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

Frank Ball, University of Nottingham*

Tom Britton, Uppsala University†

Owen Lyne, University of Nottingham‡

September 20, 2002

Abstract

This paper is concerned with a stochastic model for the spread of an SIR (susceptible $\rightarrow$ infective $\rightarrow$ removed) epidemic among a closed, finite population that contains several types of individuals and is partitioned into households. Previously obtained probabilistic and inferential results for the model are used to estimate the threshold parameter R_*, which determines whether or not a major outbreak can occur, both before and after vaccination. 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

*School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, England. *E-mail*: fgb@maths.nott.ac.uk

†Department of Mathematics, Uppsala University, P.O. Box 480, SE-751 06 Uppsala, Sweden. *E-mail*: tom.britton@math.uu.se

‡School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, England. *E-mail*: Owen.Lyne@maths.nottingham.ac.uk

value of one 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: stochastic epidemic, household epidemic, multitype epidemic, threshold parameter, vaccination, estimation, outbreak data.

1 Introduction

Epidemic models have a long history going back at least to Bernoulli [19], 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 of Kermack and McKendrick [28], and, in modern terminology, states that a major epidemic can occur only if the basic reproduction number R_0 (see, for example, Heesterbeek and Dietz [25]) 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 [6], Anderson and May [3] and Andersson [4], to mention just a few. In order to determine the critical vaccination coverage in practice, estimates of model parameters are required. Thus, alongside modelling, 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 and May [3], Becker [13], Becker and Britton [14] and Andersson and Britton [5].

One departure from homogeneous mixing, that has received considerable interest recently and has an important impact on model behaviour, is that owing to the household structure of most human populations (see, for example, Becker and Dietz [15] and Ball et al. [12]). 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-household infections. However, it is well known that heterogeneities, such as those owing to age, sex and response to vaccine, can have a significant effect on disease spread. Ball and Lyne [9] studied the

probabilistic behaviour of a stochastic multitype SIR (susceptible $\rightarrow$ infective $\rightarrow$ removed) model for the spread of an epidemic among a closed community in which individuals reside in households and, in particular, derived a threshold parameter R_* (the households model equivalent of the basic reproduction number R_0) that determines whether or not a major outbreak can occur; see also Becker and Hall [16]. Statistical inference for this model, from final outcome data (possibly only for a sample of households in the community), is considered by Ball and Lyne [11]. However, it turns out that the between-household 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 and Dietz [22] and Britton [20] for multitype epidemics without household structure.

In the present paper, estimation of the above epidemiologically important parameters is studied for the first time 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. [35] and, following Halloran et al. [23], 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 the secure vaccination coverage, that reduces the upper bound for R_* to one, to be estimated. Further, 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 for these epidemiologically important parameters. 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 non-linear optimisation problem has to be solved, unless between-household

infection is proportionate mixing (see Section 2.3.1).

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

2 Model, threshold behaviour and vaccination

2.1 Model

The model under consideration in this paper is that of Ball and Lyne [9] for the spread of an SIR (susceptible $\rightarrow$ infective $\rightarrow$ removed) epidemic among a closed, finite population that contains J classes of individuals, labelled $1, 2, \ldots, J$, and is partitioned into households. Let $\mathcal{J} = \{1, 2, \ldots, J\}$ and $\mathcal{N}_0 = \{\boldsymbol{n} = (n_1, n_2, \ldots, n_J) \in \mathbb{Z}^J : n_j \geq 0 \ (j \in \mathcal{J}), \ |\boldsymbol{n}| = \sum_{j=1}^{J} n_j \geq 1\}$. Suppose that, for $\boldsymbol{n} \in \mathcal{N}_0$, the population contains $m_{\boldsymbol{n}}$ households of category $\boldsymbol{n}$, where a household of category $\boldsymbol{n}$ contains n_j individuals of class j $(j \in \mathcal{J})$. Let $m = \sum_{\boldsymbol{n} \in \mathcal{N}_0} m_{\boldsymbol{n}}$ denote the total number of households in the population, $N_j = \sum_{\boldsymbol{n} \in \mathcal{N}_0} n_j m_{\boldsymbol{n}}$ denote the total number of individuals of class j in the population $(j \in \mathcal{J})$ and $N = \sum_{\boldsymbol{n} \in \mathcal{N}_0} |\boldsymbol{n}| m_{\boldsymbol{n}}$ denote the total number of individuals in the population. Assume that N, and hence N_j $(j \in \mathcal{J})$ and m, is finite. This implies that $m_{\boldsymbol{n}} = 0$ for all but finitely many $\boldsymbol{n}$. Let $\mathcal{N} = \{\boldsymbol{n} \in \mathcal{N}_0 : m_{\boldsymbol{n}} > 0\}$.

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 $j \in \mathcal{J}$, 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 t_j. For $i, j \in \mathcal{J}$,

throughout its 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, it makes *local* contacts with any given susceptible of class j in its own household at the points of a homogeneous Poisson process having rate λ_{ij}^L. 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 are no infectives present in the population.

For ease of exposition, it is assumed that there is no latent period and that an infectious individual can make both local and global contacts with susceptibles in its own household. However, these are no real restrictions given the purpose of the paper. The threshold behaviour of an SIR epidemic model is a function of its final outcome, the distribution of which is invariant to very general assumptions concerning a latent period (see, for example, Ludwig [29], Ball [7] and, in a households setting, Ball et al. [12]). Also, it may seem more natural to formulate the model so that global contacts can only occur between individuals from distinct households. However, if the model is formulated in that way then, provided that any individual to individual local contact rate is larger than its corresponding global contact rate (a very plausible condition in practice), it is straightforward to recast the model into the above form (cf. Ball et al. [12], Section 3.1).

2.2 Threshold behaviour and final outcome

2.2.1 Threshold parameter

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 $\mathcal{J}$, where the type of an infected household is given by the class of its initial (globally contacted) infective.

The above approximation of the epidemic process by a multitype branching process can be made mathematically fully rigorous by considering a sequence of epidemics in which $m \to \infty$ and using a coupling argument, see Ball and Lyne [9]. A threshold theorem for the epidemic process can then be obtained by saying that a *global epidemic* occurs if, in the limit as $m \to \infty$, the epidemic infects infinitely many households, i.e. if the branching process does not go extinct. Let $M = [m_{ij}]$, where for $i, j \in \mathcal{J}$, 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, i.e. $0 \leq m_{ij} < \infty$ $(i, j \in \mathcal{J})$ 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, by standard multitype branching process theory (for example, Mode [30], Chapter 1, Theorem 7.1), for large m, a global epidemic occurs with non-zero probability if and only if $R_* > 1$. Thus R_* is a threshold parameter for the multitype households epidemic model.

In order to compute R_*, expressions for m_{ij} $(i, j \in \mathcal{J})$ are required. For $\boldsymbol{n} \in \mathcal{N}$, let $\alpha_{\boldsymbol{n}} = m_{\boldsymbol{n}}/m$ denote the proportion of households in the population that have category $\boldsymbol{n}$ and, for $i \in \mathcal{J}$ and $\boldsymbol{n} \in \mathcal{N}$, let $\alpha_i(\boldsymbol{n}) = n_i m_{\boldsymbol{n}}/N_i$ be the probability that a class i individual chosen at random in the population resides in household of category $\boldsymbol{n}$. Consider a completely susceptible household of category $\boldsymbol{n}$ 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 large population (branching process) limit all global contacts are with individuals in completely susceptible households. For $j \in \mathcal{J}$, 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 $\mu_{\boldsymbol{n},i,j}(\Lambda^L) = E[Y_j]$, where $\Lambda^L = [\lambda_{ij}^L]$, and note that by Wald's identity for multitype SIR epidemics (Ball [7], Corollary 3.2), $E[T_j^A] =$

$E[T_I^{(j)}]\mu_{\boldsymbol{n},i,j}(\Lambda^L) = t_j\mu_{\boldsymbol{n},i,j}(\Lambda^L)$. During the above single household epidemic, for $k \in \mathcal{J}$, each class k infective makes class j global contacts at total rate λ_{kj}^G, so the total number of global class j global contacts that emanate from this single household epidemic follows a Poisson distribution with random mean $\sum_{k\in\mathcal{J}} T_k^A\lambda_{kj}^G$. Thus the expected total number of such class j global contacts is $\sum_{k\in\mathcal{J}} t_k\mu_{\boldsymbol{n},i,k}(\Lambda^L)\lambda_{kj}^G$. Finally, conditioning on the household category of a typical type i infected household yields, as in Ball and Lyne [9], Section 4.3, that

$$m_{ij} = \sum_{\boldsymbol{n}\in\mathcal{N}} \alpha_i(\boldsymbol{n}) \sum_{k\in\mathcal{J}} \mu_{\boldsymbol{n},i,k}(\Lambda^L)t_k\lambda_{kj}^G \quad (i,j \in \mathcal{J}). \tag{2.1}$$

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

2.2.2 Final outcome in the event of a global epidemic

Still assuming that the total number of households m is large and that the number of initial infectives is small, suppose that a global epidemic occurs. For $i \in \mathcal{J}$, let z_i denote the expected proportion of class i susceptibles that are ultimately infected and let T_i denote the sum of the infectious periods of all the class i infectives present during the epidemic. Fix attention on a household that did not contain any initial infectives. For $i \in \mathcal{J}$, the probability that a given class i individual avoids global infection throughout the entire epidemic is given by

$$\pi_i = \exp(-\sum_{j\in\mathcal{J}} T_j\lambda_{ji}^G/N_i).$$

For $i \in \mathcal{J}$, let $\gamma_i = N_i/N$ be the proportion of individuals in the population that are of class i. Now, since m is large, for $j \in \mathcal{J}$, T_j is approximately $N_j z_j E[T_I^{(j)}] = N_j z_j t_j$, so the probability that a given class i individual avoids global infection during the epidemic is approximately given by

$$\pi_i = \exp\left(-\sum_{j\in\mathcal{J}} \gamma_j z_j t_j \lambda_{ji}^G/\gamma_i\right) \quad (i \in \mathcal{J}). \tag{2.2}$$

Further, for large m, distinct individuals avoid global infection approximately independently of each other. Thus the ultimate spread of infection within the household under

consideration is approximately distributed as that of a multitype single household epidemic model, studied by Addy et al. [1], in which, in addition to local infection, 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 susceptible ($i \in \mathcal{J}$). For $i \in \mathcal{J}$, let $\mu_{\boldsymbol{n},i}(\Lambda^L, \boldsymbol{\pi})$ be the expected number of class i individuals that are ultimately infected by this epidemic, where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_J)$ and $\boldsymbol{n}$ denotes the category of the household under consideration.

