J. R. Statist. Soc. B (2001)
63, *Part* 4, *pp.* 705–715

Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage

Tom Britton

Uppsala University, Sweden

[Received January 2000. Final revision May 2001]

Summary. A stochastic multitype model for the spread of an infectious disease in a community of heterogeneous individuals is analysed. In particular, estimates of R_0 (the basic reproduction number) and the critical vaccination coverage are derived, where estimation is based on final size data of an outbreak in the community. It is shown that these key parameters cannot be estimated consistently from data; only upper and lower bounds can be estimated. Confidence regions for the upper bounds are derived, thus giving conservative estimates of R_0 and the fractions necessary to vaccinate.

Keywords: Basic reproduction number; Consistency; Final size data; Multitype epidemic; Vaccination coverage; Vaccine efficacy

1. Introduction

The main practical motivation for the study of epidemic models lies in the insights that they provide about the control of infectious diseases. These insights attain practical relevance only when the model on which they are based captures the essential characteristics of disease transmission in a real community and the available data enable estimation of the model parameters. One feature that is known to play an important role in the propagation of infectious diseases is that of heterogeneities between individuals. For example, transmission rates for measles and rubella are found to depend substantially on the age of individuals (Grenfell and Anderson, 1985) and the rate of transmission for influenza type A is much higher within households than between (Addy *et al.*, 1991), as would be expected for all transmittable diseases.

In the present paper we treat estimation procedures of the basic reproduction number R_0, where inference is based on final size data from one outbreak in the community. The estimates are derived from stochastic models, thus allowing confidence bounds. The results are then interpreted in terms of vaccination policies: what are the necessary criteria for a vaccination policy to prevent future outbreaks, i.e. to be above the critical vaccination coverage? Such a community state is known as herd immunity since then everyone in the community is protected from future outbreaks, even those who are not vaccinated. The problems stated above are analysed for a so-called multitype epidemic model in which individuals are separated into different *types* with arbitrary transmission rates between each pair of types, i.e. with no restrictions on the ‘who acquires infection from whom’ matrix

Address for correspondence: Tom Britton, Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden.
E-mail: tom.britton@math.uu.se

1369–7412/01/63705

(Anderson and May, 1991). The types may for example reflect age groups, gender or the previous history of the disease and/or vaccination.

A special feature of the statistical inference is that the basic reproduction number R_0 and the critical vaccination coverage v_c cannot, in general, be estimated consistently. Instead estimates of the lower and upper bound for R_0 are given, bounds which also induce lower and upper bounds on v_c, and so only vaccination strategies with higher coverage than the upper bound will surely prevent future epidemics. The reason for this ambiguity is that the model contains more parameters (transmission rates) than the dimension of the observed data vector, thus not enabling an estimation of all the parameters (Anderson and May, 1984).

Greenhalgh and Dietz (1994) treated similar estimation problems for a deterministic model of an open population, i.e. with births and deaths, in which heterogeneity is caused by age. Sufficient data for estimation come from a cross-sectional survey from a population at 'equilibrium'. They derived expressions for upper and lower bounds of R_0 similar to those of the present paper, both under general transmission rates as well as for several submodels. They also considered different vaccination strategies and their effect on the equilibrium, and in particular whether the disease will become extinct. The present paper differs from Greenhalgh and Dietz (1994) in several ways. Its main merit is that the model is stochastic, thus giving confidence intervals for the estimates. Further we allow heterogeneities of other sorts than age, e.g. caused by previous history of vaccination or disease or gender. In Greenhalgh and Dietz (1994) such heterogeneities are not treated, with the effect that the problem of the optimal vaccination strategy is trivial: vaccinate only in the youngest age group(s). A drawback with the present analysis, compared with Greenhalgh and Dietz (1994), is the assumption of a closed population, with the effect that individuals cannot change type over time as is natural with age cohorts observed over longer periods of time. Of course no population is really closed. However, when considering a short epidemic outbreak, perhaps lasting a few months, the community may be approximated as being closed. The methods of the present paper are not suitable for long-term outbreaks or a simultaneous analysis of several different outbreaks. The reason for not treating a stochastic epidemic model for an open population is the complicated quasi-stationary behaviour of such models; see Nåsell (1999). The estimators for open populations are usually the same as in a closed population but the standard errors are different. Farrington *et al.* (2001) also treat a deterministic model for an open population allowing for various heterogeneities. By using available contact parameters from other related disease outbreaks, assuming some relationship between the contact rates for the diseases, they could estimate R_0 consistently. See Section 2 in Greenhalgh and Dietz (1994) for an excellent survey of related work in the analysis of epidemics. A short note treating problems similar to those of the present paper, but in a deterministic framework, has appeared recently (Britton, 1998a).

In Section 2 we define the multitype epidemic model and present asymptotic results for it. In Section 3 we derive estimates, including confidence bounds, of the fundamental parameter R_0. In Section 4 we use these results to construct vaccination programmes that prevent future outbreaks. Section 5 illustrates the results with an example.

