.

Andersson, H. & Britton, T. (2000). *Stochastic Epidemic Models and their Statistical Analysis*, Lecture Notes in Statistics 151. New York: Springer.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*, 2nd ed. London: Griffin.

Ball, F. G. (1986). A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Adv. Appl. Prob.* **18**, 289–310.

Ball, F. G. & Lyne, O. D. (2001). Stochastic multitype SIR epidemics among a population partitioned into households. *Adv. Appl. Prob.* **33**, 99–123.

Ball, F. G. & Lyne, O. D. (2002). Optimal vaccination policies for stochastic epidemics among a population of households. *Math. Biosci.* **177–8**, 333–54.

Ball, F. G., Mollison, D. & Scalia-Tomba, G. (1997). Epidemic with two levels of mixing. *Ann. Appl. Prob.* **7**, 46–89.

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

Becker, N. G. & Britton, T. (1999). Statistical studies of infectious disease incidence. *J. R. Statist. Soc.* B **61**, 287–307.

Becker, N. G. & Dietz, K. (1995). The effect of household distribution on transmission and control of highly infectious diseases. *Math. Biosci.* **127**, 207–19.

Becker, N. G. & Hall, R. (1996). Immunization levels for preventing epidemics in a community of households made up of individuals of various types. *Math. Biosci.* **132**, 205–16.

Becker, N. G. & Marschner, I. C. (1990). The effect of heterogeneity on the spread of disease. In *Stochastic Processes in Epidemic Theory*, Lecture Notes in Biomathematics **86**, Ed. J.-P. Gabriel, C. Lefèvre and P. Picard, pp. 90–103. Berlin: Springer-Verlag.

Becker, N. G. & Starczak, D. N. (1997). Optimal vaccination strategies for a community of households. *Math. Biosci.* **139**, 117–32.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des advantages de l'inoculation pour la prévenir. *Mém. Math. Phys. Acad. R. Sci. Paris* 1–45.

Britton, T. (2001). Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage. *J. R. Statist. Soc.* B **63**, 705–15.

Britton, T. & Becker, N. G. (2000). Estimating the immunity coverage required to prevent epidemics in a community of households. *Biostatistics* **1**, 389–402.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Greenhalgh, D. & Dietz, K. (1994). Some bounds on estimates for reproductive ratios derived from the age-specific force of infection. *Math. Biosci.* **124**, 9–57.

Halloran, M. E., Haber, M. & Longini, I. M. (1992). Interpretation and estimation of vaccine efficacy under heterogeneity. *Am. J. Epidemiol.* **136**, 328–43.

Halloran, M. E., Longini, I. M. & Struchiner, C. J. (1999). Design and interpretation of vaccine field studies. *Epidemiol. Rev.* **21**, 73–88.

Heesterbeek, J. A. P. & Dietz, K. (1996). The concept of R_0 in epidemic theory. *Statist. Neer.* **50**, 89–110.

Hethcote, H. W. & Van Ark, J. W. (1987). Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. *Math. Biosci.* **84**, 85–118.

Jagers, P. (1975). *Branching Processes with Biological Applications*. Wiley: London.

Kermack, W. O. & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond.* A **115**, 700–21.

Lyne, O. D. & Ball, F. G. (1999). Parameter estimation for SIR epidemics in households. Bull. Int. Statist Inst. 52nd Session, Contributed Papers, Tome LVIII, Book 2, pp. 251–2. Int. Statist Inst.

Mode, C. J. (1971). *Multitype Branching Processes — Theory and Applications*. Elsevier: New York.

Monto, A. S., Koopman, J. S. & Longini, I. M. (1985). Tecumseh study of illness, XIII, Influenza infection and disease, 1976–1981. *Am. J. Epidemiol.* **121**, 811–22.

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo Methods. *Appl. Statist.* **49**, 517–42.

Smith, P. G., Rodrigues, L. C. & Fine, P. E. M. (1984). Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *Int. J. Epidemiol.* **13**, 87–93.

[*Received October* 2002. *Revised July* 2003]