For $i \in \mathcal{J}$, z_i can be interpreted as the probability that an initial class i susceptible chosen at random from the population is ultimately infected by the epidemic. By conditioning on the category of household in which this initial susceptible resides and noting that if it resides in a household of category $\boldsymbol{n}$ then its chance of ultimate infection is $\mu_{\boldsymbol{n},i}(\Lambda^L, \boldsymbol{\pi})/n_i$, it follows that

$$z_i = \sum_{\boldsymbol{n} \in \mathcal{N}} \alpha_i(\boldsymbol{n}) \mu_{\boldsymbol{n},i}(\Lambda^L, \boldsymbol{\pi})/n_i \quad (i \in \mathcal{J}), \tag{2.3}$$

which, together with (2.2), is a set of J implicit equations for $\boldsymbol{z} = (z_1, z_2, \ldots, z_J)$. Note that $\boldsymbol{z} = \boldsymbol{0}$ is a root of (2.3). It is shown in Ball and Lyne [9], Section 5.2, that, provided the $J \times J$ matrix A having elements $a_{ij} = \sum_{k \in \mathcal{J}} t_i \lambda^G_{ik} \sum_{\boldsymbol{n} \in \mathcal{N}} \alpha_k(\boldsymbol{n}) \mu_{\boldsymbol{n},k,j}(\Lambda^L)$ $(i, j \in \mathcal{J})$ is positively regular, if $R_* \leq 1$ then $\boldsymbol{z} = \boldsymbol{0}$ is the only solution of (2.3) in $[0,1]^J$, while if $R_* > 1$ then there is a unique second root, with $z_i > 0$ $(i \in \mathcal{J})$, yielding the expected proportion of individuals of different classes that are infected by a global epidemic.

The above approximations become exact in the limit as $m \to \infty$ in an appropriate manner and the heuristic arguments presented here can be made fully rigorous by adapting the embedding technique of Scalia-Tomba [33, 34]; see Ball and Lyne [9] for details. An algorithm for computing $\mu_{\boldsymbol{n},i}(\Lambda^L, \boldsymbol{\pi})$ $(\boldsymbol{n} \in \mathcal{N}, i \in \mathcal{J})$ is given in the Appendix.

2.3 Vaccination

2.3.1 All or nothing vaccines

Suppose that the vaccine either renders its recipient completely immune or it has no effect, and that vaccinated individuals are rendered immune independently, with probability

ϵ_i for a class i individual ($i \in \mathcal{J}$). For $\boldsymbol{n} \in \mathcal{N}$ and $\mathbf{0} \leq \boldsymbol{r} = (r_1, r_2, \ldots, r_J) \leq \boldsymbol{n}$, where inequalities between vectors are to be interpreted elementwise, let $v_{\boldsymbol{n},\boldsymbol{r}}$ denote the proportion of households of category $\boldsymbol{n}$ that have had $\boldsymbol{r}$ members vaccinated, and let $\boldsymbol{v} = \{v_{\boldsymbol{n},\boldsymbol{r}} : \boldsymbol{n} \in \mathcal{N}, \mathbf{0} \leq \boldsymbol{r} \leq \boldsymbol{n}\}$.

For $i, j \in \mathcal{J}$, let $m_{ij}(\boldsymbol{v})$ 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 $\boldsymbol{n}$ having $\boldsymbol{r}$ members vaccinated is $\alpha_i(\boldsymbol{n})v_{\boldsymbol{n},\boldsymbol{r}}$. For $\boldsymbol{n} - \boldsymbol{r} \leq \boldsymbol{k} \leq \boldsymbol{n}$, such a household has $\boldsymbol{k}$ susceptible individuals if $\boldsymbol{n} - \boldsymbol{k}$ of the vaccinations are successful, which happens with probability $\binom{\boldsymbol{r}}{\boldsymbol{n}-\boldsymbol{k}}\boldsymbol{\epsilon}^{\boldsymbol{n}-\boldsymbol{k}}(\mathbf{1}-\boldsymbol{\epsilon})^{\boldsymbol{r}-\boldsymbol{n}+\boldsymbol{k}}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_J)$, $\binom{\boldsymbol{r}}{\boldsymbol{n}-\boldsymbol{k}} = \prod_{l=1}^{J}\binom{r_l}{n_l-k_l}$, $\mathbf{1}$ denotes the row vector of J ones and, for two row vectors $\boldsymbol{x}, \boldsymbol{y}$ of length J, $\boldsymbol{x}^{\boldsymbol{y}} = \prod_{l=1}^{J} x_l^{y_l}$. Further, given that $\boldsymbol{k}$ 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 k_i/n_i. Hence, for $i, j \in \mathcal{J}$,

$$m_{ij}(\boldsymbol{v}) = \sum_{\boldsymbol{n}\in\mathcal{N}} \alpha_i(\boldsymbol{n}) \sum_{\boldsymbol{r}=\mathbf{0}}^{\boldsymbol{n}} v_{\boldsymbol{n},\boldsymbol{r}} \sum_{\boldsymbol{k}=\boldsymbol{n}-\boldsymbol{r}}^{\boldsymbol{n}} \binom{\boldsymbol{r}}{\boldsymbol{n}-\boldsymbol{k}} \boldsymbol{\epsilon}^{\boldsymbol{n}-\boldsymbol{k}}(\mathbf{1}-\boldsymbol{\epsilon})^{\boldsymbol{r}-\boldsymbol{n}+\boldsymbol{k}} \frac{k_i}{n_i} \sum_{l\in\mathcal{J}} \mu_{\boldsymbol{k},i,l}(\Lambda^L) t_l \lambda^G_{lj}, \tag{2.4}$$

where, for example, $\sum_{\boldsymbol{r}=\mathbf{0}}^{\boldsymbol{n}} = \sum_{r_1=0}^{n_1}\sum_{r_2=0}^{n_2}\cdots\sum_{r_J=0}^{n_J}$.

Let $M(\boldsymbol{v}) = [m_{ij}(\boldsymbol{v})]$ and $R_*^{AoN}(\boldsymbol{v})$ be the maximal eigenvalue of $M(\boldsymbol{v})$. Then $R_*^{AoN}(\boldsymbol{v})$ is a threshold parameter for the epidemic after vaccination with an all or nothing (AoN) vaccine, in the sense that a global epidemic can occur only if $R_*^{AoN}(\boldsymbol{v}) > 1$. Consequently, a vaccination scheme $\boldsymbol{v}$ having $R_*^{AoN}(\boldsymbol{v}) \leq 1$ is protective for the whole community, the aim of launching a vaccination programme. It may seem more natural to consider instead $\tilde{R}_*^{AoN}(\boldsymbol{v})$, the maximal eigenvalue of the matrix $\tilde{M}(\boldsymbol{v}) = [\tilde{m}_{ij}(\boldsymbol{v})]$, where $\tilde{m}_{ij}(\boldsymbol{v})$ is the expected number of class j individuals that will be infected globally from a typical type i infected household at the start of the epidemic. Thus $\tilde{R}_*^{AoN}(\boldsymbol{v})$ is based on global infections (i.e. global contacts with individuals who are susceptible to infection), whilst $R_*^{AoN}(\boldsymbol{v})$ is based on global contacts, irrespective of the vaccine status of contacted individuals. It follows after a little algebra that the matrices $M(\boldsymbol{v})$ and $\tilde{M}(\boldsymbol{v})$ are similar, and thus

possess the same eigenvalues, so $R_*^{AoN}(\boldsymbol{v}) = \tilde{R}_*^{AoN}(\boldsymbol{v})$, as would be expected on intuitive grounds. However, $M(\boldsymbol{v})$ is easier to calculate and analyse than $\tilde{M}(\boldsymbol{v})$. Let $\boldsymbol{v}_0$ denote the vaccination scheme having $v_{\boldsymbol{n},\boldsymbol{r}} = 1$ if $\boldsymbol{r} = \boldsymbol{0}$ and 0 otherwise ($\boldsymbol{n} \in \mathcal{N}$), so there is no vaccination. Then $M(\boldsymbol{v}_0) = M$ and $R_*^{AoN}(\boldsymbol{v}_0) = R_*$, as it should.

In general, there is no closed form expression for $R_*^{AoN}(\boldsymbol{v})$. However, if the global infection rates take the proportionate mixing form (see, for example, Hethcote and Van Ark [26] or Becker and Marschner [17]) $\lambda_{ij}^G = \alpha_i^G \beta_j^G$ ($i, j \in \mathcal{J}$), then the matrix $M(\boldsymbol{v})$ has rank one, so $R_*^{AoN}(\boldsymbol{v})$ is given by its trace, i.e.

$$R_*^{AoN}(\boldsymbol{v}) = \sum_{i \in \mathcal{J}} \sum_{\boldsymbol{n} \in \mathcal{N}} \alpha_i(\boldsymbol{n}) \sum_{\boldsymbol{r}=\boldsymbol{0}}^{\boldsymbol{n}} v_{\boldsymbol{n},\boldsymbol{r}} \sum_{\boldsymbol{k}=\boldsymbol{n}-\boldsymbol{r}}^{\boldsymbol{n}} \binom{\boldsymbol{r}}{\boldsymbol{n}-\boldsymbol{k}} \boldsymbol{\epsilon}^{\boldsymbol{n}-\boldsymbol{k}} (\boldsymbol{1}-\boldsymbol{\epsilon})^{\boldsymbol{r}-\boldsymbol{n}+\boldsymbol{k}} \frac{k_i}{n_i} \sum_{l \in \mathcal{J}} \mu_{\boldsymbol{k},i,l}(\Lambda^L) t_l \alpha_l^G \beta_i^G. \tag{2.5}$$

2.3.2 Leaky vaccines

Suppose now that, instead of vaccinated individuals acquiring either complete immunity or no immunity at all, all vaccinees respond by acquiring partial immunity. To be more specific, assume that, for all $j \in \mathcal{J}$, all infection rates to class j individuals are reduced by a factor ϵ_j. Hence, for $i, j \in \mathcal{J}$, the rate at which a class i infective has global contact with a vaccinated class j individual is $\lambda_{ij}^G(1-\epsilon_j)/N_j$ and the corresponding local contact rate is $\lambda_{ij}^L(1-\epsilon_j)$. Note that the *average* vaccine efficacy is the same as in the all or nothing case. As before, a vaccination scheme is specified by $\boldsymbol{v} = \{v_{\boldsymbol{n},\boldsymbol{r}} : \boldsymbol{n} \in \mathcal{N}, \boldsymbol{0} \leq \boldsymbol{r} \leq \boldsymbol{n}\}$, where $v_{\boldsymbol{n},\boldsymbol{r}}$ is the proportion of category $\boldsymbol{n}$ households that have $\boldsymbol{r}$ individuals vaccinated.

Just as in the all or nothing case, it is necessary to derive expressions for $m_{ij}(\boldsymbol{v})$ corresponding to (2.4) and ultimately for $R_*^{Le}(\boldsymbol{v})$, the maximal eigenvalue of the matrix with elements $m_{ij}(\boldsymbol{v})$ assuming a leaky vaccine. In order to do this, it is convenient to introduce some new notation. After a vaccination scheme, there may be $2J$ classes of individual in the population, i.e. vaccinated and unvaccinated individuals for each of the J original classes. Let $\mu_{\boldsymbol{n}-\boldsymbol{r},\boldsymbol{r},\mathrm{u}:i,l}(\Lambda^L, \boldsymbol{\epsilon})$ ($\mu_{\boldsymbol{n}-\boldsymbol{r},\boldsymbol{r},\mathrm{v}:i,l}(\Lambda^L, \boldsymbol{\epsilon})$) denote the expected number of infected class l individuals, counting *both* vaccinated and unvaccinated individuals, in a category $\boldsymbol{n}$ household having $\boldsymbol{r}$ vaccinated, and hence $\boldsymbol{n}-\boldsymbol{r}$ unvaccinated, individuals, initiated by an infectious unvaccinated (vaccinated) class i individual, neglecting further

outside infections.