2. The model

2.1. Definition

The model that we now define is a stochastic susceptible–infected–removed epidemic model for a closed multitype population (e.g. Ball and Clancy (1993)). Consider a closed population of size n consisting of k different types of individuals, labelled 1, . . ., k, and let n_i denote the number of i-individuals and $\pi_i = n_i/n$ the corresponding proportion. If an i-individual becomes infected he or she becomes infectious, possibly after a latency period with arbitrary distribution. During the infectious period an i-individual has 'close contact' with any given j-individual at rate β_{ij}/n, where a close contact is defined as a contact which results in infection if the other individual is susceptible; otherwise the contact has no effect. The matrix $\{\beta_{ij}\}$ of contact intensities is assumed to be irreducible, thus omitting the possibility of a major outbreak for some but not all types of individual. The infectious period I_i has distribution F_i with mean μ_i and standard deviation σ_i. For future use we define $\lambda_{ij} = \mu_i\beta_{ij}$, implying that $\lambda_{ij}\pi_j$ denotes the expected number of close contacts which an i-individual has with j-individuals during the infectious period. When the infectious period is over, the individual recovers and becomes immune, and we say that the individual is removed. The epidemic evolves until there are no infectious individuals in the population. Then no-one can become infected and the epidemic has entered its *final state*. All contact processes and infectious periods are defined to be mutually independent.

As pointed out by Ball and Clancy (1993) the model can be generalized without affecting the distribution of the final state. The final state depends only on the distribution of the 'total infection forces' $\{n^{-1}\beta_{ij}I_i\}$ for the different type combinations. Instead of assuming constant contact rates over the infectious period we may allow for a time-varying infectivity, including an initial latency period. This is modelled by a stochastic process $\{I_i(t);\ t \geqslant 0\}$, where $I_i(t)$ is the infectivity t time units after infection of an i-individual and it falls under the model defined above simply by letting F_i denote the distribution of $\int_0^\infty I_i(t)\,\mathrm{d}t$.

2.2. Asymptotic properties of the model

The asymptotic properties of the model above, for a large population, have been analysed extensively by Ball and Clancy (1993). Starting with few initially infectious individuals in a large, otherwise susceptible, population, the epidemic can either take off and give a large outbreak or it may die out and infect very few, a general phenomenon for epidemic models. During the initial stages the epidemic can be approximated by a multitype branching process because infectious individuals infect new individuals virtually independently of each other since the probability that they will contact the same individual is negligible. For the model of the present paper the fundamental parameter R_0, the basic reproduction number, is defined as the largest positive eigenvalue of the matrix $(\lambda_{ij}\pi_j)$. Note that $\lambda_{ij}\pi_j = (\lambda_{ij}/n)n_j$ is the expected number of close contacts which an infectious i-individual has with j-individuals during the infectious period. In the branching process $(\lambda_{ij}\pi_j)$ thus corresponds to the matrix of mean offspring distribution. The approximating branching process is subcritical, critical or supercritical depending on whether R_0 is smaller than, equal to or larger than 1. It hence follows that, asymptotically, the probability of a large outbreak in a completely susceptible population is positive if and only if $R_0 > 1$ (e.g. Ball and Clancy (1993)).

If a proportion $1 - s_j$ of all j-individuals are initially immune, so the proportion s_j are susceptible, then the *effective* reproduction number R_e is the largest positive eigenvalue of the matrix $(\lambda_{ij}\pi_j s_j)$ and a major outbreak is possible if and only if $R_\mathrm{e} > 1$. A simple argument for the last result is that we may neglect the immune individuals by introducing new notation: $n'_i = n_i s_i$, the number of susceptible i-individuals, $n' = \Sigma_i n'_i$, the total number of susceptible individuals, and $\pi'_i = n'_i/n' = \pi_i s_i/s$, the proportion of susceptible individuals that are of type i (where $s = \Sigma_i s_i \pi_i$ is the overall proportion susceptible). The contact parameter is unchanged and equals $\beta_{ij}/n = s\beta_{ij}/n'$, so by introducing $\beta'_{ij} = s\beta_{ij}$ and similarly $\lambda'_{ij} = s\lambda_{ij}$ it follows from the result above that the effective reproduction number R_e is the largest positive eigenvalue of the matrix $(\lambda'_{ij}\pi'_j) = (\lambda_{ij}\pi_j s_j)$.

When making inferences we shall always assume that a major outbreak has occurred and that the initial number of infective individuals is small. The results are thus conditional on a major outbreak — otherwise there is not enough information for consistent estimation — which implicitly assumes that $R_e > 1$ from the properties stated above.