As in Section 2.3.1, for $i, j \in \mathcal{J}$, let $m_{ij}(\boldsymbol{v})$ 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 individuals 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 $(1-\epsilon_i)$, 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 $r_i(1-\epsilon_i)/n_i$. Consequently, in this leaky vaccine case,

$$m_{ij}(\boldsymbol{v}) = \sum_{\boldsymbol{n}\in\mathcal{N}} \alpha_i(\boldsymbol{n}) \sum_{\boldsymbol{r}=\boldsymbol{0}}^{\boldsymbol{n}} v_{\boldsymbol{n},\boldsymbol{r}} \sum_{k\in\mathcal{J}} \left(\frac{n_i - r_i}{n_i} \mu_{\boldsymbol{n}-\boldsymbol{r},\boldsymbol{r},\mathrm{u}:i,k}(\Lambda^L, \boldsymbol{\epsilon}) + \frac{r_i(1-\epsilon_i)}{n_i} \mu_{\boldsymbol{n}-\boldsymbol{r},\boldsymbol{r},\mathrm{v}:i,k}(\Lambda^L, \boldsymbol{\epsilon}) \right) t_k \lambda^G_{kj}, \tag{2.6}$$

and $R^{Le}_*(\boldsymbol{v})$ is the maximal eigenvalue of the matrix $M(\boldsymbol{v}) = [m_{ij}(\boldsymbol{v})]$. As before, a global epidemic can occur only if $R^{Le}_*(\boldsymbol{v}) > 1$, implying that the main goal of a vaccination scheme is to make $R^{Le}_*(\boldsymbol{v}) \leq 1$. Note that if Λ^G takes the proportionate mixing form, $\lambda^G_{ij} = \alpha^G_i \beta^G_j$ $(i, j \in \mathcal{J})$, then there is an explicit expression for $R^{Le}_*(\boldsymbol{v})$, analagous to (2.5).

2.3.3 Optimal vaccination schemes

In the previous two subsections, two different types of vaccine responses, and their effect on the threshold parameter R_* when part of the community (specified by $\boldsymbol{v}$) is vaccinated, have been considered. As noted above, the main aim of any vaccination scheme is to bring the threshold parameter below one, i.e. to ensure that $R_*(\boldsymbol{v}) \leq 1$. (The threshold parameter following the vaccination scheme $\boldsymbol{v}$ is referred to generically as $R_*(\boldsymbol{v})$; $R_*(\boldsymbol{v}) = R^{AoN}_*(\boldsymbol{v})$ if the vaccine is all or nothing and $R^{Le}_*(\boldsymbol{v})$ if it is leaky.) Therefore, for a given community and a given vaccine response, the vaccination scheme $\boldsymbol{v}$ is said to be *preventive* (written $\boldsymbol{v} \in P$) if the induced threshold parameter satisfies $R_*(\boldsymbol{v}) \leq 1$.

If the vaccine response, or efficacy, $\boldsymbol{\epsilon}$ is not large enough, it could happen that no vaccination scheme is preventive. That is, even with $\boldsymbol{v}$ satisfying $v_{\boldsymbol{n},\boldsymbol{n}} = 1$ for all $\boldsymbol{n}$ and

$v_{\boldsymbol{n},\boldsymbol{r}} = 0$ $(\boldsymbol{r} \neq \boldsymbol{n})$, meaning that all individuals in all households are vaccinated, it may happen that $R_*(\boldsymbol{v}) > 1$. If this is the case a better vaccine or some other preventive measure, such as improving sanitary conditions, is the only way to surely prevent future global outbreaks.

On the other hand, if the vaccine response is large enough there will be many different vaccination schemes $\boldsymbol{v}$ satisfying $R_*(\boldsymbol{v}) \leq 1$. It is then important to determine which such scheme is the best in the sense that it requires the fewest vaccinations. Accordingly, if

$$S(\boldsymbol{v}) = \frac{\sum_{\boldsymbol{n}\in\mathcal{N}} \sum_{\boldsymbol{r}=\boldsymbol{0}}^{\boldsymbol{n}} |\boldsymbol{r}| v_{\boldsymbol{n},\boldsymbol{r}} \alpha_{\boldsymbol{n}}}{\sum_{\boldsymbol{n}\in\mathcal{N}} |\boldsymbol{n}| \alpha_{\boldsymbol{n}}} \tag{2.7}$$

denotes the proportion of the population that are vaccinated (i.e. the overall vaccination coverage) under the scheme $\boldsymbol{v}$, then any scheme

$$\boldsymbol{v}_{\text{opt}} \in \underset{\boldsymbol{v}\in P}{\text{argmin}}\{S(\boldsymbol{v})\} = \{\boldsymbol{v}' \in P : S(\boldsymbol{v}') \leq S(\boldsymbol{v}) \text{ for all } \boldsymbol{v} \in P\}.$$

is optimal. The definition of $\boldsymbol{v}_{\text{opt}}$ could be generalised to incorporate costs associated with the practical implementation of a vaccination scheme, for example by including an additional cost per household having individuals vaccinated (cf. Ball and Lyne [10]).

It is a non-trivial problem to derive $\boldsymbol{v}_{\text{opt}}$, particularly since, in general, $R_*(\boldsymbol{v})$ does not admit a closed-form expression. However, if the global infection rates take the proportionate mixing form then $R_*(\boldsymbol{v})$ and $S(\boldsymbol{v})$ are both linear functions of $\boldsymbol{v}$, so determining the allocation of vaccines which (a) minimises $R_*(\boldsymbol{v})$ subject to an upper bound on $S(\boldsymbol{v})$ or (b) minimises $S(\boldsymbol{v})$ subject to $R_*(\boldsymbol{v}) \leq 1$ are both linear programming problems, cf. Becker and Starczak [18] and Ball et al. [8]. Note that there are further (linear) constraints on $\boldsymbol{v}$ implicit in the above formulations, specifically that, for $\boldsymbol{n} \in \mathcal{N}$, $v_{\boldsymbol{n},\boldsymbol{r}} \geq 0$ $(\boldsymbol{0} \leq \boldsymbol{r} \leq \boldsymbol{n})$ and $\sum_{\boldsymbol{r}=\boldsymbol{0}}^{\boldsymbol{n}} v_{\boldsymbol{n},\boldsymbol{r}} = 1$.

3 Estimation

3.1 Estimation of local and global infection parameters

In order to estimate the threshold parameter $R_*(\boldsymbol{v})$ 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 $T_I^{(i)}$ $(i \in \mathcal{J})$ are assumed known from previous epidemiological studies.

Suppose that the final outcome of the previous outbreak is observed in a sample of households. The following method for estimating (Λ^L, Λ^G), where $\Lambda^G = [\lambda_{ij}^G]$, is studied in Ball and Lyne [11]. Label the m households in the population $1, 2, \ldots, m$. For $i = 1, 2, \ldots, m$, let $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \ldots, t_{iJ})$, where t_{ij} is the number of class j susceptibles ultimately infected in household i, let $\boldsymbol{n}(i)$ be the category of household i and let $\delta_i = 1(0)$ if household i is observed (unobserved). For $\boldsymbol{n} \in \mathcal{N}$ and $\boldsymbol{0} \le \boldsymbol{t} \le \boldsymbol{n}$, let $p_{\boldsymbol{n}}(\boldsymbol{t}|\Lambda^L, \boldsymbol{\pi})$ be the probability that the multitype single household epidemic model with outside infection, studied by Addy et al. [1] and described in Section 2.2.2, has final outcome $\boldsymbol{t}$. For $\boldsymbol{n} \in \mathcal{N}$, a triangular system of linear equations governing $p_{\boldsymbol{n}}(\boldsymbol{t}|\Lambda^L, \boldsymbol{\pi})$ $(\boldsymbol{0} \le \boldsymbol{t} \le \boldsymbol{n})$ is given in the Appendix.

Suppose that a global epidemic occurs. Then (2.2) and (2.3) implicitly determine $\boldsymbol{\pi}$ as a function of (Λ^L, Λ^G), so write $\boldsymbol{\pi} = \boldsymbol{\pi}(\Lambda^L, \Lambda^G)$. Let $\boldsymbol{t}_D = \{\boldsymbol{t}_i : \delta_i = 1\}$ denote the observed data. There does not exist a feasible method for computing the likelihood of (Λ^L, Λ^G) given $\boldsymbol{t}_D$, so consider estimating (Λ^L, Λ^G) by maximising the pseudolikelihood

$$L(\Lambda^L, \Lambda^G|\boldsymbol{t}_D) = \prod_{i=1}^{m} \Big\{ p_{\boldsymbol{n}(i)}\big(\boldsymbol{t}_i|\Lambda^L, \boldsymbol{\pi}(\Lambda^L, \Lambda^G)\big) \Big\}^{\delta_i}. \tag{3.1}$$

Note that (3.1) is a pseudolikelihood, and not a likelihood, since the outcomes in different households are not independent. These outcomes become independent in the limit as $m \to \infty$ but they are weakly dependent (their covariance is of order $1/m$) for large but finite m.

The pseudolikelihood (3.1) can be maximised by first maximising it as a function of $(\Lambda^L, \boldsymbol{\pi})$, to yield the estimate $(\hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$, then obtaining an estimate, $\hat{\boldsymbol{z}}$ say, of $\boldsymbol{z}$ by substituting $(\hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$ in the right hand side of (2.3), and finally solving (2.2), with $(\boldsymbol{\pi}, \boldsymbol{z})$ replaced by $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{z}})$ for Λ^G. For single type epidemics $(J = 1)$, the resulting estimate of (Λ^L, Λ^G), which are both scalars, corresponds to that described in Ball et al. [12], Section 5.1, 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} $(i, j \in \mathcal{J})$, so Λ^G is not identifiable from the observed data using this approach and the threshold parameters before and after vaccination, R_* and $R_*(\boldsymbol{v})$, cannot be estimated consistently. It is possible that the local infection rates Λ^L may also be unidentifiable, for example if for some $i, j \in \mathcal{J}$ 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_*, $R_*^{AoN}(v)$ and $R_*^{Le}(v)$

In the previous subsection final size data, from a sample of households of one epidemic outbreak, were used to derive estimates of the matrix Λ^L and the vectors $\boldsymbol{\pi}$ and $\boldsymbol{z}$. Estimation of the, epidemiologically more important parameters R_* and $R_*(\boldsymbol{v})$ is considered now, under the assumption that the population structure is sufficiently rich for Λ^L to be identifiable. The vaccination effect $\boldsymbol{\epsilon}$ and the type (all or nothing or leaky) of the vaccine are assumed known, as are the distributions of $T_I^{(i)}$, $i = 1, \ldots, \mathcal{J}$. 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 $R_*(\boldsymbol{v})$ for some future epidemic in a community with different household structure. Note that this should only be done if it is considered reasonable to extrapolate parameter estimates from the sample to the future population. Let $\tilde{\alpha}_{\boldsymbol{n}}$ denote the proportion of households in the future population that have category $\boldsymbol{n}$ (for $\boldsymbol{n} \in \tilde{\mathcal{N}}$, with the obvious definition of $\tilde{\mathcal{N}}$) and, for $i \in \mathcal{J}$ and $\boldsymbol{n} \in \tilde{\mathcal{N}}$, let $\tilde{\alpha}_i(\boldsymbol{n})$ be the probability that a class i individual chosen at random in the future population resides in household of category $\boldsymbol{n}$. For the remainder of Section 3 and for Section 4, it is assumed that estimation of R_* and $R_*(\boldsymbol{v})$ is for a population with household structure given by $\tilde{\alpha}_{\boldsymbol{n}}$ $(\boldsymbol{n} \in \tilde{\mathcal{N}})$. Further, when referring to formulae in Section 2 for the mean matrices M and $M(\boldsymbol{v})$, it is assumed implicitly that $\alpha_i(\boldsymbol{n})$ has been replaced

by $\tilde{\alpha}_i(\boldsymbol{n})$.