Consider the model defined above in a population with type distribution $\{\pi_i\}$ and initial proportions susceptible given by $\{s_i\}$. Let $\tilde{p}_i$ denote the random proportion among the initially susceptible i-individuals who become infected during the course of the epidemic. Applying the results in Ball and Clancy (1993) then shows, assuming few initial infective individuals and a major outbreak, that the vector $\{\tilde{p}_i\}$ converges in probability to $\{p_i\}$ as $n \to \infty$, where $\{p_i\}$ is the unique positive solution to the system of equations

$$1 - p_j = \exp\Big(- \sum_i \pi_i s_i p_i \lambda_{ij}\Big), \qquad j = 1, \ldots, k. \tag{1}$$

Equations (1) have a natural interpretation: the proportion that escape infection equals the probability of escaping infection from the aggregated total infection forces. A central limit theorem in Ball and Clancy (1993) shows that the vector $\{\surd(n_j s_j)(\tilde{p}_j - p_j)\}$ is asymptotically Gaussian with mean vector $\mathbf{0}$ and variance matrix

$$\Sigma = S^{\mathrm{T}^{-1}} \Xi S^{-1},$$

where the matrices S and Ξ have elements

$$S_{ij} = \delta_{ij} - \surd(\pi_i s_i \pi_j s_j)\lambda_{ij}(1 - p_j),$$
$$\Xi_{ij} = p_i(1 - p_j)\delta_{ij} + \surd(\pi_i s_i \pi_j s_j)(1 - p_i)(1 - p_j) \sum_k \pi_k s_k p_k \lambda_{ki} \lambda_{kj} (\sigma_k/\mu_k)^2,$$

where δ_{ij} denotes the Kronecker delta function ($\delta_{ii} = 1$ and $\delta_{ij} = 0$, $i \neq j$).

2.3. Modelling vaccination

Suppose that a vaccine is available having efficacy r_j among j-individuals, $j = 1, \ldots, k$. It could for example be that all infection rates λ_{ij}, $i = 1, \ldots, k$, are reduced by a factor r_j (the so-called leaky effect), or that a proportion r_j become completely immune and the rest are unaffected by the vaccine (the all-or-nothing effect). The case $r_j = 1$ for all j corresponds to a perfect vaccine and $r_j = 0$ for all j to a useless vaccine. See for example Halloran *et al.* (1992) for more about vaccine efficacy.

The propagation of disease transmission in a partly vaccinated community also having initially immune individuals can be described using the present model. Consider the same population as before having proportions initially susceptible given by $\{s_i\}$ and suppose that a proportion v_j of all initially susceptible j-individuals are vaccinated with such a vaccine before the epidemic season. The expected number of close contacts that an infectious i-individual has with initially susceptible j-individuals is then reduced from $\lambda_{ij}\pi_j s_j$ to

$$\lambda_{ij}\pi_j s_j\{(1 - v_j) + v_j(1 - r_j)\} = \lambda_{ij}\pi_j s_j(1 - v_j r_j).$$

This is true because a proportion v_j among the j-individuals have reduced their susceptibility by a factor r_j. The effective reproduction number after vaccination, R_{ev}, is then the largest eigenvalue of the matrix $(\lambda_{ij}\pi_j s_j(1 - v_j r_j))$ and the vaccinations performed will surely prevent an outbreak in the population if $R_{\mathrm{ev}} \leqslant 1$. Vaccinating with the effect of surely preventing an outbreak when the entire population is initially susceptible is of main interest, because this vaccination programme will be preventive whatever proportion is immune and will remain so if the disease-acquired immunity wanes with time. In what follows we shall thus focus on studying vaccination programmes for which $R_{\mathrm{v}} \leqslant 1$, where R_{v} is the largest eigenvalue of the matrix $(\lambda_{ij}\pi_j(1 - v_j r_j))$. Vaccination programmes aiming at reducing R_{ev} below 1, for some other specified susceptibility levels $\{s_i\}$, can be derived by using identical arguments.

3. Estimation of R_0

In this section we derive estimates of R_0 for the multitype epidemic defined in the previous section. Remember that R_0 is the largest eigenvalue of the matrix $(\lambda_{ij}\pi_j)$, i.e. for a completely susceptible population. The parameters are assumed to be unknown and are estimated with data from one epidemic outbreak $(\tilde{p}_1, \ldots, \tilde{p}_k)$ which may have occurred in a community containing initially immune individuals. The proportions immune before the outbreak $\{1 - s_i\}$, the community structure $\{\pi_i\}$ and the community size n are assumed to be known. It is important to take into account the presence of initially immune individuals when making inference on the reproduction number from an epidemic outbreak. If this is neglected the resulting estimates will underestimate the true parameters, with the effect that the suggested vaccination coverage may not be preventive. This is very different from assuming that the community to be vaccinated is completely susceptible, as assumed in the previous section. In the latter case the suggested proportions to vaccinate are preventive even when the assumption fails.