3.2.1 Estimation of R_*

Recall that R_* is the maximal eigenvalue of the matrix M having elements m_{ij} given by (2.1). In the expression for m_{ij} the quantities $\tilde{\alpha}_i(\boldsymbol{n})$ and $t_k = E(T_I^{(k)})$ are known, and $\mu_{\boldsymbol{n},i,k}(\Lambda^L)$ is estimated consistently by $\mu_{\boldsymbol{n},i,k}(\hat{\Lambda}^L)$. However, for $J > 1$, the matrix $\Lambda^G = [\lambda_{ij}^G]$ (and hence R_* and $R_*(\boldsymbol{v})$, which are functions of (Λ^L, Λ^G)) cannot be estimated consistently. Nevertheless, Λ^G is known to satisfy the constraints given by (2.2), where $\boldsymbol{\pi}$ and $\boldsymbol{z}$ can be estimated consistently. Thus the Perron-Frobenius theorem is used to obtain bounds on R_* and $R_*(\boldsymbol{v})$, which are functions of $(\boldsymbol{\pi}, \boldsymbol{z})$ and thus can be estimated consistently. Similar methods were used for a multitype epidemic model without household structure by Britton [20].

By the Perron-Frobenius theorem (e.g. Jagers [27], page 92) it follows that there is a unique (up to normalisation) vector $(x_1, x_2, \ldots, x_J)$ satisfying

$$R_* x_j = \sum_{i=1}^{J} x_i m_{ij} \qquad (j = 1, 2, \ldots, J).$$

Substituting the expression (2.1) for m_{ij} yields

$$R_* x_j = \sum_{i,\boldsymbol{n},k} x_i \tilde{\alpha}_i(\boldsymbol{n}) t_k \mu_{\boldsymbol{n},i,k}(\Lambda^L) \lambda_{kj}^G = \sum_{k=1}^{J} \frac{1}{\gamma_k z_k} \gamma_k z_k t_k \lambda_{kj}^G \sum_{i,\boldsymbol{n}} x_i \tilde{\alpha}_i(\boldsymbol{n}) \mu_{\boldsymbol{n},i,k}(\Lambda^L).$$

Define the final sum by $r_k = \sum_{i,\boldsymbol{n}} x_i \tilde{\alpha}_i(\boldsymbol{n}) \mu_{\boldsymbol{n},i,k}(\Lambda^L)$. Further, let $A = \max_k\{r_k/\gamma_k z_k\}$ and assume that the maximum is attained for $k = k_0$. Then,

$$R_* x_j = \gamma_j \sum_{k=1}^{J} \frac{r_k}{\gamma_k z_k} \gamma_k z_k t_k \lambda_{kj}^G / \gamma_j \leq A \gamma_j \sum_{k=1}^{J} \gamma_k z_k t_k \lambda_{kj}^G / \gamma_j = A \gamma_j (-\log \pi_j) \quad (j \in \mathcal{J}),$$

where the final equality follows from (2.2). Hence, recalling the definition of r_k,

$$\frac{R_* r_k}{\gamma_k z_k} = \frac{1}{\gamma_k z_k} \sum_{i,\mathbf{n}} R_* x_i \tilde{\alpha}_i(\boldsymbol{n}) \mu_{\boldsymbol{n},i,k}(\Lambda^L) \leq \frac{A}{\gamma_k z_k} \sum_{i,\boldsymbol{n}} \gamma_i (-\log \pi_i) \tilde{\alpha}_i(\boldsymbol{n}) \mu_{\boldsymbol{n},i,k}(\Lambda^L).$$

In particular, for $k = k_0$ this yields

$$R_* \leq \frac{1}{\gamma_{k_0} z_{k_0}} \sum_{i,\boldsymbol{n}} \gamma_i (-\log \pi_i) \tilde{\alpha}_i(\boldsymbol{n}) \mu_{\boldsymbol{n},i,k_0}(\Lambda^L) \leq \max_k \frac{1}{\gamma_k z_k} \sum_{i,\boldsymbol{n}} \gamma_i (-\log \pi_i) \tilde{\alpha}_i(\boldsymbol{n}) \mu_{\boldsymbol{n},i,k}(\Lambda^L).$$

Identical arguments yield a similar lower bound for R_*, so

$$\min_k \frac{1}{\gamma_k z_k} \sum_{i,\boldsymbol{n}} \gamma_i(-\log \pi_i)\tilde{\alpha}_i(\boldsymbol{n})\mu_{\boldsymbol{n},i,k}(\Lambda^L) \leq R_* \leq \max_k \frac{1}{\gamma_k z_k} \sum_{i,\boldsymbol{n}} \gamma_i(-\log \pi_i)\tilde{\alpha}_i(\boldsymbol{n})\mu_{\boldsymbol{n},i,k}(\Lambda^L). \tag{3.2}$$

Note that the upper and lower bounds in (3.2) contain only known or estimable quantities. Of course, estimates of the bounds are obtained by replacing the unknown quantities π_i, z_k and $\mu_{\mathbf{n},i,k}(\Lambda^L)$ by their estimates.

Just like in the multitype case without household structure treated in Britton [20], these bounds are sharp, in that there exists $\Lambda^G = [\lambda^G_{ij}]$ satisfying (2.2), such that the corresponding maximal eigenvalue equals the right hand side of (3.2), and similarly for the lower bound. To see this, suppose the maximum on the right hand side of (2.2) is obtained for $k = k_1$. For each j, define $\lambda^G_{k_1 j} = (-\log \pi_j)\gamma_j/(\gamma_{k_1} z_{k_1} t_{k_1})$ and $\lambda^G_{kj} = 0$ for all other k. First, note that this choice for λ^G_{ij} satisfies (2.2). Second, inserting this choice into (2.1) yields

$$m_{ij} = \sum_{\boldsymbol{n}} \tilde{\alpha}_i(\boldsymbol{n})\mu_{\boldsymbol{n},i,k_1}(\Lambda^L)\frac{\gamma_j(-\log \pi_j)}{\gamma_{k_1} z_{k_1}} \quad (i,j \in \mathcal{J}),$$

from which it is evident that m_{ij} can be written as a product $a_i b_j$ where the first factor is independent of j and the second of i. Thus $M = [m_{ij}]$ has rank one and its maximal eigenvalue is given by its trace $\sum_i a_i b_i$ (e.g. [17]), which for these specific a_i's and b_j's is

$$\frac{1}{\gamma_{k_1} z_{k_1}} \sum_{\boldsymbol{n},i} \gamma_i(-\log \pi_i)\tilde{\alpha}_i(\boldsymbol{n})\mu_{\boldsymbol{n},i,k_1}(\Lambda^L).$$

But this is exactly the upper bound for R_*. A similar argument shows that the lower bound can also be attained. Further, linear interpolation between the two extreme choices for Λ^G, shows that R_* can take any value in the interval given by (3.2). Note from the above construction that the upper bound is attained when all global infections are caused by one single class of individual, viz. the class which maximizes the right hand side of (3.2).

3.2.2 Estimation of R_*^{AoN}

The same methods as for R_* can be applied to obtain bounds for $R_*^{AoN}(\boldsymbol{v})$, where now vaccinations have been performed according to the scheme $\boldsymbol{v}$, and the vaccine is assumed to have an all or nothing effect. Recall that $R_*^{AoN}(\boldsymbol{v})$ is the maximal eigenvalue of the matrix $M(\boldsymbol{v})$ with elements $m_{ij}(\boldsymbol{v})$ given by (2.4). Note that (2.4) can be written as

$$m_{ij}(\boldsymbol{v}) = \sum_{\boldsymbol{n}\in\tilde{\mathcal{N}}} \tilde{\alpha}_i(\boldsymbol{n}) \sum_{l\in\mathcal{J}} b_{il}^{\boldsymbol{n},\boldsymbol{v}} t_l \lambda_{lj}^G, \tag{3.3}$$

where

$$b_{il}^{\boldsymbol{n},\boldsymbol{v}} = b_{il}^{\boldsymbol{n},\boldsymbol{v}}(\Lambda^L, \boldsymbol{\epsilon}) = \sum_{\boldsymbol{r}=\boldsymbol{0}}^{\boldsymbol{n}} v_{\boldsymbol{n},\boldsymbol{r}} \sum_{\boldsymbol{k}=\boldsymbol{n}-\boldsymbol{r}}^{\boldsymbol{n}} \binom{\boldsymbol{r}}{\boldsymbol{n}-\boldsymbol{k}} \boldsymbol{\epsilon}^{\boldsymbol{n}-\boldsymbol{k}} (\boldsymbol{1}-\boldsymbol{\epsilon})^{\boldsymbol{r}-\boldsymbol{n}+\boldsymbol{k}} \frac{k_i}{n_i} \mu_{\boldsymbol{k},i,l}(\Lambda^L). \tag{3.4}$$

The only difference between (3.3) and (2.1), with k replaced by l, is that $\mu_{\boldsymbol{n},i,l}(\Lambda^L)$ has been replaced by $b_{il}^{\boldsymbol{n},\boldsymbol{v}}$. Further, $b_{il}^{\boldsymbol{n},\boldsymbol{v}}$ contains only known or estimable quantities. A similar argument as for R_* then shows that R_*^{AoN} cannot be estimated consistently but that it can be bounded by

$$\min_k \frac{1}{\gamma_k z_k} \sum_{i,\boldsymbol{n}} \gamma_i(-\log\pi_i)\tilde{\alpha}_i(\boldsymbol{n}) b_{ik}^{\boldsymbol{n},\boldsymbol{v}} \le R_*^{AoN}(\boldsymbol{v}) \le \max_k \frac{1}{\gamma_k z_k} \sum_{i,\boldsymbol{n}} \gamma_i(-\log\pi_i)\tilde{\alpha}_i(\boldsymbol{n}) b_{ik}^{\boldsymbol{n},\boldsymbol{v}}. \tag{3.5}$$

The bounds contain only known quantities and estimable parameters, so estimates of the bounds are obtained simply by replacing the unknown parameters by their estimates. A vaccination scheme for this type of vaccine is surely preventive (i.e. secure) only if the upper limit does not exceed 1.

3.2.3 Estimation of R_*^{Le}

Bounds on $R_*^{Le}(\boldsymbol{v})$, the maximal eigenvalue of the matrix $M(\boldsymbol{v})$ having elements given by (2.6), i.e. assuming a leaky vaccine, are derived in a similar fashion to those for R_* and R_*^{AoN}. Note that $m_{ij}(\boldsymbol{v})$ can be rewritten as

$$m_{ij}(\boldsymbol{v}) = \sum_{\boldsymbol{n}\in\tilde{\mathcal{N}}} \tilde{\alpha}_i(\boldsymbol{n}) \sum_{l\in\mathcal{J}} c_{il}^{\boldsymbol{n},\boldsymbol{v}} t_l \lambda_{lj}^G,$$

where

$$c_{il}^{\boldsymbol{n},\boldsymbol{v}} = c_{il}^{\boldsymbol{n},\boldsymbol{v}}(\Lambda^L, \boldsymbol{\epsilon}) = \sum_{\boldsymbol{r}=\boldsymbol{0}}^{\boldsymbol{n}} v_{\boldsymbol{n},\boldsymbol{r}} \left(\frac{n_i - r_i}{n_i} \mu_{\boldsymbol{n}-\boldsymbol{r},\boldsymbol{r},\mathrm{u}:i,l}(\Lambda^L, \boldsymbol{\epsilon}) + \frac{r_i(1-\epsilon_i)}{n_i} \mu_{\boldsymbol{n}-\boldsymbol{r},\boldsymbol{r},\mathrm{v}:i,l}(\Lambda^L, \boldsymbol{\epsilon}) \right). \tag{3.6}$$

Thus, arguing as for R_*, $R_*^{Le}(\boldsymbol{v})$ can be bounded by

$$\min_k \frac{1}{\gamma_k z_k} \sum_{i,\boldsymbol{n}} \gamma_i(-\log \pi_i) \tilde{\alpha}_i(\boldsymbol{n}) c_{ik}^{\boldsymbol{n},\boldsymbol{v}} \leq R_*^{Le}(\boldsymbol{v}) \leq \max_k \frac{1}{\gamma_k z_k} \sum_{i,\boldsymbol{n}} \gamma_i(-\log \pi_i) \tilde{\alpha}_i(\boldsymbol{n}) c_{ik}^{\boldsymbol{n},\boldsymbol{v}}. \tag{3.7}$$

Again, these bounds contain only known or estimable parameters, and hence are easily estimated. The vaccination scheme $\boldsymbol{v}$ is secure if the upper limit is below 1.

3.3 Estimation of the optimal vaccination scheme

In Section 2.3.3, an optimal vaccination scheme $\boldsymbol{v}_{\mathrm{opt}}$ was defined as a scheme $\boldsymbol{v}$ which minimises the overall vaccination coverage $S(\boldsymbol{v}) = \sum_{\boldsymbol{n},\boldsymbol{r}} |\boldsymbol{r}| v_{\boldsymbol{n},\boldsymbol{r}} \tilde{\alpha}_{\boldsymbol{n}} / \sum_{\boldsymbol{n}} |\boldsymbol{n}| \tilde{\alpha}_{\boldsymbol{n}}$ among those schemes that are preventive (a scheme $\boldsymbol{v}'$ is preventive if $R_*(\boldsymbol{v}') \leq 1$). Since $R_*(\boldsymbol{v})$ cannot be estimated consistently, be it a leaky or an all or nothing vaccine, it follows that $\boldsymbol{v}_{\mathrm{opt}}$ cannot be estimated consistently either. Instead, vaccination schemes with associated upper bound for $R_*(\boldsymbol{v}) \leq 1$ are considered.

Suppose that the vaccine is all or nothing. (The leaky case is similar and hence omitted.) Let

$$R_*^{(k)}(\boldsymbol{v}) = \frac{1}{\gamma_k z_k} \sum_{i \in \mathcal{J}} \sum_{\boldsymbol{n} \in \tilde{\mathcal{N}}} \gamma_i(-\log \pi_i) \tilde{\alpha}_i(\boldsymbol{n}) b_{ik}^{\boldsymbol{n},\boldsymbol{v}} \quad (k \in \mathcal{J}) \tag{3.8}$$

and

$$R_*^{max}(\boldsymbol{v}) = \max_k R_*^{(k)}(\boldsymbol{v}).$$

Then, from (3.5), any vaccination scheme $\boldsymbol{v}$ with $R_*^{max}(\boldsymbol{v}) \leq 1$ is preventive, irrespective of the underlying parameter Λ^G consistent with the data, whilst for any vaccination scheme $\boldsymbol{v}$ with $R_*^{max}(\boldsymbol{v}) > 1$ there exists Λ^G, consistent with the data, so that $R_*(\boldsymbol{v}) > 1$. Thus it is appropriate to consider minimisation of the vaccine coverage $S(\boldsymbol{v})$ subject to the constraints $R_*^{(k)}(\boldsymbol{v}) \leq 1$ $(k = 1, 2, \ldots, J)$. Note that this is a linear programming

problem since, by (2.7), (3.4) and (3.8), the objective function $S(\boldsymbol{v})$ and the constraints $R_*^{(k)}(\boldsymbol{v}) \leq 1$ $(k = 1, 2, \ldots, J)$ are all linear functions of the optimising variables $\boldsymbol{v}$; see Ball et al. [8] for discussion of the form of associated optimal vaccination schemes. Let $\boldsymbol{v}_{\rm opt}$ denote a solution to this minimisation problem and let $c_v = S(\boldsymbol{v}_{\rm opt})$ be the corresponding vaccination coverage. Thus c_v is the secure vaccination coverage required to be sure of preventing a future global outbreak. As noted previously, $R_*^{(k)}(\boldsymbol{v})$ is estimated by replacing the unknown parameters in the right hand side of (3.8) by their estimates, yielding $\hat{R}_*^{(k)}(\boldsymbol{v})$ say. Thus, $R_*^{max}(\boldsymbol{v})$ is estimated by $\hat{R}_*^{max}(\boldsymbol{v}) = \max_k \hat{R}_*^{(k)}(\boldsymbol{v})$ and c_v is estimated by solving the above linear programming problem, with $R_*^{(k)}(\boldsymbol{v})$ replaced by $\hat{R}_*^{(k)}(\boldsymbol{v})$ $(k \in \mathcal{J})$, yielding $\hat{\boldsymbol{v}}_{\rm opt}$ and $\hat{c}_v = S(\hat{\boldsymbol{v}}_{\rm opt})$.

4 Uncertainty

In this section, standard errors for the estimates $\hat{R}_*^{max}(\boldsymbol{v})$ and $\hat{c}_v$, and associated confidence intervals for $R_*^{max}(\boldsymbol{v})$ and c_v, are considered, in the situation when the number of households in the observed sample is large.

Let $\boldsymbol{\theta} = [vec(\Lambda^L), \boldsymbol{\pi}]$ $(= (\theta_1, \theta_2, \ldots, \theta_{J(J+1)}))$, where $vec(\Lambda^L)$ is the row vector representation of Λ^L, and let $\hat{\boldsymbol{\theta}}$ be the maximum pseudolikelihood estimate of $\boldsymbol{\theta}$. It is shown in Ball and Lyne [11] that, given the occurrence of a global epidemic, as the population and sample tend to infinity in an appropriate fashion, $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ and

$$m^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N\big(\mathbf{0}, \Sigma(\Lambda^L, \Lambda^G)\big) \text{ as } m \to \infty. \tag{4.1}$$

Note that the variance matrix $\Sigma(\Lambda^L, \Lambda^G)$ depends on Λ^G rather than on $\boldsymbol{\pi}$. The matrix $\Sigma(\Lambda^L, \Lambda^G)$ also depends on $\alpha_{\boldsymbol{n}}, \beta_{\boldsymbol{n}}$ $(\boldsymbol{n} \in \mathcal{N})$, where, for $\boldsymbol{n} \in \mathcal{N}$, $\beta_{\boldsymbol{n}}$ denotes the proportion of households of category $\boldsymbol{n}$ in the population that are in the observed sample. This latter dependence is suppressed for ease of notation. Calculation of $\Sigma(\Lambda^L, \Lambda^G)$ is described in Ball and Lyne [11] and is not reproduced here as it is rather complicated. Sufficient conditions for the limit (4.1) to hold are also given in Ball and Lyne [11]. These include the important practical case when the proportions $\alpha_{\boldsymbol{n}}, \beta_{\boldsymbol{n}}$ $(\boldsymbol{n} \in \mathcal{N})$ are held fixed as the number of households $m \to \infty$.

Recall from Section 3.3 that $R_*^{max}(\boldsymbol{v}) = \max_k R_*^{(k)}(\boldsymbol{v}, \Lambda^L, \boldsymbol{\pi})$, where $R_*^{(k)}(\boldsymbol{v}, \Lambda^L, \boldsymbol{\pi})$ ($k \in \mathcal{J}$) are given by the right hand side of (3.8) and their dependence on the unknown parameters Λ^L and $\boldsymbol{\pi}$ are shown explicitly. (Note that $(\Lambda^L, \boldsymbol{\pi})$ determines $\boldsymbol{z}$ by (2.3).) Let k_1 denote the k which maximises $R_*^{(k)}(\boldsymbol{v}, \Lambda^L, \boldsymbol{\pi})$ ($k \in \mathcal{J}$), and suppose, for ease of exposition, that k_1 is unique. Then $R_*^{max}(\boldsymbol{v})$ is estimated consistently by $\hat{R}_*^{max}(\boldsymbol{v}) = R_*^{(\hat{k}_1)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$, where $\hat{k}_1$ maximises $\hat{R}_*^{(k)}(\boldsymbol{v})$ ($k \in \mathcal{J}$). Further, let $\boldsymbol{d}(\boldsymbol{v}, \Lambda^L, \boldsymbol{\pi})$ be the $J(J+1)$ dimensional row vector whose ith element is the partial derivative of $R_*^{(k_1)}(\boldsymbol{v}, \Lambda^L, \boldsymbol{\pi})$ with respect to θ_i. Then an application of the delta-method (see, for example, Andersen et al [2]) shows that

$$m^{1/2}\big(\hat{R}_*^{max}(\boldsymbol{v}) - R_*^{max}(\boldsymbol{v})\big) \xrightarrow{D} N\big(0, \sigma^2(\boldsymbol{v}, \Lambda^L, \Lambda^G)\big) \text{ as } m \to \infty, \tag{4.2}$$

where

$$\sigma^2(\boldsymbol{v}, \Lambda^L, \Lambda^G) = \boldsymbol{d}(\boldsymbol{v}, \Lambda^L, \boldsymbol{\pi})\Sigma(\Lambda^L, \Lambda^G)\boldsymbol{d}(\boldsymbol{v}, \Lambda^L, \boldsymbol{\pi})^\top \tag{4.3}$$

and $\top$ denotes transpose.

In order to use (4.2) to obtain a confidence interval for $R_*^{max}(\boldsymbol{v})$, an estimate of Λ^G is required. Now $R_*(\boldsymbol{v})$ 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

$$\hat{\lambda}_{k_1 j}^G = (-\log \hat{\pi}_j)\gamma_j/(\gamma_{k_1}\hat{z}_{k_1}t_{k_1}) \quad (j \in \mathcal{J}), \tag{4.4}$$

where $\hat{z}_j$ is obtained by setting $(\Lambda^L, \boldsymbol{\pi}) = (\hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$ in (2.3). A one-sided $1-\alpha$ confidence interval for $R_*^{max}(\boldsymbol{v})$ is then given by $\big(0, \hat{R}_*^{max}(\boldsymbol{v}) + m^{-1/2}z_\alpha\sigma(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\Lambda}^G)\big)$, where z_α is the $(1-\alpha)$-quantile of the standard normal distribution. The asymptotic variance $\sigma^2(\boldsymbol{v}, \Lambda^L, \Lambda^G)$ may be larger for other choices of Λ^G consistent with $(\Lambda^L, \boldsymbol{\pi})$, but such choices will have $R_*(\boldsymbol{v}) < R_*^{max}(\boldsymbol{v})$. Thus the above confidence interval is asymptotically conservative.