As mentioned in the previous section the vector $(\tilde{p}_1, \ldots, \tilde{p}_k)$ converges in probability to $(p_1, \ldots, p_k)$ defined in equations (1) as $n \to \infty$. We hence start by treating the deterministic limit before dealing with uncertainty.

3.1. Deterministic limit

With given vectors $\{p_j\}$ and $\{s_j\}$ satisfying equations (1) but $\{\lambda_{ij}\}$ otherwise arbitrary, R_0 is not determined uniquely, as has been observed previously (e.g. Greenhalgh and Dietz (1994) and Britton (1998b)). In fact R_0 can attain any value in an interval, as the following lemma shows.

Lemma 1. Let $\{p_j\}$, $\{s_j\}$ and $\{\pi_j\}$ be defined as above and let $\{\tau_j\}$ be any given vector with positive elements. Then the largest positive eigenvalue of the matrix $(\lambda_{ij}\pi_j\tau_j)$, where $\{\lambda_{ij}\}$ satisfies equations (1), lies in the closed interval $[\rho^{\min}, \rho^{\max}]$, where

$$\rho^{\min} = \min_i\{-\tau_i \log(1 - p_i)/s_i p_i\},$$
$$\rho^{\max} = \max_i\{-\tau_i \log(1 - p_i)/s_i p_i\}.$$

All values in the interval can be attained.

Remark 1. If lemma 1 is applied with $\tau_i = 1$, $i = 1, \ldots, k$, the largest eigenvalue specifies R_0. It hence follows that R_0 lies between

$$R_0^{\min} = \min_i\{-\log(1 - p_i)/s_i p_i\}$$

and

$$R_0^{\max} = \max_i\{-\log(1-p_i)/s_i p_i\}.$$

Proof. Denote the largest eigenvalue of the matrix $(\lambda_{ij}\pi_j\tau_j)$ by ρ. By the Perron–Frobenius theorem it follows that there is a vector $\{x_i\}$ with positive components, unique up to normalization, such that

$$L(j) := \rho x_j = \sum_i x_i \lambda_{i,j} \pi_j \tau_j =: R(j) \qquad j = 1, \ldots, k \tag{2}$$

(e.g. Jagers (1975), pages 92–93). Define $M = \max_{1 \leqslant j \leqslant k}(x_j/\pi_j s_j p_j)$, and suppose that the maximum is attained for $j = j_0$, i.e. $x_{j_0}/\pi_{j_0} s_{j_0} p_{j_0} = M$. Then, by the definition of the left-hand side in equation (2) we have $L(j_0) = \rho M \pi_{j_0} s_{j_0} p_{j_0}$. The right-hand side can be dominated as follows:

$$R(j_0) = \sum_i \frac{x_i}{\pi_i s_i p_i} \pi_i s_i p_i \lambda_{i,j_0} \pi_{j_0} \tau_{j_0} \leqslant M \sum_i \pi_i s_i p_i \lambda_{i,j_0} \pi_{j_0} \tau_{j_0} = M \pi_{j_0} \tau_{j_0} \{-\log(1-p_{j_0})\},$$

where the last equality is equation (1). Since $\rho M \pi_{j_0} s_{j_0} p_{j_0} = L(j_0) = R(j_0)$ this gives the upper bound $\rho \leqslant -\tau_{j_0} \log(1-p_{j_0})/s_{j_0} p_{j_0} \leqslant \max_i\{-\log(1-p_i)/s_i p_i\} = \rho^{\max}$. An identical argument shows that $\rho \geqslant \rho^{\min}$. Below are some observations showing that the end points of the interval can be obtained (when $\tau_i = 1$). Finally, any point in the interval can for example be obtained by a linear combination of the two extremes. □

In the restricted parameter space known as separable mixing (e.g. Hethcote and Van Ark (1987)), i.e. $\lambda_{ij} = \alpha_i \beta_j$, the parameters may be interpreted as infectivity and susceptibility respectively. (Sometimes this is called proportional mixing, but most often proportional mixing is used for the stronger assumption that $\lambda_{ij} = \alpha_i \alpha_j$.) Then the basic reproduction number has the explicit expression

$$R_0 = \sum_i \alpha_i \beta_i \pi_i \tag{3}$$

(e.g. Becker and Marschner (1990)). This can be used to verify the following observations.

(a) $R_0^{\max}$, the 'worst scenario', is attained in the separable mixing case if $\beta_i = -\log(1-p_i)$, $i = 1, \ldots, k$, and $\alpha_i = 0$ for all i except for the type i_0 maximizing $-\log(1-p_i)/s_i p_i$, for which $\alpha_{i_0} = 1/\pi_{i_0} s_{i_0} p_{i_0}$. This choice gives $R_0^{\max}$ when inserted in equation (3) and condition (1) is also satisfied.