Turn now to estimation of the secure vaccination coverage c_v under the optimal vaccination strategy, outlined in Section 3.3. For $c \in (0,1)$, let $\boldsymbol{v}^{opt}(c, \boldsymbol{\theta}) = \{v_{\boldsymbol{n},\boldsymbol{r}}^{opt}(c, \boldsymbol{\theta}) : \boldsymbol{n} \in \tilde{\mathcal{N}}, \boldsymbol{0} \leq \boldsymbol{r} \leq \boldsymbol{n}\}$ denote an optimal vaccination scheme, given that a proportion c of the population are to be vaccinated, where dependence on the unknown parameters

$\boldsymbol{\theta} = [vec(\Lambda^L), \boldsymbol{\pi}]$ is shown explicitly; i.e. $\boldsymbol{v}^{opt}(c, \boldsymbol{\theta})$ minimises $R_*^{max}(\boldsymbol{v})$ subject to $S(\boldsymbol{v}) \leq c$. Let $R(c, \boldsymbol{\theta}) = R_*^{max}\big(\boldsymbol{v}^{opt}(c, \boldsymbol{\theta})\big)$. Then, for fixed $\boldsymbol{\theta}$, $R(c, \boldsymbol{\theta})$ is a strictly decreasing, piecewise linear function of c and the secure vaccination coverage $c_v = c_v(\boldsymbol{\theta})$ is obtained by solving $R(c, \boldsymbol{\theta}) = 1$. In practice, $\boldsymbol{\theta}$ is unknown and c_v is estimated by $\hat{c}_v = c_v(\hat{\boldsymbol{\theta}})$.

Let $\boldsymbol{d}_c(\Lambda^L, \boldsymbol{\pi})$ be the $J(J+1)$ dimensional row vector whose ith element is $\partial c_v(\boldsymbol{\theta})/\partial\theta_i$. Then, by the delta-method,

$$m^{1/2}(\hat{c}_v - c_v) \xrightarrow{D} N\big(0, \sigma_c^2(\Lambda^L, \Lambda^G)\big) \quad \text{as} \quad m \to \infty,$$

where

$$\sigma_c^2(\Lambda^L, \Lambda^G) = \boldsymbol{d}_c(\Lambda^L, \boldsymbol{\pi})\Sigma(\Lambda^L, \Lambda^G)\boldsymbol{d}_c(\Lambda^L, \boldsymbol{\pi})^\top.$$

A one-sided $1-\alpha$ confidence interval for c_v is then given by $\big(0, \hat{c}_v + m^{-1/2} z_\alpha \sigma_c(\hat{\Lambda}^L, \hat{\Lambda}^G)\big)$, with $\hat{\Lambda}^G$ being given by (4.4). The lack of an explicit expression for $R(c, \boldsymbol{\theta})$ means that, unless $J = 1$ (cf. Britton and Becker [21]), the derivatives $\partial c_v(\boldsymbol{\theta})/\partial\theta_i$ need to be evaluated numerically, which is straightforward since $c_v(\boldsymbol{\theta})$ arises from the solution of a linear programming problem.

The above confidence interval for $R_*^{max}(\boldsymbol{v})$ assumes that the vaccination scheme $\boldsymbol{v}$ is fixed, whereas, in practice, for a given coverage c, a confidence interval may be required for $R(c, \boldsymbol{\theta})$, the post-vaccination threshold parameter assuming that the vaccines are allocated optimally. To obtain such an interval using the delta-method, the partial derivatives $\partial R(c, \boldsymbol{\theta})/\partial\theta_i$ $(i = 1, 2, \ldots, J(J+1))$ are required. These are difficult to obtain directly, even numerically, since the optimisation problem underlying $R(c, \boldsymbol{\theta})$ is a linear programming problem only when $J = 1$. For fixed $r \in (0, \infty)$, let $c_v^{(r)} = c_v^{(r)}(\boldsymbol{\theta})$ satisfy $R(c_v^{(r)}, \boldsymbol{\theta}) = r$. Then by the implicit function theorem

$$\frac{\partial c_v^{(r)}(\boldsymbol{\theta})}{\partial\theta_i} = -\frac{\partial R(c_v^{(r)}, \boldsymbol{\theta})}{\partial\theta_i} \Bigg/ \frac{\partial R(c_v^{(r)}, \boldsymbol{\theta})}{\partial c} \quad \big(i = 1, 2, \ldots, J(J+1)\big). \tag{4.5}$$

Now $c_v^{(r)}$ arises from the solution of the linear programming problem "minimise $S(\boldsymbol{v})$ subject to $R_*^{(k)} \leq r$ $(k = 1, 2, \ldots, J)$", enabling $\partial c_v^{(r)}(\boldsymbol{\theta})/\partial\theta_i$ and $\partial c_v^{(r)}(\boldsymbol{\theta})/\partial r$ to be calculated numerically. Also, $\partial R(c_v^{(r)}, \boldsymbol{\theta})/\partial c = 1/\big(\partial c_v^{(r)}(\boldsymbol{\theta})/\partial r\big)$, so $\partial R(c, \boldsymbol{\theta})/\partial\theta_i$ can be found by letting $r = R(c, \boldsymbol{\theta})$ and using (4.5).

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

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

Then $\big(0, R_*^U(\boldsymbol{v}, \alpha)\big)$ is a $1 - \alpha$ confidence interval for $R_*^{max}(\boldsymbol{v})$ 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 $\big(0, R_*^U(\boldsymbol{v}, \alpha)\big)$ be used in practice. A similar comment applies to the confidence intervals for c_v and $R(c, \boldsymbol{\theta})$. Indeed, the numerical examples in Section 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 (see Monto et al. [31]), 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, so the latter is difficult to estimate. 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) and higher (less susceptible) 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 ($\geq$ 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). 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 which 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. [1], 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 individuals and the category of a household is determined by the number of low and high titre individuals it contains. The following estimates are obtained, where class 1 is low antibody (titre) level and class 2 is high:

$$\hat{\Lambda}^L = \begin{pmatrix} 0.0536 & 0.0291 \\ 0.0000 & 0.0052 \end{pmatrix}, \hat{\boldsymbol{\pi}} = (0.8172\ 0.9196) \text{ and } \hat{\boldsymbol{z}} = (0.2339\ 0.1015).$$

Define $R_*^{(k)}(\hat{\Lambda}^L, \hat{\boldsymbol{\pi}}) = R_*^{(k)}(\boldsymbol{v}_0, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$, where $\boldsymbol{v}_0$ denotes the null vaccination scheme introduced in Section 2.3.1. The bounds $R_*^{(k)}(\hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$ can be calculated as 1.0868 and 1.1801, for $k = 1, 2$, respectively. An upper bound for the threshold parameter is $R_*^U(\boldsymbol{v}_0, 0.05) = 1.2832$, leading to a 95% confidence interval of $(0, 1.2832)$. Assuming an all or nothing vaccine with efficacy $\boldsymbol{\epsilon} = (0.7\ 0.7)$, the secure vaccination coverage c_v is estimated to be 0.0877 with a 95% confidence interval of $(0, 0.2424)$. (The current killed influenza vaccine has an efficacy of about 0.7, irrespective of prior immunity, Ira M. Longini, personal communication.) For the leaky case $\hat{c}_v = 0.0897$ with 95% confidence interval $(0, 0.2441)$, so slightly more vaccine is required than in the all or nothing case. The leaky case is illustrated in figure 1 (the all or nothing case would look very similar and is hence omitted), where $R_*^{(k)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$ $(k = 1, 2)$ and $R_*^U(\boldsymbol{v}, 0.05)$ are plotted against coverage $S(\boldsymbol{v})$, with $\hat{c}_v$ and the upper limit of its associated 95% confidence

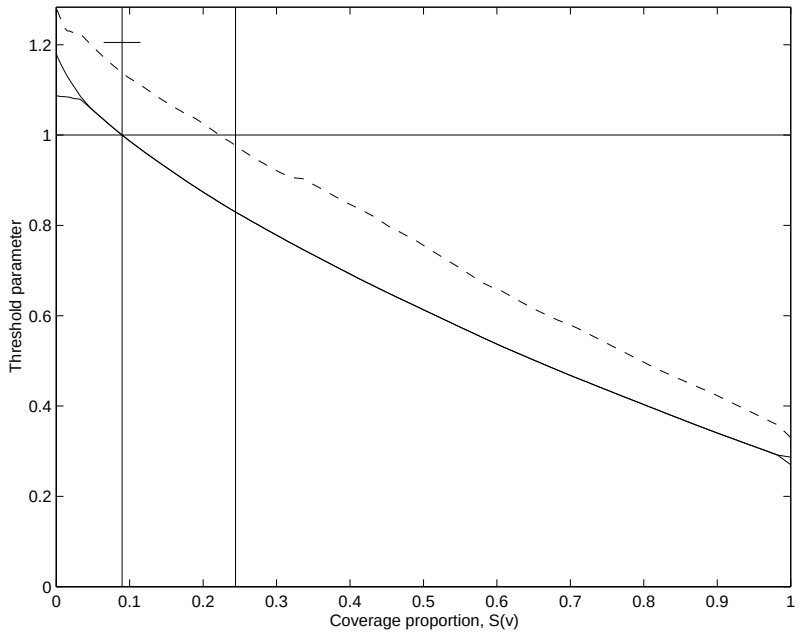

Figure 1: Reduction of threshold through optimal vaccination for two-type model, with a leaky vaccine of efficacy $\boldsymbol{\epsilon} = (0.7\ 0.7)$. The solid lines, which coincide for most values of $S(\boldsymbol{v})$, are $R_*^{(1)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$ and $R_*^{(2)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$ and the dashed line is $R_*^U(\boldsymbol{v}, 0.05)$; see text for further details. The vertical lines mark $\hat{c}_v$ and the upper limit of its associated 95% confidence interval. The horizontal bar marks the upper limit of the 95% confidence interval for $R(c, \boldsymbol{\theta})$ when $c = c_v$.

interval also marked. (The figures in this section are obtained by solving the linear programming problem "minimise $S(\boldsymbol{v})$ subject to $R_*^{(k)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}}) \leq r$ $(k = 1, 2, \ldots, J)$" for a grid of values for r. However, the confidence intervals are calculated assuming that the vaccination scheme $\boldsymbol{v}$ is fixed. For comparison, the confidence interval for $R(c, \boldsymbol{\theta})$ when $c = c_v$ is also shown in the figures. It would be numerically prohibitive to calculate the latter confidence interval for all values of c.) Note that for most values of $S(\boldsymbol{v})$ (except for $S(\boldsymbol{v})$ close to zero or one) the optimal strategy results in $R_*^{(1)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}}) = R_*^{(2)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$. The vaccination scheme that results in $R_*^{(1)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}}) = R_*^{(2)}(\boldsymbol{v}, \hat{\Lambda}^L, \hat{\boldsymbol{\pi}}) = 1$ concentrates most vaccination on low titre individuals residing in large households, but the full details of the scheme would take up too much space to present here.

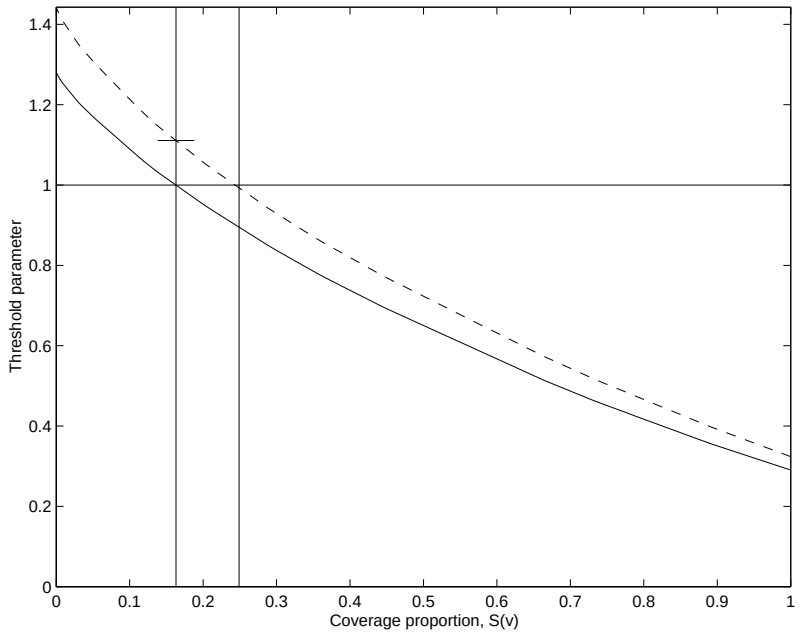

Figure 2: Reduction of threshold through optimal vaccination for two-type model, predicting on all low titre, with an all or nothing vaccine of efficacy $\epsilon_1 = 0.7$. Solid and dashed lines have the same meaning as in Figure 1.

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 $\hat{\Lambda}^L$, $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{z}}$ obtained above. The bounds $R_*^{(k)}(\hat{\Lambda}^L, \hat{\boldsymbol{\pi}})$ can then be calculated as 1.2810 and 0 for $k = 1, 2$, respectively. An upper bound for the threshold parameter is $R_*^U(\boldsymbol{v}, 0.05) = 1.4430$. Assuming an all or nothing vaccine with $\epsilon_1 = 0.7$, the secure vaccination coverage c_v is estimated to be 0.1631 with a 95% confidence interval of $(0, 0.2491)$ and is illustrated in figure 2. A leaky vaccine with $\epsilon_1 = 0.7$ requires coverage 0.1673 (95% confidence interval (0, 0.2544)). Note that again slightly more vaccine is required in the leaky case than in the all or nothing case. Also, the assumption that the population is all low titre has increased appreciably the estimate of c_v for both kinds of vaccine. Observe that $R_*^U(\boldsymbol{v}, 0.05)$ and the upper limit of the 95% confidence interval for $R(c, \boldsymbol{\theta})$ when $c = c_v$ coincide. This usually happens when, as in this instance, the population to be vaccinated is single type.

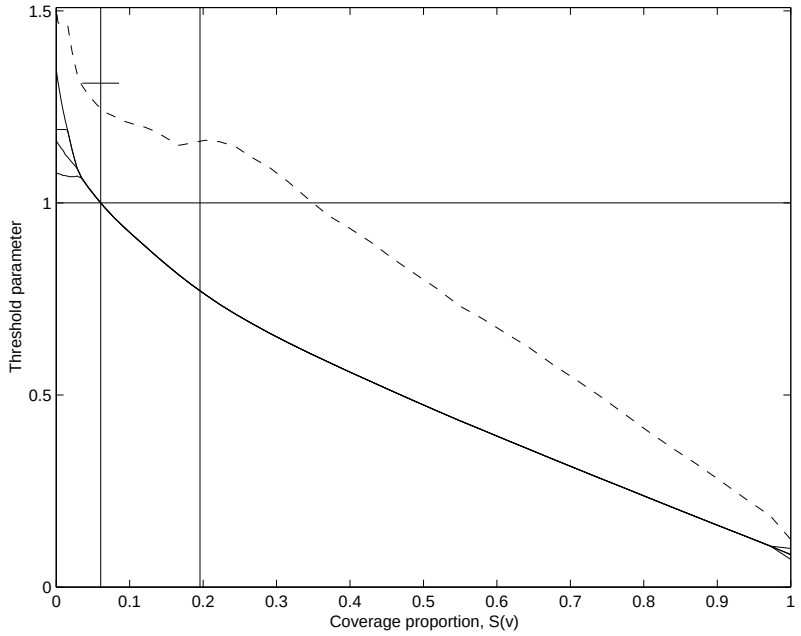

Figure 3: Reduction of threshold through optimal vaccination for four-type model, with an all or nothing vaccine of efficacy $\boldsymbol{\epsilon} = (0.9, 0.9, 0.9, 0.9)$. Solid and dashed lines have the same meaning as in Figure 1.

Consider now the case when individuals are also classified as adults or children. This gives 4 classes of individuals and estimates of $\hat{\Lambda}^L$, $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{z}}$ can be obtained. Figure 3 shows the reduction of threshold with an all or nothing vaccine with $\boldsymbol{\epsilon} = (0.9, 0.9, 0.9, 0.9)$, chosen as it illustrates the following points more clearly than $\boldsymbol{\epsilon} = (0.7, 0.7, 0.7, 0.7)$. 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 $R_*^U(\boldsymbol{v}, 0.05)$ to be below 1 (0.3495). Also note that $R_*^U(\boldsymbol{v})$ is not monotonic.

If all the individuals in a future population were in fact low titre and the vaccine was all or nothing with $\boldsymbol{\epsilon} = (0.7, 0.7, 0.7, 0.7)$, the required coverage is 0.2567 (95% confidence interval $(0, 0.3356)$) and is achieved entirely through vaccinating children. A leaky vaccine of the same efficacy would require coverage 0.2881 (95% confidence interval (0, 0.3995)) and is illustrated in figure 4. The leaky vaccine requires somewhat more vaccine than in the all or nothing case. Also note that the four-type model used here to predict on

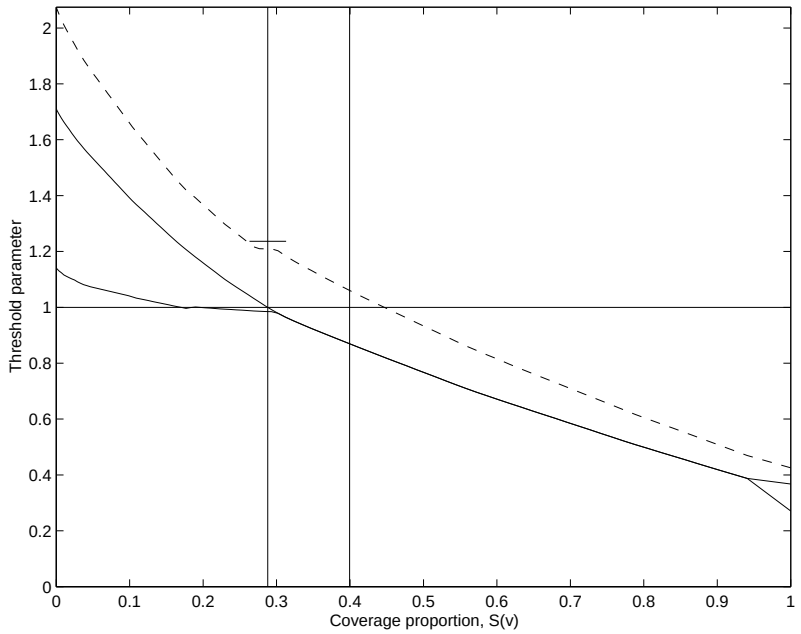

Figure 4: Reduction of threshold through optimal vaccination for four-type model, predicting on all low titre, with a leaky vaccine of efficacy $\boldsymbol{\epsilon} = (0.7, 0.7, 0.7, 0.7)$. Solid and dashed lines have the same meaning as in Figure 1.

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 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 [24].

The model used in the paper allows for heterogeneities owing to 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. Further, there are other social structures which surely affect the spread of disease, such as schools and workplaces, which clearly act as clusters where there is a higher contact rate between individuals than elsewhere. Nevertheless, it is believed that households, in combination with having different types of individual, capture the most important departures from homogeneity, so models admitting these two forms of heterogeneity should not be too far removed from reality. Needless to say, to capture all heterogeneities in a community into a mathematical model is impossible.

As noted already in Section 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, either expressed 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 (MCMC) methods; see O'Neill et al. [32] for an application of MCMC methods in a simpler epidemic setting.

Finally, although linear programming provides a means for computing optimal vaccination allocations, it would be 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 and Lyne [10] 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. [12], in which vaccines are allocated sequentially, always to a household that contains the greatest number of unvaccinated individuals. In the multitype case, the form of an optimal vaccination schemes does not admit such a simple characterisation and will be investigated in a separate paper (Ball et al. [8]), where it will also be proved that the leaky vaccine leads to less reduction in the spread of disease than the corresponding all or nothing vaccine.

Acknowledgements

This research was supported by the UK Engineering and Physical Sciences Research Council (EPSRC), under research grants GR/N09091 and GR/R08292, the Swedish Research Council and the Royal Swedish Academy of Sciences.

A Appendix

In the appendix, algorithms for calculating $\mu_{\boldsymbol{n},i,j}(\Lambda^L)$, $\mu_{\boldsymbol{n},i}(\Lambda^L,\boldsymbol{\pi})$ and $p_{\boldsymbol{n}}(\boldsymbol{t}|\Lambda^L,\boldsymbol{\pi})$ are presented. These are required to determine the threshold parameter R_*, the proportions of individuals of different classes infected by a global epidemic $\boldsymbol{z}$ and the estimate $(\hat{\Lambda}^L,\hat{\boldsymbol{\pi}})$, respectively.

For $\boldsymbol{n}=(n_1,n_2,\ldots,n_J)$ and $\boldsymbol{a}=(a_1,a_2,\ldots,a_J)$, let $E_{\boldsymbol{n},\boldsymbol{a}}(\Lambda^L,\boldsymbol{\pi})$ denote the multitype single household epidemic model studied by Addy et al. [1], in which initially there are a_i infectives and n_i susceptibles of class i ($i\in\mathcal{J}$), 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 $\phi_i(\theta)=E[\exp(-\theta T_I^{(i)})]$ ($\theta\geq 0$). Throughout its infectious period, a given class i infective contacts a given class j susceptible at the points of a homogeneous Poisson

process with rate λ^L_{ij}. For $i \in \mathcal{J}$, let $\tilde{S}_i$ denote the number of initial class i susceptibles that are uninfected at the end of the epidemic and let $\mu_{\boldsymbol{n},\boldsymbol{a},i}(\Lambda^L, \boldsymbol{\pi}) = E[\tilde{S}_i]$. A recursive expression for $\mu_{\boldsymbol{n},\boldsymbol{a},i}(\Lambda^L, \boldsymbol{\pi})$ is presented, from which expressions for $\mu_{\boldsymbol{n},i,j}(\Lambda^L)$ and $\mu_{\boldsymbol{n},i}(\Lambda^L, \boldsymbol{\pi})$ are easily obtained. Specifically, if $\mathbf{0}$ and $\mathbf{1}$ denote the J-dimensional row vectors consisting of all zeroes and all ones, respectively, and for $i \in \mathcal{J}$, $\boldsymbol{a}^{(i)}$ denotes the J-dimensional row vector whose ith element is one and all of whose other elements are zero, then $\mu_{\boldsymbol{n},i,j}(\Lambda^L) = n_j - \mu_{\boldsymbol{n}-\boldsymbol{a}^{(i)},\boldsymbol{a}^{(i)},j}(\Lambda^L, \mathbf{1})$ $(i, j \in \mathcal{J})$ and $\mu_{\boldsymbol{n},i}(\Lambda^L, \boldsymbol{\pi}) = n_i - \mu_{\boldsymbol{n},\mathbf{0},i}(\Lambda^L, \boldsymbol{\pi})$ $(i \in \mathcal{J})$.