(b) $R_0^{\min}$, the 'best scenario', is attained in the separable mixing case if $\beta_i = -\log(1-p_i)$, $i = 1, \ldots, k$, $\alpha_{i_1} = 1/\pi_{i_1} s_{i_1} p_{i_1}$ for the type i_1 minimizing $-\log(1-p_i)/s_i p_i$, and $\alpha_i = 0$ for all other is.

(c) Under the assumption of equal infectivity ($\lambda_{ij} = \alpha\beta_j$) equations (1) determine the parameters uniquely, $\alpha\beta_j = -\log(1-p_j)/\Sigma_i \pi_i s_i p_i$, so R_0 is completely specified and equals $\Sigma_j \pi_j\{-\log(1-p_j)\}/\Sigma_i \pi_i s_i p_i$.

3.2. *Stochastic model*

In the previous subsection bounds on the basic reproduction number were derived for the deterministic limit of the multitype epidemic. In a finite community the observed proportions infected $\{\tilde{p}_i\}$ are random, so the corresponding quantities $\tilde{R}_0^{\max}$ and $\tilde{R}_0^{\min}$, where p_i is replaced by $\tilde{p}_i$, are random estimates of these quantities. We now derive a one-sided confidence interval for $R_0^{\max}$ which will be used in the next section when deriving how many need to be vaccinated to obtain herd immunity with some given certainty.

As noted in the previous subsection R_0 was maximized when one type caused all infections, the type i_0 which maximizes $-\log(1-p_i)/s_ip_i$. (When all types have equal proportions initially susceptible, e.g. $s_i = 1$ for all i, this is the type with highest proportion infected p_i.) We therefore assume this to hold when constructing confidence bounds and thus overcome the fact that all parameters are not identifiable. The variance of an estimated R_0 may of course be larger for other parameter configurations, but these configurations all have a limiting R_0 smaller than $R_0^{\max}$, so then our confidence bound for $R_0^{\max}$ will still be conservative. In Section 2.2 the asymptotic variance matrix of $(\surd(n_is_i)(\tilde{p}_i - p_i))$, denoted Σ, was given. For the case when $\lambda_{ij} = 0$, $i \neq i_0$, $\Sigma_{i_0i_0}$ is explicit and equals

$$\begin{aligned}\Sigma_{i_0i_0} &= \frac{p_{i_0}(1-p_{i_0})\{1+(\pi_{i_0}s_{i_0}\lambda_{i_0i_0})^2(1-p_{i_0})(\sigma_{i_0}/\mu_{i_0})^2\}}{\{1-\pi_{i_0}s_{i_0}\lambda_{i_0i_0}(1-p_{i_0})\}^2}\\ &= \frac{p_{i_0}(1-p_{i_0})[p_{i_0}^2+\{\log(1-p_{i_0})\}^2(1-p_{i_0})(\sigma_{i_0}/\mu_{i_0})^2]}{[p_{i_0}+(1-p_{i_0})\log(1-p_{i_0})]^2}. \end{aligned} \tag{4}$$

The second equality follows from the assumption that $\lambda_{ij} = 0$, $i \neq i_0$, implying that $\pi_{i_0}s_{i_0}\lambda_{i_0i_0} = -\log(1-p_{i_0})/p_i$ for condition (1) to hold. The asymptotic variance of $\tilde{p}_{i_0}$ is $\Sigma_{i_0i_0}/n_{i_0}s_{i_0}$. If we perform this substitution and replace p_{i_0} by $\tilde{p}_{i_0}$ we obtain an explicit standard error

$$\mathrm{se}(\tilde{p}_{i_0}) = \frac{\surd(\tilde{p}_{i_0}(1-\tilde{p}_{i_0})[p_{i_0}^2+\{-\log(1-\tilde{p}_{i_0})\}^2(1-\tilde{p}_{i_0})(\sigma_{i_0}/\mu_{i_0})^2])}{\surd(n_{i_0}s_{i_0})\{\tilde{p}_{i_0}+(1-\tilde{p}_{i_0})\log(1-\tilde{p}_{i_0})\}}. \tag{5}$$

The quantity σ_{i_0}/μ_{i_0} appearing in equation (5) denotes the coefficient of variation of the length of the infectious period for type i_0. This quantity must be known or else estimated using prior information; final size data carry no information about any temporal quantities. Our estimate $\hat{R}_0^{\max} = -\log(1-\tilde{p}_{i_0})/s_{i_0}\tilde{p}_{i_0}$ is increasing in $\tilde{p}_{i_0}$. Replacing $\tilde{p}_{i_0}$ by an upper confidence limit will thus produce an upper confidence limit for $R_0^{\max}$. We summarize our results in the following theorem.

Theorem 1. Let i_0 be defined as the index maximizing $-\log(1-\tilde{p}_i)/s_i\tilde{p}_i$, assumed to be asymptotically unique. Then, for the multitype epidemic model,

$$\hat{R}_0^{\max} = -\log(1-\tilde{p}_{i_0})/s_{i_0}\tilde{p}_{i_0} = \max_i\{-\log(1-\tilde{p}_i)/s_i\tilde{p}_i\} \tag{6}$$

is a consistent and asymptotically Gaussian estimator for $R_0^{\max}$ (defined in lemma 1). The asymptotic variance of the estimator is

$$\mathrm{var}(\hat{R}_0^{\max}) = \frac{p_{i_0}^2+\{-\log(1-p_{i_0})\}^2(1-p_{i_0})(\sigma_{i_0}/\mu_{i_0})^2}{n_{i_0}s_{i_0}^3p_{i_0}^3(1-p_{i_0})}. \tag{7}$$

A $1-\alpha$ upper confidence bound for $R_0^{\max}$ is given by

$$\max_i\{-\log(1-\tilde{p}_i^+)/s_i\tilde{p}_i^+\}, \tag{8}$$

where $\tilde{p}_i^+ = \tilde{p}_i + z_\alpha\,\mathrm{se}(\tilde{p}_i)$. Here $\mathrm{se}(\tilde{p}_i)$ is defined as in equation (5) only replacing i_0 by i, and z_α is the $(1-\alpha)$-quantile in the normal distribution.

Remark 2. If uniqueness is not assumed the estimator is still consistent. However, then the variance is not correct as, in the limit, the index i_0 varies. The confidence bound given by expression (8) may be used as a confidence bound for R_0 for any set of underlying parameters and is then conservative.

Proof. Consistency and asymptotic normality are a direct consequence of the asymptotic results stated in Section 2.2, together with the assumption of asymptotic uniqueness of i_0. The variance formula is obtained by the delta method on $\hat{R}_0^{\max} = f(\tilde{p}_{i_0})$ viewed as a function of $\tilde{p}_{i_0}$. It follows that

$$\operatorname{var}(\hat{R}_0^{\max}) = f'(\tilde{p}_{i_0})^2 \operatorname{var}(\tilde{p}_{i_0})$$

plus terms of smaller order. It follows from simple algebra that this equals equation (7). The standard error for $\tilde{p}_i$ is a consistent estimate for the standard deviation since the observed quantities converge to the limits defined by equations (1). The asymptotic normality then implies that the upper confidence bound defined by expression (8) is correct.

4. Control

We now return to controlling the spread of disease by means of vaccination as discussed in Section 2.3, only now the model parameters are assumed unknown and are estimated by using methods presented in the previous section.

4.1. Deterministic limit

Suppose that a vaccination programme, for which the vaccine has known efficacy $\{r_j\}$, is to be carried out. The contact matrix (λ_{ij}) is known to satisfy condition (1). As mentioned previously the necessary vaccination levels will be derived assuming a completely susceptible community, this being a conservative assumption. If a proportion v_i of the i-individuals are vaccinated, $i = 1, \ldots, k$, then the resulting reproduction number R_v is given by the largest eigenvalue of the matrix $(\lambda_{ij}\pi_j(1 - v_j r_j))$; see Section 2.3. Applying lemma 1 with $\tau_j = 1 - v_j r_j$ then shows that R_v is contained in the interval

$$\left[\min_{1 \leqslant i \leqslant k} \left\{ (1 - r_i v_i) \frac{-\log(1 - p_i)}{s_i p_i} \right\}, \max_{1 \leqslant i \leqslant k} \left\{ (1 - r_i v_i) \frac{-\log(1 - p_i)}{s_i p_i} \right\} \right].$$

Herd immunity is obtained if $R_\text{v} \leqslant 1$. This is surely the case only when the upper end of the interval does not exceed 1, or equivalently $v_i \geqslant r_i^{-1}\{1 + s_i p_i / \log(1 - p_i)\}$, for each i. The optimal vaccination strategy for the multitype epidemic, meaning the vaccination programme vaccinating the smallest number of individuals among all vaccination strategies that surely prevent future outbreaks, is thus given by

$$v_i = \frac{1}{r_i} \left\{ 1 - \frac{s_i p_i}{-\log(1 - p_i)} \right\}, \qquad i = 1, \ldots, k. \tag{9}$$

These proportions will surely prevent future outbreaks. Each estimate is conservative in that the proportion is derived under the assumption that all infectivity comes from that specific type. Unless the vaccine is perfect, i.e. $r_i = 1$ for all i, it may happen that $v_i > 1$ for some j. This implies that the community is not surely protected from future outbreaks even when every such individual is vaccinated, i.e. the vaccine is not sufficiently effective to obtain herd immunity.