Before describing the method of computing $\mu_{\boldsymbol{n},\boldsymbol{a},i}(\Lambda^L, \boldsymbol{\pi})$, some more notation is required. For J-vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, let $\boldsymbol{x}^{\boldsymbol{y}} = \prod_{i=1}^J x_i^{y_i}$. Let $\mathcal{N}_0' = \mathcal{N}_0 \cup \{\mathbf{0}\}$. For $\boldsymbol{i}, \boldsymbol{j} \in \mathcal{N}_0'$ with $\boldsymbol{i} \leq \boldsymbol{j}$ (inequalities between vectors are to be interpreted elementwise), let $\binom{\boldsymbol{j}}{\boldsymbol{i}} = \prod_{k=1}^J \binom{j_k}{i_k}$. For $\boldsymbol{n} \in \mathcal{N}_0'$, let $\sum_{\boldsymbol{k}=\mathbf{0}}^{\boldsymbol{n}} = \sum_{k_1=0}^{n_1} \sum_{k_2=0}^{n_2} \cdots \sum_{k_J=0}^{n_J}$. For $\boldsymbol{i} \in \mathcal{N}_0'$, let $\boldsymbol{h}(\boldsymbol{i}) = \big(h_1(\boldsymbol{i}), h_2(\boldsymbol{i}), \ldots, h_J(\boldsymbol{i})\big)$, where $h_j(\boldsymbol{i}) = \sum_{k=1}^J i_k \lambda^L_{jk}$. Finally, for $\boldsymbol{\theta} \in \mathbb{R}^J$, let $\boldsymbol{\phi}(\boldsymbol{\theta}) = \big(\phi_1(\theta_1), \phi_2(\theta_2), \ldots, \phi_J(\theta_J)\big)$.

An expression for the joint probability generating function of $\boldsymbol{S} = (S_1, S_2, \ldots, S_J)$ for the epidemic $E_{\boldsymbol{n},\boldsymbol{a}}(\Lambda^L, \mathbf{1})$ is given (in different notation) by Theorem 3.5 of Ball [7]. Appropriate differentiation of that expression shows that, for $i \in \mathcal{J}$,

$$\mu_{\boldsymbol{n},\boldsymbol{a},i}(\Lambda^L, \mathbf{1}) = \sum_{\boldsymbol{k}=\mathbf{0}}^{\boldsymbol{n}} \binom{\boldsymbol{n}}{\boldsymbol{k}} \alpha_{\boldsymbol{k}}^{(i)} \boldsymbol{\phi}\big(\boldsymbol{h}(\boldsymbol{k})\big)^{\boldsymbol{a}+\boldsymbol{n}-\boldsymbol{k}}, \tag{A.1}$$

where $\alpha_{\boldsymbol{k}}^{(i)}$ $(\boldsymbol{k} \geq \mathbf{0})$ are determined by

$$\sum_{\boldsymbol{k}=\mathbf{0}}^{\boldsymbol{n}} \binom{\boldsymbol{n}}{\boldsymbol{k}} \alpha_{\boldsymbol{k}}^{(i)} \boldsymbol{\phi}\big(\boldsymbol{h}(\boldsymbol{k})\big)^{\boldsymbol{n}-\boldsymbol{k}} = n_i \quad (\boldsymbol{n} \geq \mathbf{0}).$$

The distribution of the ultimate spread of the epidemic $E_{\boldsymbol{n},\boldsymbol{a}}(\Lambda^L, \boldsymbol{\pi})$ can be obtained by conditioning on the numbers of initial susceptibles of the J classes that avoid infection from outside the household, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_J)$ say, and considering the epidemic $E_{\boldsymbol{n}-\boldsymbol{Y},\boldsymbol{a}+\boldsymbol{Y}}(\Lambda^L, \mathbf{1})$ in which there is no outside infection. Hence, for $i \in \mathcal{J}$,

$$\mu_{\boldsymbol{n},\boldsymbol{a},i}(\Lambda^L, \boldsymbol{\pi}) = \sum_{\boldsymbol{l}=\mathbf{0}}^{\boldsymbol{n}} \binom{\boldsymbol{n}}{\boldsymbol{l}} \boldsymbol{\pi}^{\boldsymbol{l}} (\mathbf{1}-\boldsymbol{\pi})^{\boldsymbol{n}-\boldsymbol{l}} \mu_{\boldsymbol{n},\boldsymbol{a},i}(\Lambda^L, \mathbf{1}). \tag{A.2}$$

Substituting (A.1) into (A.2) and reversing the order of summation shows after a little algebra that

$$\mu_{\boldsymbol{n},\boldsymbol{a},i}(\Lambda^L,\boldsymbol{\pi}) = \sum_{\boldsymbol{k}=\boldsymbol{0}}^{\boldsymbol{n}} \binom{\boldsymbol{n}}{\boldsymbol{k}} \alpha_{\boldsymbol{k}}^i \boldsymbol{\phi}\big(\boldsymbol{h}(\boldsymbol{k})\big)^{\boldsymbol{a}+\boldsymbol{n}-\boldsymbol{k}} \boldsymbol{\pi}^{\boldsymbol{k}} \quad (i \in \mathcal{J}). \tag{A.3}$$

Recall from Section 3.1 that $p_{\boldsymbol{n}}(\boldsymbol{t}|\Lambda^L,\boldsymbol{\pi})$ $(\boldsymbol{0} \le \boldsymbol{t} \le \boldsymbol{n})$ is the total size distribution of the epidemic $E_{\boldsymbol{n},\boldsymbol{0}}(\Lambda^L,\boldsymbol{\pi})$. It follows, using Addy et al. [1], equation (4), that

$$\sum_{\boldsymbol{t}=\boldsymbol{0}}^{\boldsymbol{j}} \binom{\boldsymbol{n}-\boldsymbol{t}}{\boldsymbol{j}-\boldsymbol{t}} p_{\boldsymbol{n}}(\boldsymbol{t}|\Lambda^L,\boldsymbol{\pi}) \Big/ \left[\Big\{ \boldsymbol{\phi}\big(\boldsymbol{h}(\boldsymbol{n}-\boldsymbol{j})\big) \Big\}^{\boldsymbol{t}} \boldsymbol{\pi}^{\boldsymbol{n}-\boldsymbol{j}} \right] = \binom{\boldsymbol{n}}{\boldsymbol{j}} \quad (\boldsymbol{0} \le \boldsymbol{j} \le \boldsymbol{n}). \tag{A.4}$$

The triangular system of linear equations (A.4) determines $p_{\boldsymbol{n}}(\boldsymbol{t}|\Lambda^L,\boldsymbol{\pi})$ $(\boldsymbol{0} \le \boldsymbol{t} \le \boldsymbol{n})$.

References

[1] C. L. Addy, I. M. Longini and M. Haber, A generalized stochastic model for the analysis of infectious disease final size data, Biometrics 47 (1991) 961–974.

[2] P. K. Andersen, Ø. Borgan, R. D. Gill and N. Keiding, Statistical Models Based on Counting Processes, Springer-Verlag, New York, 1993.

[3] R. M. Anderson, R. M. May, Infectious diseases of humans: dynamic and control, Oxford, Oxford University press, 1991.

[4] H. Andersson, Epidemic models and social networks, Math. Scientist 24 (1999) 128–147.

[5] H. Andersson, T. Britton, Stochastic models and their statistical analysis. Springer Lecture Notes in Statistics 151. New York, Springer, 2000.

[6] N. T. J. Bailey, The Mathematical Theory of Infectious Diseases and its Applications, 2nd edition, Griffin, London, 1975.

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

[8] F. G. Ball, T. Britton and O. D. Lyne, Stochastic multitype epidemics in a community of households: form of optimal vaccination scheme, in preparation.

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

[10] F. G. Ball, O. D. Lyne, Optimal vaccination policies for stochastic epidemics among a population of households, Math. Biosci. 177–178 (2002) 333–354.

[11] F. G. Ball, O. D. Lyne, Statistical inference for epidemics among a population of households, in preparation.

[12] F. G. Ball, D. Mollison, G. Scalia-Tomba, Epidemics with two levels of mixing, Ann. Appl. Prob. 7 (1997) 46–89.

[13] N. G. Becker, Analysis of infectious disease data, London, Chapman and Hall, 1989.

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

[15] N. G. Becker, K. Dietz, The effect of the household distribution on transmission and control of highly infectious diseases, Math. Biosci. 127 (1995) 207–219.

[16] N. G. Becker, R. Hall, Immunization levels for preventing epidemics in a community of households made up of individuals of various types, Math. Biosci. 132 (1996) 205–216.

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

[18] N. G. Becker, D. N. Starczak, Optimal vaccination strategies for a community of households, Math. Biosci. 139 (1997) 117–132.

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

[20] T. Britton, Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage, J. R. Statist. Soc. B 63 (2001) 705–715.

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

[22] D. Greenhalgh, K. Dietz, Some bounds on estimates for reproduction ratio derived from age-specific force of infection, Math. Biosci. 124 (1994) 9–57.

[23] M. E. Halloran, M. Haber and I. M. Longini, Interpretation and estimation of vaccine efficacy under heterogeneity, Am. J. Epidemiol. 136 (1992) 328–343.

[24] M. E. Halloran, I. M. Longini, C. J. Struchiner, Design and interpretation of vaccine field studies. Epidemiologic Reviews 21 (1999), 73–88.

[25] J. A. P. Heesterbeek, K. Dietz, The concept of R_0 in epidemic theory, Statistica Neerlandica 50 (1996) 89–110.

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

[27] P. Jagers, Branching Processes with Biological Applications, Wiley, London, 1975.

[28] W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. Lond. A 115 (1927), 700–721

[29] D. Ludwig, Final size distributions for epidemics, Math. Biosci. 23 (1975) 33–46.

[30] C. J. Mode, Multitype Branching Processes, Elsevier, New York, 1971.

[31] A. S. Monto, J. S. Koopman, I. M. Longini, Tecumseh study of illness, XIII, Influenza infection and disease, 1976–1981, Am. J. Epidemiol. 121 (1985) 811–822.

[32] P. D. O'Neill, D. J. Balding, N. G. Becker, M. Eerola, D. Mollison, Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo Methods. Appl. Statist. 49 (2000), 517–542.

[33] G. Scalia-Tomba, Asymptotic final size distribution for some chain-binomial processes, Adv. Appl. Prob. 17 (1985) 477–495.

[34] G. Scalia-Tomba, On the asymptotic final size distribution of epidemics in heterogeneous populations, in J.-P. Gabriel, C. Lefèvre and P. Picard (Eds.) Stochastic Processes in Epidemic Theory, Lecture Notes in Biomathematics 86, Springer-Verlag, Berlin, 1990, 189–196.

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