4.2. Uncertainty in estimates

The vaccination coverages for different types are estimated from equation (9) simply by replacing the limits $\{p_i\}$ by the observed proportions infected $\{\tilde{p}_i\}$. As mentioned above, v_i, and hence also its estimate $\hat{v}_i$, was obtained assuming that *i*-individuals were responsible for all infections. The uncertainty of the estimate should thus be obtained under this assumption. Using arguments that are identical with those for the variance of $\tilde{p}_{i_0}$ presented in Section 3.2 it follows that the asymptotic variance of $\tilde{p}_i$ assuming $\lambda_{kj} = 0$, $k \neq i$, equals $\Sigma_{ii}/n_i s_i$ where Σ_{ii} is defined as in equation (4) only replacing i_0 by i. Since $\tilde{v}_i$ is increasing in $\tilde{p}_i$ upper confidence bounds for $\hat{v}_i$ can be obtained by replacing the estimate $\tilde{p}_i$ by the upper confidence bound $\tilde{p}_i^+$ defined in theorem 1. We summarize the results in the following theorem.

Theorem 2. The estimates defined by

$$\hat{v}_i = \frac{1}{r_i}\left\{1 - \frac{s_i\tilde{p}_i}{-\log(1-\tilde{p}_i)}\right\}, \qquad i = 1, \ldots, k, \tag{10}$$

are consistent and asymptotically Gaussian estimates of the critical vaccination coverage of the multitype epidemic defined by equation (9). The asymptotic variance for $\hat{v}_i$ is

$$\operatorname{var}(\hat{v}_i) = s_i^2 \frac{p_i^2 + \log(1-p_i)^2(1-p_i)(\sigma_i/\mu_i)^2}{r_i^2 n_i s_i p_i (1-p_{i_0})\{\log(1-p_i)\}^4}, \tag{11}$$

and a $1-\alpha$ upper confidence bound for v_i is given by

$$\hat{v}_i^+ = \frac{1}{r_i}\left\{1 - \frac{s_i\tilde{p}_i^+}{-\log(1-\tilde{p}_i^+)}\right\}, \tag{12}$$

where $\tilde{p}_i^+$ is defined in theorem 1.

Proof. Consistency and asymptotic normality follow by using arguments that are similar to those in the proof of theorem 1. The variance expression (11) is obtained by using the delta method and the upper confidence bound (12) is derived simply by inserting an upper confidence estimate $\tilde{p}_i^+$ for the unknown quantity p_i. □

The vaccination coverages defined above are preventive when the community is completely susceptible. Note that the epidemic outbreak on which inference is based could still have contained initially immune individuals. If the community is not completely susceptible the vaccination levels may be lower. In case the susceptible proportions are known and equal to $\{s_i'\}$ say, then the vaccination programme should vaccinate enough individuals such that R_{ev} does not exceed 1. Recall that R_{ev}, the effective reproduction number after vaccination, was defined as the largest eigenvalue of the matrix $(\lambda_{ij}\pi_j s_j'(1-v_j r_j))$. Lemma 1 can be applied, with a different choice of $\{\tau_j\}$, to solve this problem using the same methods as above. In fact, the community structure may also be altered from that of the epidemic outbreak, to $\{\pi_j'\}$ say. We could thus estimate vaccination programmes for a community that is different from the community on which the outbreak inference is based. However, these estimates will only be valid if the contact parameters $\{\lambda_{ij}\}$ are the same for the two communities.

5. An example

A simple example with two types of individual illustrates the methods in the paper. Suppose that a community consists of 1000 individuals, 300 children and 700 adults. Before the epidemic individuals are tested for antibodies and it is found that 90% of the children and 60% of the adults were susceptible. An epidemic outbreak then occurs, resulting in 80% of the susceptible children and 20% of the susceptible adults becoming infected. Further we assume that the coefficient of variation of the length of the infectious period is the same for both types and equals 3/7, e.g. 7 days on average and 3 days standard deviation.

With the terminology of the paper we have (children being type 1) $n_1 = 300$, $s_1 = 0.9$, $\tilde{p}_1 = 0.8$, $n_2 = 700$, $s_2 = 0.6$, $\tilde{p}_2 = 0.2$ and $\sigma_1/\mu_1 = \sigma_2/\mu_2 = 3/7$. This implies that $-\log(1-\tilde{p}_1)/s_1\tilde{p}_1 = 2.235$ and $-\log(1-\tilde{p}_2)/s_2\tilde{p}_2 = 1.859$. From equation (6) it hence follows that $\hat{R}_0^{\max} = 2.235$ which is estimated for the worst case where children cause all infections. Using equation (5) the standard errors for $\tilde{p}_1$ and $\tilde{p}_2$ are 0.044 and 0.198 respectively. A 95% upper confidence bound for $R_0^{\max}$ equals 2.618, computed using expression (8).

Suppose now that a vaccine having 90% efficacy, the same for both types so $r_1 = r_2 = 0.9$, is available. The necessary proportion to vaccinate to avoid future epidemics depends on the proportion of individuals who are immune to the disease. For example, directly after the epidemic the community is protected without any vaccination. However, as time passes the disease-acquired immunity usually wanes and an increasing proportion must be vaccinated to obtain herd immunity. The necessary proportions to vaccinate when the entire community is susceptible, estimated from equation (10), are $\hat{v}_1 = 0.614$ and $\hat{v}_2 = 0.514$. Taking uncertainty into account by giving 95% upper confidence bounds changes the estimates to $\hat{v}_1^+ = 0.687$ and $\hat{v}_2^+ = 0.641$ using equation (12). We conclude that at least 69% of the children and 64% of the adults should be vaccinated to prevent future outbreaks. This level of vaccination will keep the community protected even when the entire community is initially susceptible, but also if some individuals are immune. The same levels also apply to any other community having the same community and contact structure.

6. Discussion

In the present paper it is shown that fundamental parameters, such as R_0 and the critical vaccination coverage, cannot be estimated consistently from final size data in a simple multitype epidemic model. It is shown that a range of parameter configurations (having different R_0!) are consistent with data. The largest range of possible values of R_0 appears when the proportion infected among different types varies greatly. In particular it cannot be determined who causes further infections. However, by estimating 'the worst case' it is still possible to derive vaccination programmes which surely prevent future outbreaks. The suggested vaccination coverage consists of vaccinating the proportion of a type such that the whole community is 'safe', herd immune, even if this type causes all further infections. The paper is meant to serve as an example showing that consistent estimation is often not possible even in simple epidemic models. If temporal data from an outbreak are available, a topic which is not treated in the present paper, all the parameters are often, but not always, identifiable (Britton, 1998b).

To derive expressions for the critical vaccination coverage is not only of academic interest. Of course, health practitioners aim for complete vaccination coverage but this is hardly ever possible to achieve. For this reason expressions for the critical vaccination coverage may serve as the lowest acceptable coverage. The results of the present paper also show that it is not enough to vaccinate only in the most susceptible subgroups unless prior information about infectivity among subgroups is available: all groups must be partly vaccinated for the community to have herd immunity surely.

The model is still some way from being realistic in that it does not allow mixing at different levels (see Ball *et al.* (1997)) which is natural when social structures are present. If both individual heterogeneities and social structures such as households are acknowledged, then estimation quickly becomes cumbersome (e.g. Addy *et al.* (1991) and Britton and Becker (2000)). A thorough study for such models remains to be performed.

In real life the proportions of various types, the proportions immune and the proportions infected during an epidemic are not fully known, only estimates thereof. This means that only *estimates* of $\{\pi_i, s_i\}$ and the random quantities $\{\tilde{p}_i\}$ are available. This will result in more uncertainty when estimating R_0 and the vaccination coverage. If the uncertainty in measurement error is quantified it is possible to derive how this affects the uncertainty in the estimates by using the delta method. In the present paper measurement error has been neglected and such an analysis remains to be performed.

A different approach that is worthy of exploration is to use Markov chain Monte Carlo methods (see O'Neill *et al.* (2000) for an application of Markov chain Monte Carlo methods to epidemic models). The fact that R_0 is unidentifiable will also have consequences for such an approach. However, if the relative infectivities of the different types are equipped with informative prior distributions, this will induce a posterior distribution for R_0 indicating which part of the range of possible values is most likely.

Acknowledgements

This work was initiated by stimulating discussions with Sergei Utev and Niels Becker while visiting La Trobe University. Financial support from the Swedish Natural Science Research Council is gratefully acknowledged. Constructive suggestions from the Associate Editor and three referees have improved the paper considerably.

References

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

Anderson, R. M. and May, R. M. (1984) Spatial, temporal and genetic heterogeneity in host populations and the design of immunisation programs. *IMA J. Math. Appl. Med. Biol.*, **1**, 233–266.

——— (1991) *Infectious Diseases of Humans; Dynamic and Control.* Oxford: Oxford University Press.

Ball, F. and Clancy, D. (1993) The final size and severity of a generalised stochastic multitype epidemic model. *Adv. Appl. Probab.*, **25**, 721–736.

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

Becker, N. G. and Marschner, I. C. (1990) The effect of heterogeneity on the spread of disease. *Lect. Notes Biomath.*, **86**, 90–103.

Britton, T. (1998a) Preventing epidemics in heterogeneous communities. In *Proc. 19th Int. Biometric Conf.*, invited papers, pp. 109–115. Cape Town: International Biometric Society.

——— (1998b) Estimation in multitype epidemics. *J. R. Statist. Soc.* B, **60**, 663–679.

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

Farrington, C. P., Kanaan, M. N. and Gay, N. J. (2001) Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Appl. Statist.*, **50**, 251–292.

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

Grenfell, B. T. and Anderson, R. M. (1985) The estimation of age-related rates of infection from case notifications and serological data. *J. Hyg. Camb.*, **94**, 419–436.

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

Hethcote, H. W. and 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.* London: Wiley.

Nåsell, I. (1999) On the time to extinction in recurrent epidemics. *J. R. Statist. Soc.* B, **61**, 309–330.

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